Predicting the targets of mRNA-binding proteins
+ GeneMANIA & ISOLATE

Quaid Morris
Banting and Best Department of Medical Research,
Departments of Computer Science and Molecular Genetics
Outline

mRNA-binding proteins

Collaborators: Paul Boutros & Syed Haider (OICR), Blaise Clarke (UHN)

ISOLATE for in silico purification of tumour samples

Collaborators: Tim Hughes & Deb Ray (Donnelly), Howard Lipshitz (U of T, MoGen), Wei Jiao (U of T, MoGen)

Ang Cui, Amit Deshwar (no pictures avail)

Collaborators: Paul Boutros & Syed Haider (OICR), Blaise Clarke (UHN)
GeneMANIA: gene function prediction for the masses

Gary Bader

Sara Mostafavi

David Warde-Farley

Khalid Zuberi

Christian Lopes

Jason Montojo

Harold Rodriguez

Max Franz

Chris Grouios

Farzana Kazi

Deb Ray
Other projects in the lab

Cancer networks (w/ Sara and Gerald)
GWAS & QTLs of complex disease (w/ David)

Anna Goldenberg, PhD

microRNA target prediction

Hossein Radfar
mRNA-binding proteins
Post-transcriptional regulation (PTR)

transcription

splicing

poly-adenylation

export

localization

mRNA stability

mRNA translation rate

genomic DNA

pre-mRNA

spliced mRNA

mature mRNA

Nucleus

Cytoplasm
Patterns of mRNA localization before the MZT

Red: Nucleus
Green: mRNA

Adapted from LeCuyer et al, (2008) Cell 131
Recipe for modeling PTR

• Identify the trans-acting regulators:
  – RNA-binding proteins and other ncRNAs ✔

• Identify sequences recognized by ncRNAs (e.g. miRNAs):
  – TargetScan, PicTar, miRanda, GenMiR, PITA, rna22, etc. ✔

• Identify sequences recognized by RNA-binding proteins:
  – RBPDB (http://rbpdb.ccbr.utoronto.ca) is a start ✔

• Design algorithms to predict post-transcriptional fate given cis-elements and the mRNA sequence. ✔
RNAcompete Method

1. Agilent microarray
2. Primer Extension
3. Linker Ligation
4. Strip ssDNA
5. dsDNA Pool
6. PCR
7. Hybridize To Array/Computational Analysis
8. Label pulldown RNA
9. RNA Pool
10. Unstructured RNA
11. Structured RNA (7loop)
12. Structured RNA (8loop)
13. RNA Pool
14. Affinity Matrix
15. RBP
16. Label RNA pool
17. Label pulldown RNA
18. Hilal Kazan

Deb Ray and Tim Hughes
Validation of RNAcompete

10 highest ratio probes
AGACCUUCAUGUUCUGUUCUUUUUACUGCUUUGGUGGU
AGAAAGUUUGUUUGGUGUGUGCUCUCUGUGUGCA
AGAUUUUAAUCUUCAUGCUUACGUACCUCUCUGUUGGAA
AGGCUUUCGUUUGUGUACUCUCCUAGGUUUCUUUGGC
AGGAACUUAGUUUGGGUUUACUUUGUGCCUCACCGU
AGGAUGACCUAGGCUAAGUUGCUAUGGUAGGGCCU
AGAAUAGUUGCUAAUUGAAGGUUGUCUAGGCCACAC
AGAUCAGAGAUAACCGUAGUAGUAUGUUUGCU
AGAAUCCAGAUAUAGCACUGUAGUAGUAGGUAGGU
AGAAUUGUUAGGUUGUGACCAUCUGCUUAGGU
AGAUCGUAAACUCGUUUGUUGCUCAACGUUGCA

(RED indicates motif match)

HuR

Log affinity from Set A (a.u.)

0
6
12
18

Log affinity from Set B (a.u.)

0
6
12
18

Vts1

10 highest ratio probes
AGG
AGGCUCAAG
AGGGGG
AGGGCC
AGGCG
AGGCC
AGGC
AGGC
AGGA
AGGA

(RED indicates motif match
BLUE indicates intended paired bases
GREEN indicates predicted paired bases)
SetA and SetB 7-mers are highly correlated for all 9 RBPs profiled and reflect known sequence binding preferences where known.
RIP-chip for detecting in vivo RBP targets
**YB1** RNAcompete predicts its *in vivo* targets

YB1 co-immunoprecipitates (cDNAs) from (Dong, J. et al. RNA Biol 2009), 95 positives, 99 negatives
SF2/ASF RNAcompete predicts its *in vivo* targets

Clip-seq data from (Sanford et al Genome Research 2009) 296 positives, 314 negatives
RNAcompete summary

• Measures in vitro affinity of an RBP for ~250k custom-designed 35-mer RNA sequences,

• These affinities can be summarized into highly reproducible 7-mer scores (or affinities),

• RNAcompete 7-mer scores predict *in vivo* targets for the RBPs that we’ve looked at.
Good news:

We’re doing RNAcompete for all Drosophila and human RBPs!
Bad news:

RBP sequence binding preferences alone might not be good enough.
Most potential Puf3p sites are not Puf3p-bound

- mRNAs whose 3'UTR contains a Puf3p site
- Puf3p bound mRNAs
- mRNAs in control but not Puf3p bound

Sequence-specific RBPs bind ssRNA

DNA: B-Form Helix
RNA: A-Form Helix
Other RBPs recognize different RNA structures

Weeks et al, Cell (1992) 70:1069
1. Does the secondary structure of naked mRNA constrain RBP binding?

2. Are mRNA secondary structure predictions made with current tools useful?
Accessibility: Probability being unpaired in the thermodynamic ensemble

mRNA

Ensemble:

Target Site Accessibility

5' 1 0.5 0 3'

5' 3'
RNAplfold: local thermodynamic folding

mRNA transcript

Local folding windows
Target site accessibility predicts Puf3p binding

Area under the ROC curve: 74%

Median site accessibility

(P < 2 x 10^{-14}, ranksum test)
Scoring accessibility for RBPs with >1 binding sites

**RIP-chip result**

- **Bound**
- **Unbound**

**Consensus sites**

**Ranking by total accessibility (#ATS)**

- #ATS Score
  - Bound: 2.7
  - Unbound: 1.7

**Ranking by site number (#TS)**

- #TS Score
  - Bound: 3 and 3
  - Unbound: 2 and 2

**ROC curve for predicting in vivo binding**

- Area under curve (AUROC): 1.0 vs. 0.5
Accessibility significantly increases predictive accuracy for 71% of RBPs tested

* indicates a significant improvement in AUROC (FDR < 0.05, Delong-Delong-Clarke-Pearson test, BH correction)
1. Does the secondary structure of naked mRNA constrain RBP binding?

Yes.

2. Are mRNA secondary structure predictions made with current tools useful?

Yes, much more useful than expected.

Also, we have developed an in silico benchmark for testing predictions of mRNA secondary structure.
RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins

Hilal Kazan¹, Debasish Ray², Esther T. Chan³, Timothy R. Hughes²,3,4, Quaid Morris¹,2,3,4*

¹ Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, ² Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada, ³ Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada, ⁴ Donnelley Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada

Abstract

Metazoan genomes encode hundreds of RNA-binding proteins (RBPs). These proteins regulate post-transcriptional gene expression and have critical roles in numerous cellular processes including mRNA splicing, export, stability and translation. Despite their ubiquity and importance, the binding preferences for most RBPs are not well characterized. In vitro and in vivo studies, using affinity selection-based approaches, have successfully identified RNA sequence associated with specific RBPs; however, it is difficult to infer RBP sequence and structural preferences without specifically designed motif finding methods. In this study, we introduce a new motif-finding method, RNAcontext, designed to elucidate RBP-specific sequence and structural preferences with greater accuracy than existing approaches. We evaluated RNAcontext on recently published in vitro and in vivo RNA affinity selected data and demonstrate that RNAcontext identifies known binding preferences for several control proteins including HuR, PTB, and Vts1p and predicts new RNA structure preferences for SF2/ASF, RBM4, FUSIP1 and SLM2. The predicted preferences for SF2/ASF are consistent with its recently reported in vivo binding sites. RNAcontext is an accurate and efficient motif finding method ideally suited for using large-scale RNA-binding affinity datasets to determine the relative binding preferences of RBPs for a wide range of RNA sequences and structures.

Citation: Kazan H, Rav D, Chan ET, Hughes TR, Morris O (2010) RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-
Representing structural context by annotating bases

- Hairpin loop
- Internal loop
- Hairpin loop
- Multiloop
- Bulge loop
- External region
- Stem
- P: paired
- U: unpaired
- L: loop
- M: miscellaneous
- 5' 3'
Representing ensembles of structures using our structural annotation (i)

AGACGCGCGCGUUCGCCGCGCUCGGCGCAUGC

P: paired  L: loop  M: miscellaneous  U: unstructured

AGACGCGCGCGUUCGCCGCGCUCGGCGCAUGC
1  UPPMPPPPPLLLLLPPMPMMMMMPPPPLLLLLPP
2  UUUUPMPPPPPLLLLLPPMPMMMMMPUUUUUUUUUU
3  UUUUUPMPPPPMPPMMMPPLLLLLPPMPMMMMMPUUUUU
4  UPMPPLLLLLPPMPPPPPLLLLLLPPPPUUUUUUU
5  UUUUUUUUUPPPPMPPPPPPLLLLLPPPPPMMPPPP

...
RNAcontext inputs and outputs

RBP affinities

sequences and annotation profiles

GGGAUACCC

GGGCUACCG

CCCAUACCC

GGGAAACCC

RNAcontext

base preference

Φ

structural preference

Γ

relative preference

P

L

M

U

0
RNAcontext motif model

\[
\text{binding site score} = \Phi \text{base identity} \times \Gamma \text{structural context}
\]

## Comparison with other motif models

### Average precision on held-out data

<table>
<thead>
<tr>
<th>Proteins</th>
<th>RNAcontext</th>
<th>MEMERIS</th>
<th>MatrixREDUCE</th>
<th>Precision improvement</th>
<th>Reduction in error</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBM4</td>
<td>91%</td>
<td>43%</td>
<td>63%</td>
<td>28%</td>
<td><strong>75.7%</strong></td>
</tr>
<tr>
<td>Fusip</td>
<td>53%</td>
<td>31%</td>
<td>32%</td>
<td>21%</td>
<td><strong>30.9%</strong></td>
</tr>
<tr>
<td>VTS1</td>
<td>65%</td>
<td>58%</td>
<td>56%</td>
<td>7%</td>
<td><strong>16.7%</strong></td>
</tr>
<tr>
<td>YB1</td>
<td>17%</td>
<td>7%</td>
<td>11%</td>
<td>6%</td>
<td><strong>6.8%</strong></td>
</tr>
<tr>
<td>SLM2</td>
<td>81%</td>
<td>49%</td>
<td>77%</td>
<td>4%</td>
<td><strong>17.4%</strong></td>
</tr>
<tr>
<td>SFRS1 (SF2/ASF)</td>
<td>70%</td>
<td>50%</td>
<td>66%</td>
<td>4%</td>
<td><strong>11.8%</strong></td>
</tr>
<tr>
<td>HuR</td>
<td>96%</td>
<td>74%</td>
<td>94%</td>
<td>2%</td>
<td><strong>33.3%</strong></td>
</tr>
<tr>
<td>PTB</td>
<td>69%</td>
<td>26%</td>
<td>67%</td>
<td>2%</td>
<td>6.1%</td>
</tr>
<tr>
<td>U1A</td>
<td>30%</td>
<td>27%</td>
<td>21%</td>
<td>3%</td>
<td>4.1%</td>
</tr>
</tbody>
</table>
i) Predicted motifs and structural contexts

<table>
<thead>
<tr>
<th>Domain(s)</th>
<th>Previously reported binding site</th>
<th>RNAcontext predicted motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vts1p</strong></td>
<td>one SAM domain</td>
<td><img src="image1" alt="RNA motif" /></td>
</tr>
<tr>
<td></td>
<td>source: [16]</td>
<td>context: P L L L L L L</td>
</tr>
<tr>
<td><strong>SLM2</strong></td>
<td>one KH domain</td>
<td><img src="image2" alt="RNA motif" /></td>
</tr>
<tr>
<td><strong>RBM4</strong></td>
<td>full-length; two RRM domains</td>
<td><img src="image3" alt="RNA motif" /></td>
</tr>
<tr>
<td><strong>SF2/ASF</strong></td>
<td>two RRM domains</td>
<td><img src="image4" alt="RNA motif" /></td>
</tr>
<tr>
<td></td>
<td>source: [42]</td>
<td>context: M M M M U</td>
</tr>
<tr>
<td><strong>FUSIP1</strong></td>
<td>one RRM domain</td>
<td><img src="image5" alt="RNA motif" /></td>
</tr>
<tr>
<td><strong>HuR</strong></td>
<td>full-length; three RRM domains</td>
<td><img src="image6" alt="RNA motif" /></td>
</tr>
<tr>
<td><strong>PTB</strong></td>
<td>full-length; four RRM domains</td>
<td><img src="image7" alt="RNA motif" /></td>
</tr>
<tr>
<td></td>
<td>source: [43]</td>
<td>context: U U U U U U U</td>
</tr>
</tbody>
</table>

Legend

- **Paired**
- **Loop**
- **Misc**
- **Unstructured**

ii) Relative structural preferences
RNA-map from SFRS1 knock-down

Alternatively spliced exon sequences

Change in splicing level
- more exclusion
- no change
- more inclusion

RNAcontext predicted sequence preferences

Structural preference:
- Loop: 1.0
- Multi/Internal: 0.5
Summary

RNAcontext:

• is a new discriminative motif model for finding RBP binding preferences

• incorporates structural information with a new way of representing secondary structure

• recovers known sequence and structure binding preferences of most RBPs
Acknowledgements: mRNA

Morris Lab
Hilal Kazan
Xiao Li, Gerald Quon
Timothy Hughes
Debashish Ray
Kate Cook
Esther Chan: *PWMs*
Shaheynoor Talukder
Howard Lipshitz

Benjamin Blencowe
Arneet Saltzman: SF2 knockdown
Sidharth Chaudhry: *cloning*

Helpful Feedback:
Craig Smibert (U Toronto)
Howard Lipshitz (U Toronto)
Frank Sicheri (U Toronto)
Mariana Kekis (Hughes Lab)

Reagent gifts:
Carol Lutz: U1A (UMDNJ)
Craig Smibert: Vts1
Mariano Garcia-Blanco: PTB (Duke)
Kimitoshi Kohno: YB1 (U of OEH)
The GeneMANIA project: gene function prediction for the masses
Genomics revolution, the bad news

Functional genomic datasets are:

- noisy,
- redundant,
- incomplete,
- mysterious,
- massive
How can genomic data be used in the lab?

- Microarray expression profiles
- Genetic interactions
- CHiP-chip regulatory networks
- Protein-protein interactions
Use cases:

1. What is the function of this gene?

2. Give me more genes like these.

3. Here’s a list of genes -- what does it mean?
Recommender systems

Google Sets

Automatically create sets of items from a few examples.

Enter a few items from a set of things. (example)
Next, press Large Set or Small Set and we'll try to predict other items in the set.

- edinburgh
- glasgow
- stirling
- dundee
- inverness
- falkirk
- perth
- paisley
- dunfermline
- avr
- kilmarnock
- hamilton

(clear all)

Large Set  Small Set (15 items or fewer)
Google can’t do biology

Automatically create sets of items from a few examples.

Enter a few items from a set of things. (example)
Next, press Large Set or Small Set and we’ll try to predict other items in the set.

- CDC27
- APC11
- APC4
- CDC26
- DOC1

(clear all)

Large Set  Small Set (15 items or fewer)
GeneMANIA gene scoring via label propagation

Guilt-by-association

MCA1 – CDC48 – CPR3 – TDH2

Label propagation

MCA1 – CDC48 – CPR3 – TDH2
The problem of gene multi-functionality

• Gene function could be a/the:
  – Biological process,
  – Biochemical/molecular function,
  – Subcellular/Cellular localization,
  – Regulatory targets,
  – Temporal expression pattern,
  – Phenotypic effect of deletion.

Some networks may be better for some types of gene function than others
Three rules for network weighting

– Reliability
  • Noisy networks get low weight

– Relevance
  • Irrelevant networks get low weight

– Redundancy
  • Redundant weights share their weights
http://www.genemania.org/plugin/
GeneMANIA

Available Data

<table>
<thead>
<tr>
<th>Organisms</th>
<th>Networks</th>
<th>Genes</th>
<th>Interactions</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>76</td>
<td>20247</td>
<td>9394174</td>
<td>2010-04-28</td>
</tr>
</tbody>
</table>

Choose Query Genes

Organism: C. elegans (worm)

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>unc-18 (UNC18_CAEEL)</td>
<td>unc-18 encodes the C. elegans ortholog of Saccharomyces cerevisiae SEC1 and mammalian Munc18 proteins. It plays a role in vesicle transport.</td>
</tr>
<tr>
<td>unc-30 (UNC30_CAEEL)</td>
<td>unc-30 encodes a homeodomain-containing protein that is orthologous to the Pitx family of homeodomain transcription factors. It is involved in segmentation and other processes.</td>
</tr>
<tr>
<td>unc-4 (UNC4_CAEEL)</td>
<td>The unc-4 gene encodes a paired-class homeodomain protein with homologs in Drosophila and vertebrates. It plays a role in neurodevelopment.</td>
</tr>
<tr>
<td>unc-5 (UNC5_CAEEL)</td>
<td>unc-5 encodes a netrin receptor. unc-5 activity is required for cell autonomous for dorsalward cell and pioneer cell pathways.</td>
</tr>
</tbody>
</table>

Choose Interaction Networks

Select: all, none, default.

- Co-expression (3/10)
- Co-localization (1/1)
- Genetic interactions (2/4)
- Other (0/1)
- Physical interactions (4/8)
- Predicted (0/50)
- Shared protein domains (0/2)

Find the top 10 related genes using automatic weighting. Start
http://cytoscapeweb.cytoscape.org/
Data providers

+ BioGRID
+ Pathway Commons
+ ENSEMBL
+ MODs
+ Human Protein Reference Database
+ HUMANCYC
+ Reactome
+ Memorial Sloan-Kettering Cancer Center
+ MINT
+ IntAct
+ SBCN
+ www.FunctionalNet.org
+ bioPIXIE
Who?

Students
- Sara Mostafavi
- David Warde-Farley
- Ovi Comes
- Khalid Zuberi
- Farzana Kazi

Principal Investigators
- Quaid Morris
- Gary Bader

Developers
- Christian Lopes
- Max Franz
- Jason Montojo

Outreach
- Sylva Donaldson
GeneMANIA: the next steps

1. More data, organisms, and functions
2. Grouping functions together
3. Better handling of profiling data
ISOLATE: a tool for purifying heterogeneous tumour array data
Tumour samples are heterogeneous

A) An endometrial endometrioid carcinoma with significant amounts of healthy tissue.

B) A sample from the same endometrial endometrioid carcinoma, but almost exclusively tumor.
Acknowledgements

Morrislab
Gerald Quon
Ang Cui
Khalid Zuberi
Amit Deshwar

Paul Boutros (OICR)
Syed Haider

Blaise Clarke (UHN)

Peter Zandstra (Donnelly)
Wendy Qiao

Rae Yeang (Sick Kids)

Funding

NSERC CRSNG
Ontario

Canada Foundation for Innovation
Fondation canadienne pour l’innovation
Questions?
RNA probe set / pool design

Two independent 35nt probe sets (SetA and SetB), each probe set contains:

<table>
<thead>
<tr>
<th>Sequence type</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured or weakly structured sequences</td>
<td>81% of all 10-mers &gt;= 12 repeats of all 8-mers &gt;= 64 repeats of all 7-mers</td>
</tr>
<tr>
<td>Stem-loops containing loops with &lt;= 7 bases</td>
<td>99 – 100% represented, some 6-loops are repeated. 10bp length stems</td>
</tr>
<tr>
<td>Stem-loops containing loops with 8 bases</td>
<td>59% represented 10bp length stems</td>
</tr>
</tbody>
</table>

+ ~30k replicate probes + controls on an Agilent 244k format

Now also available: 8-mer design
7-mer scores more accurately predict transcripts that co-purify with PTB

PTB co-immunoprecipitates (cDNAs) from (Gama-Carvalho, M et al GB 2006)
2,951 positives, 2,052 negatives