Living with noise

Limsoon Wong

SinFra 2012
Biology is full of noise

- Experimental noise
- Intrinsic noise
Living with noise

• When determining whether the value of a biological entity is above or below a threshold, instead of first determining its exact value and comparing that to the threshold, determine a distribution of that value and see whether it is likely to be above or below the threshold.

• Instead of identifying and eliminating noise from samples, use bootstrap re-sampling to produce many bags of samples that are enriched with less noisy samples.

• Use noise-robust logic reasoning.
Batch Effect in Gene Expression Profiles
Headaches in gene expression analysis

Low % of overlapping genes from diff expt in general

- Prostate cancer
  - Lapointe et al, 2004
  - Singh et al, 2002
- Lung cancer
  - Garber et al, 2001
  - Bhattacharjee et al, 2001
- DMD
  - Haslett et al, 2002
  - Pescatori et al, 2007

Percentage of Overlapping Genes

<table>
<thead>
<tr>
<th>Datasets</th>
<th>DEG</th>
<th>POG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>Top100</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Lung Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top100</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>DMD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Top100</td>
<td>0.54</td>
<td></td>
</tr>
</tbody>
</table>

Zhang et al, Bioinformatics, 2009

Batch Effects

- Samples from diff batches are grouped together, regardless of subtypes and treatment response

Image credit: Difeng Dong’s PhD dissertation, 2011
Bootstrap sampling suppresses noise

- Suppose there are more “good” than “bad” samples in the training set.
- Then any collection of its bootstrap replicates is likely to be enriched with bags containing more “good” than “bad” samples.
Why bagging works

- Learning algo’s are well behaved

- Given learning algo C and training set S with more “good” than “bad” samples. Let $B_1, \ldots, B_n$ be bootstrap replicates of S. Then a bagging classifier based on a majority vote of classifiers $C(B_1), \ldots, C(B_n)$ is better than $C(S)$
Significantly improves cross-batch prediction accuracy in gene expression profile analyses.
Protein Interactome Cleansing
Why Biological Networks?

- Complete genomes are now available
- Knowing the genes is not enough to understand how biology functions
- Proteins, not genes, are responsible for many cellular activities
- Proteins function by interacting with other proteins and biomolecules

Slide credit: See-Kiong Ng
Identifying true PPIs in noisy expts

- PPIs are the basis of many biological mechanisms
- But there is a lot of noise in high-throughput PPI assays

Sprinzak et al., *JMB*, 327:919-923, 2003

<table>
<thead>
<tr>
<th>Experimental method category</th>
<th>Number of interacting pairs</th>
<th>Co-localization (%)</th>
<th>Co-cellar-role (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All: All methods</td>
<td>9347</td>
<td>64</td>
<td>49</td>
</tr>
<tr>
<td>A: Small scale Y2H</td>
<td>1861</td>
<td>73</td>
<td>62</td>
</tr>
<tr>
<td>A0: GY2H Uetz et al. (published results)</td>
<td>956</td>
<td>66</td>
<td>45</td>
</tr>
<tr>
<td>A1: GY2H Uetz et al. (unpublished results)</td>
<td>516</td>
<td>53</td>
<td>33</td>
</tr>
<tr>
<td>A2: GY2H Ito et al. (core)</td>
<td>798</td>
<td>64</td>
<td>40</td>
</tr>
<tr>
<td>A3: GY2H Ito et al. (all)</td>
<td>3655</td>
<td>41</td>
<td>15</td>
</tr>
<tr>
<td>B: Physical methods</td>
<td>71</td>
<td>98</td>
<td>95</td>
</tr>
<tr>
<td>C: Genetic methods</td>
<td>1052</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>D1: Biochemical, <em>in vitro</em></td>
<td>614</td>
<td>87</td>
<td>79</td>
</tr>
<tr>
<td>D2: Biochemical, chromatography</td>
<td>648</td>
<td>93</td>
<td>88</td>
</tr>
<tr>
<td>E1: Immunological, direct</td>
<td>1025</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>E2: Immunological, indirect</td>
<td>34</td>
<td>100</td>
<td>93</td>
</tr>
<tr>
<td>2M: Two different methods</td>
<td>2360</td>
<td>87</td>
<td>85</td>
</tr>
<tr>
<td>3M: Three different methods</td>
<td>1212</td>
<td>92</td>
<td>94</td>
</tr>
<tr>
<td>4M: Four different methods</td>
<td>570</td>
<td>95</td>
<td>93</td>
</tr>
</tbody>
</table>
Can noise be removed w/o more info?

- **Some common ideas to remove noise**
  - A PPI detected by two independent assays is more likely to be true
  - Two proteins participating in same biological process are more likely to interact
  - Two proteins in the same cellular compartments are more likely to interact

- **But these need additional expt and additional info**

- **Can we do better?**
Topology of neighbourhood of real PPIs

• Suppose 20% of putative PPIs are noise
  \[ \geq 3 \] purple proteins are real partners of both A and B
  \[ \Rightarrow \] A and B are likely localized to the same cellular compartment (Why?)

• Fact: Proteins in the same cellular compartment are 10x more likely to interact than other proteins
  \[ \Rightarrow \] A and B are likely to interact
Iterated CD Distance

- **CD-distance**
  \[ S(u, v) = 1 - D(u, v) = \frac{2X}{2X + (Y + Z)} \]

- **X** is the number of common neighbours of 1\textsuperscript{st} and 2\textsuperscript{nd} proteins
- **Y/Z** is the number of unique neighbours of 1\textsuperscript{st}/2\textsuperscript{nd} protein

- These counts are noisy. ∴ Use CD-distance to weigh these counts and recompute CD-distance

\[
w^k_L(u, v) = \frac{\sum_{x \in N_u \cap N_v} w^{k-1}_L(x, u) + \sum_{x \in N_u \cap N_v} w^{k-1}_L(x, v)}{\sum_{x \in N_u} w^{k-1}_L(x, u) + \sum_{x \in N_v} w^{k-1}_L(x, v) + \lambda^k_u + \lambda^k_v}
\]
Performance wrt Functional Homogeneity

Cf. ave functional homogeneity of protein pairs in DIP < 4%
ave functional homogeneity of PPI in DIP < 33%

- Ditto wrt localization coherence (not shown)
Consistency of Proteomic Profiles
Issues in Proteomic Profiling

• Coverage
• Consistency

⇒ Thresholding
– Somewhat arbitrary
– Potentially wasteful

• By raising threshold, some info disappears
Intuitive Example

- Suppose the failure to form a protein complex causes a disease
  - If any component protein is missing, the complex can’t form

⇒ Diff patients suffering from the disease can have a diff protein component missing
  - Construct a profile based on complexes?
“Threshold-free” Principle of PSP

Consistency: Samples segregate by their classes with high confidence
References & Acknowledgements

• Materials for this talk are from joint works with my students (Kenny Chua, Wilson Goh, Chuan Hock Koh) and postdoc (Guimei Liu):