Efficient Designs for Multiple Gene Knockdown Experiments

Bobak Nazer and Robert D. Nowak
University of Wisconsin, Madison

GENSIPS 2010
Motivation: Virus-Host Interaction

Which genes does the virus hijack for reproduction?

Issue: Redundancy

• Several similar studies have tried to ascertain genes involved in virus reproduction through knockdowns.

• However, a meta-analysis revealed that there is little overlap between the sets of genes discovered by each of these studies. Bushman et al., “Host cell factors in HIV replication: Meta-analysis of genome-wide studies,” PLoS Pathogens ’09.

• Some biologists have begun to suspect that there may be some redundancy among the involved genes.


• We will investigate (theoretically) whether this number of experiments can be significantly reduced.
If there is no redundancy, number of experiments is proportional to the number of genes.
With (pairwise) redundancy, the number of experiments could be as large as the number of genes squared.

Have to knockout both redundant genes to see an effect!
Compressed Sensing

Candès-Romberg-Tao IT Trans. ’06, Candès-Tao IT Trans. ’06, Donoho IT Trans. ’06

Linear measurements: Coefficients are S-sparse:

\[ y_k = \sum_i a_{ki} x_i = a_k^T x \]
\[ y = Ax \]

\[ \| x \|_0 = S \]
\[ x \in \mathbb{R}^N \]

If the measurement matrix satisfies the restricted isometry property (RIP) with \( \delta_{2S} < \sqrt{2} - 1 \) for all S-sparse vectors:

\[ (1 - \delta_S) \| x \|_2^2 \leq \| Ax \|_2^2 \leq (1 + \delta_S) \| x \|_2^2 \]

Then we can recover the coefficients through an \( \ell_1 \) optimization:

\[ \min \| \hat{x} \|_1 \text{ subject to } A \hat{x} = y \]

Can get RIP with \( K \geq cS \log(N/S) \) measurements.
Compressed DNA Microarrays


• Novel method for determining gene expression levels.

• **Sparse signal:** Very few genes active compared to size of genome.

• **Linear measurements:** Use a mixture of probes per spot, measure total fluorescence.

\[ y = \sum_{i} a_i x_i \]

- \( a_i \): 1 if probe is present, 0 if not.
- \( x_i \): 1 if gene is active, 0 if not.

• Can apply the framework of compressed sensing to drastically reduce the number of required experiments. (Also looked at sparse measurement matrices.)

• Is this possible when there are redundancies?
Approximate the output (virus expression) with a sparse bilinear system.

\[ y = \sum_i a_i x_i^{(1)} + \sum_{i<j} a_i a_j x_{ij}^{(2)} \]

- \( x_i^{(1)} \): 1 if gene is critical to pathway, 0 if not.
- \( x_{ij}^{(2)} \): 1 if gene pair is critical to pathway, 0 if not.
- \( a_i \): 1 if gene is knocked down.
Map to Linear System

\[{x_i}^{(1)}, {x_{ij}}^{(2)}\] → \(x\)  
Write coefficients as a vector

\[{a_{ki}}, {a_{kj}}\] → \(\tilde{a}_k\)  
Write measurements as a vector

Vectorize:

\[\text{Symmetrize:} \quad \tilde{y}_k = \tilde{a}_k^T x\]

\[y_k = \tilde{y}_{2k} - \tilde{y}_{2k-1}\]

Get measurements that look zero-mean, which simplifies the analysis.

Back to a linear problem: \(y = Ax\) with potentially dependent measurements.

Can we get down to \(S \log \left(\frac{N}{S}\right)\) measurements?
Approach 1: Gershgorin’s Disc Theorem

Approach inspired by Haupt-Bajwa-Raz-Nowak ’08:

i) Control each element of the Gram matrix $\mathbf{G}_\mathcal{R} = \mathbf{A}_\mathcal{R}^T \mathbf{A}_\mathcal{R}$ via Hoeffding’s inequality. Bound probability that Gram matrix is at worst:

$$
\mathbf{G}_\mathcal{R} = \begin{bmatrix}
1 & \frac{\delta_S}{S} & \cdots & \frac{\delta_S}{S} \\
\frac{\delta_S}{S} & 1 & \cdots & \frac{\delta_S}{S} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\delta_S}{S} & \frac{\delta_S}{S} & \cdots & 1
\end{bmatrix}
$$

ii) Gershgorin’s Disc Theorem guarantees that eigenvalues lie in the range:

$$
g_{ii} - \sum_{j \neq i} |g_{ij}| \leq \lambda_i(\mathbf{G}_\mathcal{R}) \leq g_{ii} + \sum_{j \neq i} |g_{ij}|
$$

iii) Union bound over all $\binom{N}{S}$ sparse patterns. Get RIP as long as number of measurements is at least:

$$
K = c \ S^2 \log N
$$
Approach 2: Tail Bounds

Typical approach to proving that RIP holds for linear problems:

i) Measurement matrix is composed of i.i.d. subgaussian random variables so each measurement is itself subgaussian:

\[ \Pr(|y_k| > t) < \exp(-ct^2) \]

ii) Show that measurement vector concentrates rapidly for a single sparsity pattern:

\[ \Pr\left(\left| \|y\|^2 - 1 \right| > t \right) < \exp(-cKt) \]

iii) Union bound over all \( \binom{N}{S} \) sparse patterns. Get RIP with constant long as number of measurements is at least:

\[ K = c \ S \log \left( \frac{N}{S} \right) \]
How exactly do the tails behave?

Rademacher Chaos of order D:

\[ y = \sum_{i_1 < i_2 < \cdots < i_D} a_{i_1} a_{i_2} \cdots a_{i_D} x_{i_1 i_2 \cdots i_D} \]

\( a_i \) are independent and binary symmetric (Rademacher).

The combinatorial dimension measures the level of dependence introduced by the sparsity pattern \( \mathcal{T}_L \). A chaos has combinatorial dimension \( \alpha \) if

\[
\sup_{A_1, A_2, \ldots, A_D} \frac{|\mathcal{T}_L \cap (A_1 \times A_2 \times \cdots \times A_D)|}{(\max_{1 \leq j \leq D} |A_j|)^\alpha} \leq C_1
\]

\[
|\mathcal{T}_L| \geq C_2 L^\alpha
\]

Takes values between \( 1 \leq \alpha \leq D \)
Chaos Tail Bounds

Blei-Janson ’04: A Rademacher chaos with combinatorial dimension satisfies:

$$\exp \left( -c_1 \frac{t^2}{\alpha} \right) \leq \sup_L \mathbb{P} (|y_k| > t) \leq \exp \left( -c_2 \frac{t^2}{\alpha} \right)$$

Using this bound it can be shown that:

$$\mathbb{P} \left( \left\| \|y\| - 1 \right\| \right) \leq \exp \left( -c \max \left( Kt/S, K^{1/\alpha} t^{1/\alpha} \right) \right)$$

from which we can derive the number of measurements needed to get RIP for a single pattern.

Union bound technique from Baraniuk-Devore-Davenport-Wakin ’08:

$$K \geq \min \left( c \ S^2 \log(N/S), \ c \ S^\alpha \log^\alpha(N/S) \right)$$
Approach 3: Heavy-Tailed Restricted Isometries

Rudelson-Vershynin ’08, Vershynin ’10:

If the expected Gram matrix is identity, $\mathbb{E} \left[ A^T A \right] = I$, and the rows are independent, then

$$K = c \frac{S}{\epsilon^2} \left( \log \left( \frac{S}{\epsilon^2} \right) \right)^3 \log N$$

measurements suffice to get an (expected) RIP constant less than $\epsilon$

One of the key points in the proof is that it avoids using a union bound and instead bounds the RIP constants of all sparsity patterns simultaneously.

Note that if the sparsity scales linearly with the problem size, this is just

$$K = cS \log^4 N$$
These simulations suggest that the number of required measurements is much closer to the sparsity level, rather than the problem size.
## Summary and Future Directions

<table>
<thead>
<tr>
<th>Method</th>
<th>Bound</th>
<th>Union Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gershgorin</td>
<td>$S^2 \log N$</td>
<td>Second Moment</td>
</tr>
<tr>
<td>Rudelson-Vershynin</td>
<td>$S(\log^3 S)(\log N)$</td>
<td>Second Moment</td>
</tr>
<tr>
<td>Rademacher Chaos Chaos</td>
<td>$S^{\alpha} \log^{\alpha} (N/S)$</td>
<td>Tail Bounds</td>
</tr>
</tbody>
</table>

Same bounds apply to general multilinear systems with $D$ redundant terms instead of 2. Allerton 2010.

---

It is currently possible to knockdown up to two or three genes in a single experiment with some accuracy. Higher order knockdowns (which we have assumed) are beyond current technical capabilities.

Definitely need to consider **sparse** measurement inputs.

One workaround may be to randomly knockdown genes and determine which ones were affected after the experiment.