How accurate is Gene Regulatory Network inference from perturbations?

Erik Sonnhammer

Stockholm Bioinformatics Centre
Science for Life Laboratory
Dept. Biochemistry and Biophysics
Stockholm University
Sonnhammer group: protein function prediction

- Protein domain families
- Orthology
- Genome/proteome analysis
- Gene regulatory networks
- Global Networks
- Pfam
- InParanoid
- FunCoup
1. The FunCoup network
   1. Application to Disease Gene Prediction
   2. Application to Pathway Analysis

2. Gene Regulatory Network (GRN) inference
   1. Using prior information
   2. Optimising sparsity
   3. Increasing data informativeness
FunCoup

- Protein-protein interactions
- Phylogenetic profiles
- Co-expression patterns
- Domain interactions
- Orthology
- Shared transcription factor binding
- Shared miRNA targeting
- Subcellular co-localisation
- Correlated Genetic interactions
- Other Organisms

2.0 Alexeyenko et al., *NAR* 40:D821 (2012)
1.0 Alexeyenko & Sonnhammer, *Genome Research* 19:1107 (2009)
Naïve Bayesian integration of evidences for 1 network link

Raw data

Bayesian LLR score

Sum of LLR scores

Confidence value $pfc$

---

Raw scores for cel_rad-50 - cel_him-1 and their orthologs

<table>
<thead>
<tr>
<th>Species</th>
<th>Data type</th>
<th>Dataset</th>
<th>Score type</th>
<th>Score</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ath</td>
<td>MEX</td>
<td>Expression map of development (Schmid et al., 2005)</td>
<td>PLC</td>
<td>0.845</td>
<td></td>
</tr>
<tr>
<td>cel</td>
<td>SCL</td>
<td>Gene Ontology (cellular compartment) (GO)</td>
<td>WMI</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>cel</td>
<td>MEX</td>
<td>Toxicant physiological mode of action (Swain et al., 2010)</td>
<td>PLC</td>
<td>-0.2</td>
<td></td>
</tr>
<tr>
<td>cel</td>
<td>DOM</td>
<td>Unification of domain-domain Interactions (Bjorkholm &amp; Sonnhammer, 2009)</td>
<td>UniDomint sum</td>
<td>0.364</td>
<td></td>
</tr>
<tr>
<td>rfa</td>
<td>MEX</td>
<td>Canine normal tissue (Briggs et al., 2010)</td>
<td>PLC</td>
<td>0.468</td>
<td></td>
</tr>
<tr>
<td>rfa</td>
<td>MEX</td>
<td>Distinguishing features of canine hemangiosarcoma (Beth et al., 2010)</td>
<td>PLC</td>
<td>0.303</td>
<td></td>
</tr>
<tr>
<td>dre</td>
<td>MEX</td>
<td>Zebrafish development stages (Domazet-Loso, 2010)</td>
<td>PLC</td>
<td>0.701</td>
<td></td>
</tr>
<tr>
<td>gsa</td>
<td>MEX</td>
<td>Hepatic transcriptional profiling (Trakooljul et al., 2011)</td>
<td>PLC</td>
<td>-0.055</td>
<td></td>
</tr>
<tr>
<td>gsa</td>
<td>MEX</td>
<td>Regulation of the early embryonic RPE (Sun et al., 2008)</td>
<td>PLC</td>
<td>0.213</td>
<td></td>
</tr>
<tr>
<td>hsa</td>
<td>MEX</td>
<td>Derivation of septic shock subgroups (Wong et al., 2009)</td>
<td>PLC</td>
<td>0.609</td>
<td></td>
</tr>
<tr>
<td>hsa</td>
<td>MEX</td>
<td>9 different cancer tissues (Reinhold et al., 2011)</td>
<td>PLC</td>
<td>0.528</td>
<td></td>
</tr>
<tr>
<td>hsa</td>
<td>PPI</td>
<td>IntAct (Aranda et al., 2010)</td>
<td>PPI score</td>
<td>0.011</td>
<td></td>
</tr>
</tbody>
</table>
Major network component
FunCoup 3.0 network sizes
Search for functional couplings

Gene Identifier(s): Myc
Species: Homo sapiens

[Search]
FunCoup.sbc.su.se

Summary

Evidence

Species
<table>
<thead>
<tr>
<th>Interaction partners</th>
<th>Confidence</th>
<th>Network</th>
<th>Evidence types</th>
<th>Species</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Myc</strong> v-myc myelocytomatosis ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUVBL2</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUVBL1</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GNL3</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPM1</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FBL</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDAC1</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOP2A</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTL6A</td>
<td>1.000</td>
<td>Complex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDAC3</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MYBBP1A</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDK1</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOL11</td>
<td>1.000</td>
<td>Complex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NCL</td>
<td>1.000</td>
<td>Complex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAX</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOPBP1</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EIF3E</td>
<td>1.000</td>
<td>PPI</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Details**

- 1 publication: Fwing RM
- Shared locations: GO:0005654, GO:0031981, GO:0070013, GO:0043233, GO:0005730
- Correlation of 0.802 in GSE7307
- Correlation of 0.691 in GSE7404
- Correlation of 0.717 in GSE26440
- Correlation of 0.652 in GSE10246
- Correlation of 0.875 in GSE23093
- Correlation of 0.534 in GSE20113
- Correlation of 0.553 in GNF1H_RMA
- Correlation of 0.660 in GSE10538
- Correlation of 0.630 in GSE16233
- Correlation of 0.436 in GSE9954
- Correlation of 0.271 in GSE29288
- Correlation of 0.154 in GSE9745
- Correlation of 0.165 in GSE29685
- Correlation of 0.206 in GSE23760
MaxLink on-line

Maxlink search for top interactors?

Gene Identifier(s):
ART1  ATF1  ASXL1  ATM  BRD3  BRD4  CREB1  CTNNB1  CBFB  CREBBP  CARS  DDX6  DEK  Dicer1  EWSR1  ERCC3  FUS  GMPS  MEN1  MSI2  MLH1  MSH2  MSH6  MYH9  NRAS  NONO  NFKB2  NSD1  NPM1  PTEN  PBRM1  RBM15  SFPQ  SDHB  SDHD  SMARCA4  SMARCB1  TCF3  TPR

Read from file: Browse... No file selected.

Species: Homo sapiens

Confidence threshold: 0.8

Candidates: 30

Search
"Maxlink" mining for novel cancer genes

- Known cancer genes
- Maxlink ranking
- Annotation filter
- More Q links than expected
- Candidate cancer genes

MaxLink: network-based prioritization of genes tightly linked to a disease seed set.
Guala, Sjölund, Sonnhammer
Bioinformatics. 30:2689-2690 (2014)
Red: known cancer genes
Blue: 185 novel candidates with 
>= 10 links to known

"Network-based identification of novel cancer genes"
Östlund, Lindskog, and Sonnhammer
Classical Gene Enrichment vs Network Crosstalk Analysis

Gene overlap enrichment (e.g. Fisher exact test)

Gene list

Network crosstalk analysis

Gene list

Pathways
Network Crosstalk Analysis (NCA) with CrossTalkB / FunCoup

Links within/between groups in original network

A  B  C

intra-crosstalk

inter-crosstalk

Links within/between groups in random network

A  B  C

"Statistical Assessment of Crosstalk Enrichment between Gene Groups in Biological Networks"
McCormack et al.
PLoS ONE, 8:e54945 (2013)
Benchmarking NCA and GEA

Task: associate (FDR < 0.01)

- 2392 experimental gene groups (MsigDB)
- 289 pathways (KEGG)
- 691288 potential group-pathway associations

Gene overlap enrichment (Fisher exact)

- NCA: 168390
- 1785
- 12520
Novel pathway annotation example

MsigDB gene group:
SETLUR_PROSTATE_CANCER_TMPRSS2_ERG_FUSION_UP
(63 genes up-regulated in prostate cancer)

- Gene overlap enrichment: 0 overlapping KEGG pathways

- CrossTalkZ/FunCoup: 9 significant KEGG pathways, e.g.
  “Prostate cancer”
  “Pathways in cancer”
Network Crosstalk Analysis (NCA) with Network Clustering

1. Find Network Clusters
   - Gene List 1
   - Gene List 2
   - Gene List 3

2. Find Crosstalk Enrichment/Depletion
   - KEGG pathways
   - GO terms
Module finding with MGclus


Dynamical Systems Biology

- Static network
- Associations
- Undirected links

- Regulatory network
- Direction + sign
- Causal relations
Gene Regulatory Network (GRN)

- Any link (edge) is allowed, not just from TFs
- A link represents information flow between modeled nodes, and may be indirect.
Perturbations can give causal networks

Ideally moderate up or down pushes, not 100% knockout
Inference of Gene Regulatory Networks (GRNs)

• Model: \( Y = -PA^{-1} \) (inverse problem)

  \( Y \): gene expression matrix (output)
  \( A \): network matrix
  \( P \): perturbation matrix (input)

• Seek a GRN \( A \) that can predict gene expression responses \( Y \)
  after perturbations \( P \).

• >70 modelling techniques exist for regulatory network
  inference: Least Squares, LASSO, Elastic nets, ARACNE,
  CLR, NIR, Bayesian networks, etc.
Perturbation-based Gene Regulatory Network (GRN) inference

Perturb system → Measure [mRNA] responses → Inference algorithm

\[ \theta_N = \arg \min_{\theta} V_N(\theta, Z^N) \]

GRN Model
Problems with Gene Regulatory Network Inference

1. Vast search space
   - 4-gene network: 16 links, 3 states/knot
     \[ \text{16 links, 3 states/knot} \rightarrow 43 \text{ million topologies, } (3^{16}) \]
     Functional association networks as priors for gene regulatory network inference.
     Studham, Tjärnberg, Nordling, Nelander, Sonnhammer
     *Bioinformatics* (2014) 30:i130

2. Dependence on sparsity
   - Needs to be optimised.
   "Optimal sparsity criteria for network inference"
   Tjärnberg, Nordling, Studham, and Sonnhammer

3. Uninformative data
   - High noise levels and ill-conditioned data
   - Single-gene perturbations are generally insufficient to unravel all causal links.
   Make data better conditioned by iterative multi-gene perturbations to excite system in all directions.
Perturbation-based GRN inference with prior

- **Perturb system**
- **Measure [mRNA] responses**
- **Inference algorithm**

Prior

GRN Model
Synthetic Data Workflow

- Generate true networks
- Generate expression data
- Generate priors
- GRN inference
- Evaluate results
Inference methodology

- **GRN inference method:**
  \[ Y = -A^{-1}P + \mathcal{E} \]

\[
\text{minimize}_{A} \quad \|AY + P\|_F + \zeta \sum_{i} \sum_{j} (1-w_{ij})|a_{ij}|
\]

- **Y** expression data matrix
- **A** influence matrix (network)
- **P** perturbation matrix
- **E** input error
- **ζ** sparsity
- **w** prior weight, \(0 \leq w \leq 1\)
Inference accuracy improvement using simulated prior (GeneNetWeaver-generated data)

70% prior accuracy

90% prior accuracy
Biological Data Workflow

YEASTRACT

Knockout Experiments (Hu et al., 2007)

Priors from FunCoup & STRING

GRN inference

Evaluate results
Inference accuracy improvement using biological data

FunCoup

Yeastract network sparsity

STRING
How to set network sparsity?

Most methods use arbitrary sparsity.

”Optimal sparsity criteria for network inference”
Tjärnberg, Nordling, Studham, Sonnhammer EL.
Sparsity term dependence
Network Inference Pipeline

LOOCO = Leave One Out Cross-Optimisation

Build network model. Predict left-out data.

Measure prediction error

Repeat for all sparsity levels
Sparsity Optimisation

![Graph showing prediction error and similarity of topology against sparsity (penalty)]

- **Prediction error**: Y-axis
- **Similarity of Topology**: Y-axis
- **Empty network** vs. **Fully connected**

The graph illustrates how prediction error and similarity of topology change as the sparsity (penalty) increases from empty network to fully connected network.
• Less informative data:

• Uninformative:
Data informativeness and GRN inference

- Simulate expression data after perturbation of known GRN.
- Measure accuracy as MCC (TP = link in same direction).
- Vary **SNR** (Signal to Noise Ratio)
- Vary **condition number** (= Ratio between largest and smallest singular values of the network matrix)
GRN inference accuracy, low condition nr

- 10 genes
- 20 samples
- Condition nr = 1-2
GRN inference accuracy, high condition nr

- 10 genes
- 20 samples
- Condition nr = 10-180

Gardner et al.
Lorenz et al.
Properties of real perturbation expression data

<table>
<thead>
<tr>
<th></th>
<th>Gardner et al., 2003</th>
<th>Lorenz et al., 2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>#genes</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Condition nr of network matrix</td>
<td>54</td>
<td>253</td>
</tr>
<tr>
<td>Condition nr of expression matrix</td>
<td>154</td>
<td>215</td>
</tr>
<tr>
<td>SNR* worst – best</td>
<td>0.002 – 0.1</td>
<td>0.002 - 0.2</td>
</tr>
</tbody>
</table>

*) SNR depends on estimated variance, which differs between genes
GRN inference accuracy, low condition nr

- 45 genes
- Condition nr = 26-41
GRN inference accuracy, high condition nr

- 45 genes
- Condition nr = 413-505

Gardner et al.
Lorenz et al.
Why is real data uninformative?

1. Biological variation gives low Signal-to-Noise-Ratio
2. Poor experimental design gives ill-conditioned data

What to do?
How to get informative data?

1. Multi-gene perturbations. *From (Nordling and Jacobsen, 2009):*

2. Iterative perturbation design. *From (Nordling and Jacobsen, 2007):*
Perturbation-based network inference

1. Perturb system
2. Measure [mRNA]
3. Inference algorithm
4. Model

Validation
Iterative experiment design

1. Perturb system
2. Measure [mRNA]
3. Inference algorithm
4. Experiment design
5. Validation
6. Model

by Torbjörn Nordling, tn@kth.se
The MYC oncogene

- Transcription factor that binds to E-boxes
- Promotes rapid cell growth & proliferation
- Normal expression:
  - Upregulated in embryos
  - Downregulated in specialized adult cells
- Failure of regulation leads to cancer
- ”Universal amplifier”
Unravelling the MYC gene regulatory network

• **Design**
  – 45 genes
  – Single and double knockdowns (at Cell Screening facility)
  – Spiked in reference RNA

• **Results:**
  – 18432 qPCRs performed for 192 samples
  – Lower sensitivity than regular qPCR ⇒ some failed
  – Noise levels normal, network inference “hard”
  – MYC appears to go up when knocked down
MYC regulatory network
raw results
Network inference of MYC GRN

- LSCO with sparsity optimisation did not work well – the empty network had lowest error.

- Instead, used Bootstrap-LASSO:
  - 1000 bootstrap GRN inferences from sampled data.
  - 100 runs of bootstraps to gauge consistency.
  - Contrast with results from shuffled data.
Bootstrap procedure

Sample with replacement 1000 times
Infer GRNs at given sparsity

Repeat 100 times, generate consensus GRN at different bootstrap cutoffs
Overlap between bootstrap runs

Bootstrap results at sparsity 0.75

Overlap shuffled data

Support, shuffled data

Support, real data

Bootstrap support
"Best” GRN

- Consensus GRN with 157 links supported in 100 bootstrap GRNs.

- Sparsity and bootstrap cutoff (0.965) optimised to yield the largest GRN in which repeated bootstrap networks overlap more than expected by chance.

- Expected from randomly shuffled data with the same parameters: 0 links.
Validation of network

- 35 two-gene perturbations, not included for network inference were performed.
- Network-predicted and observed effect highly correlated (Spearman $r=0.87; p<10^{-85}$).
- "Trivial" prediction using mean effect of both single experiments gave $r=0.35$. 
Conclusions

1. FunCoup.sbc.su.se is a resource to discover novel functional coupling between genes.

2. FunCoup applications include Network Crosstalk Analysis for Pathway annotation and MaxLink for candidate (disease) gene prediction.

3. Gene regulatory network (GRN) inference from perturbation responses is challenging due to unknown sparsity and low informativeness.

4. GRN inference can be improved by sparsity optimisation and prior information.

5. GRN inference is expected to give low accuracy for biological data but can be improved by iterative multi-gene perturbation experiments.
Acknowledgements

Matt Studham
Andreas Tjärnberg
Torbjörn Nordling
Daniel Morgan
Christoph Ogris
Mobashir Mohammad
Thomas Schmitt
Dimitri Guala
Oliver Frings

http://FunCoup.sbc.su.se
http://Sonnhammer.org