Systematic discovery of structural elements governing stability of mammalian messenger RNAs

Hani Goodarzi, Hamed S. Najafabadi, Panos Oikonomou, Todd M. Greco, Lisa Fish, Reza Salavati, Ileana M. Cristea & Saeed Tavazoie

Nature | Vol 485 | 10 MAY 2012

Timo A. Ebeling
03.12.14

Freie Universität Berlin

Overview

- computational framework based on context-free grammars (CFGs) and mutual information (MI)
- de-novo motif discovery tool for finding informative structural elements in RNA
- Experimental validation of proposed algorithm

LETTER

doi:10.1038/nature11013

Systematic discovery of structural elements governing stability of mammalian messenger RNAs

Hani Goodarzi[1,2]†, Hamed S. Najafabadi[3,4]†, Panos Oikonomou[1,2]†, Todd M. Greco[2], Lisa Fish[5], Reza Salavati[3,4,6], Ileana M. Cristea[2] & Saeed Tavazoie[1,2]†

Decoding post-transcriptional regulatory programs in RNA is a critical step towards the larger goal of developing predictive dynamical models of cellular behaviour. Despite recent efforts[1–3], the vast landscape of RNA regulatory elements remains largely uncharacterized. A long-standing obstacle is the contribution of local RNA secondary structure to the definition of interaction

these *in silico* predictions reflect stable *in vivo* molecular conformations has not been fully explored[9]. In fact, the RNA binding proteins and complexes that interact with their target transcripts may facilitate the formation of secondary structures *in vivo*. Thus, we sought to bypass the need for predicting thermodynamically stable secondary structures by efficiently enumerating a large space of potential struc-

Overview

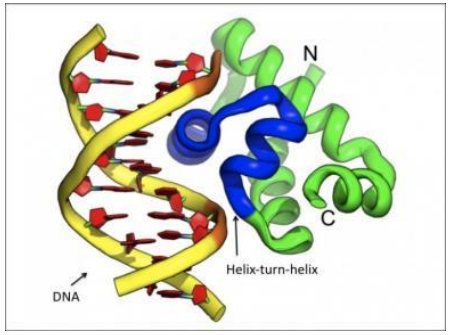
- Motivation
- Context-free grammars
- mRNA stability measurements
- Mutual information (MI)
- TEISER
- Hands-On: TEISER
- Experiment and Validation
- Summary

Nomenclature

- **Linear motif:**
  - short *protein sequences* mediating protein-protein interations
  - Sequence: VTLYDVAEYAGVSYQTVSRVVN
    4 25

Linear Helix-turn-helix motif in the lactose operon of ecoli

- **Structural motif:**
  - Short *segments of protein secondary structure*
  - e.g „helix-turn-helix"- or „zinc finger"- motif



Helix-turn-helix motif |
http://www.ebi.ac.uk/training/online/course/biomacromolecular-structures-introduction-ebi-reso/proteins/structural-motifs

Motivation

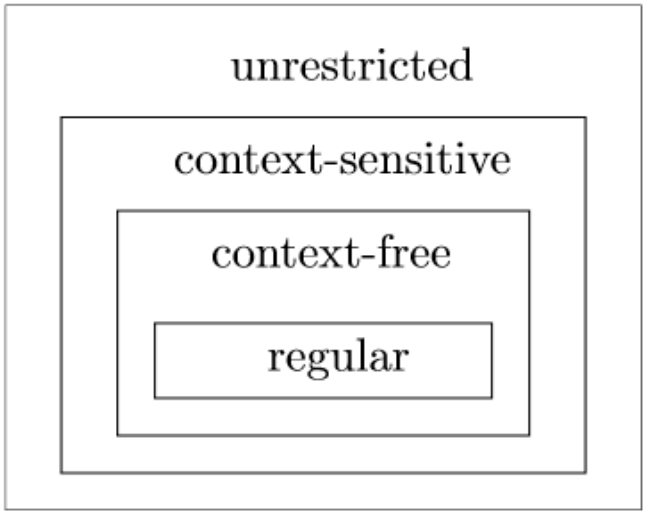
- Decoding of regulatory programs in RNA leads to models of cellular behaviors
- Presence of structural or regulatory element dictates alternative splicing patterns or affects other aspects of RNA biology
- Vast landscape of RNA regulatory elements still remains uncharacterized

Context-free grammars (CFGs)

- Formalization how all possible sentences can be enumerated in a (natural) language
- Generative grammars:
  - able to generate a string that belongs to a encoded language
- A grammar consists of:
  - A set of abstract non-terminal symbols , e.g {S}
  - A set of rewriting rules, e.g $S \rightarrow aS$, $S \rightarrow bS$, $S \rightarrow \emptyset$
  - A set of terminal symbols that appear in a word of the language e.g. {a, b}
- From left to right we replace S with a series of productions to generate a string e.g. $S \rightarrow aS \rightarrow abS \rightarrow abb$

Context-free grammars (CFGs)

- Can be used to model RNA sequences and their interactions
- Any production of the following form is allowed:
  W -> α
  α: String of terminal and non terminal symbols



Classes of grammars | https://www.mi.fu-berlin.de/wiki/pub/ABI/SS14Lecture11Materials/script.pdf

Definition:

A CFG is a 4-tupel C = (N, T, P, S)

s.t. N und T are alphabets with $N \cap T = \emptyset$

N is the nonterminal alphabet

T is the terminal alphabet

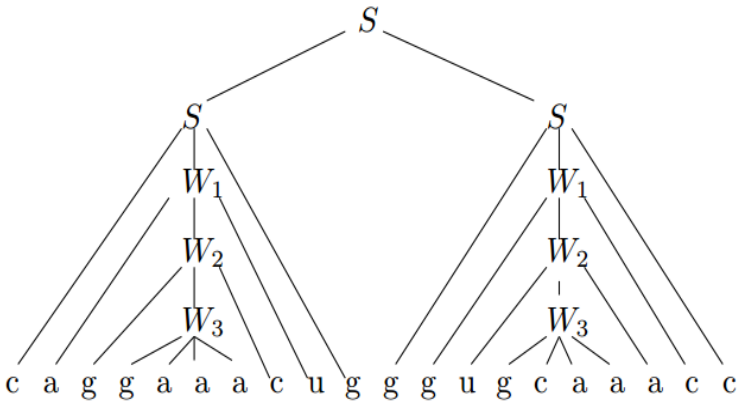
$S \in N$ is the start symbol

$P \subseteq N \times (N \cup T)^*$ is the finite set of all productions

Context-free grammars

- Consider a CFG to handle RNA hairpin loops:

$$
\begin{aligned}
&\mathrm{S} \rightarrow \mathrm{SS} \\
&\mathrm{S} \rightarrow aW_1u|cW_1g|gW_1c|uW_1a, \\
&W_1 \rightarrow aW_2u|cW_2g|gW_2c|uW_2a, \\
&W_2 \rightarrow aW_3u|cW_3g|gW_3c|uW_3a, \\
&W_3 \rightarrow gaaa|gcaa.
\end{aligned}
$$



Parse Tree of given grammar |
https://www.mi.fu-berlin.de/wiki/pub/ABI/SS14Lecture11Materials/script.pdf

- models a hairpin loops with 3 base pairs and a **gcaa** or **gaaa** loop

- Implementation: PDA

mRNA stability measurements

- Different approach than "free-energy" based methods
- Whole-genome mRNA stability measurements are performed to isolate stability from other aspects of mRNA behavior
- Used to identify cis-regulatory elements (linear and structural) that underlie transcript stability

mRNA stability measurements

- mRNA is tagged via biotinylation
  - enables discriminability
- new mRNA is synthesized
- Samples are taken after 0, 1, 2 and 4 hours
- RNA samples are labeled and hybridized to whole-genome human microarrays
- Rate at which the signal drops used as measure of decay rate (r):

$$r = -ln\frac{S_t}{S_0}/t$$

Mutual information (MI)

- Mutual information measures how much one random variables tells us about another
- E.g. multiple alignment with 2 columns
- 1. calculate for each column i of alignment, the frequency $f_i(x)$ of each base $x \in$ {A,C,G,T}
- 2. calculate the 16 joint frequencies $f_{ij}(x, y)$
- 3. calculate mutual information content H(i,j) in bits:

$$H_{ij} = \sum_{xy} f_{ij}(x, y) \cdot log_2 \frac{f_{ij}(x, y)}{f_i(x) \cdot f_j(y)}$$

TEISER

- TEISER (Tool for Eliciting Informative Structural Elements in RNA)
- Framework for identifying structural motifs that are informative of **whole-genome measurements** across all given transcripts
- Structural motifs are defined in terms of CFGs representing hairpin structures as well as primary sequence information
- MI is used to measure regulatory consequences of ~100 million different seed CFGs

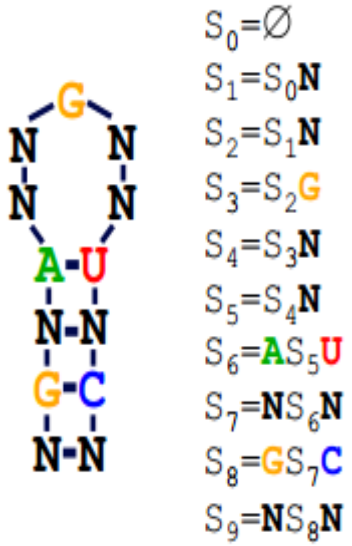
TEISER

1. Genome profile

- defined across the genes in the genome
  -> each gene is associated with a unique measurement
  -> obtained from experimental or computational sources

2. Structural motif definition

- each structural motif is defined as series of CFG statements that define sequence and structure



Structural motif discovery schematic

TEISER

3. Motif profile

- for every given motif a binary vector across all genes is created, which holds

  1: presence of that motif
  0: absence of that motif

TEISER

4. Creating seed CFGs

- A Set of CFG statements is used to represent all possible stem-loop structures that satisfy following criteria:

  Stem legth ranging from 4 bp – 7bp
  Loop length ranging from 4nt -9nt

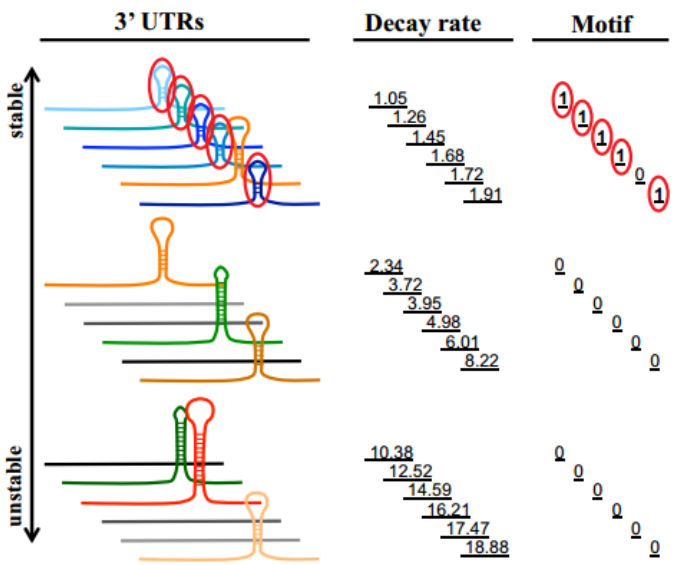
  Min 4 production rules and max 6 production rules representing non-degrading bases
  e.g. productions that are not: S->SN, S->NS, S-> NSN

  A information content of min 14 bits and max 20 bits

TEISER

5. Removing recently duplicated genes

- Duplicates that have similar values are removed



Structural motif discovery schematic

6. Calculating the mutual information values

- Mutual information (MI) is calculated between the *genome profile* and the *motif profile*

TEISER

7. Randomization-based statistical testing

- Genome profile is shuffled 1.5 million times and the corresponding MI values are calculated
- A motif is deemed significant if real MI value is greater than all of the randomly generated ones
- To minimize number of tests, structural motifs are first sorted based on MI values (high to low) and the statistical test is applied in order
- If 20 contiguous motifs in the sorted list fail the test, the procedure is terminated

TEISER

8. Optimization of the identified seeds into more informative motifs

- Initial collection of structural motifs is a raw sample of the entire solution space
- Providing a set of informative seeds which is optimized into closer representations of their actual form
- structural motifs that pass the previouse stage are further optimized and elongated

TEISER

8. Optimization of the identified seeds into more informative motifs

1. **Optimization**: select random CFG statements from the motif
   and convert the seq. information to all possible combinations of nt

   Evaluate all resulting struct. motifs and select the highest MI value

2. **Elongation**: production rules are added to the end of the CFG phrases,
   representing the motif

   Evaluate all resulting struct. motifs and select the highest MI value

TEISER

9. Detection of robust motifs

- Bootstrapping is performed to find robust motifs that are not overfitted

- For each predicted motif, 10 bootstraping steps are executed,
  - in each step $\frac{1}{3}$ of the genes are rendomly removed
  - MI value is calculated and statistical significance is evaluated

- A robustness score is defined as the number of steps in which the motif remains significant (Ranging from $\frac{0}{10}$ to $\frac{10}{10}$)

TEISER

10. Final statistical tests

- Returns motifs which are enriched at one end of the data range or the other
- Calculation of Spearman correlation between enrichment scores and the average data
- Threshould is set to 0.01 -> FDR: 10 %

TEISER

11. Inter-species conservation

- A conservation score is calucated for each motif, based on its conservation with respect to a related genome
- Orthologous transcripts in both genomes are scanned for the presence/absence of the motif
- Overlap is used in a hypergeometric test
- Conservation score is defined as 1-p, ranging from 0 to 1 (1 beeing highly conserved between two genomes)

TEISER

12. Predicting functional interactions

- Given 2 motifs:
  - functional interactions are assessible by measuring how informative the presense of one would be without the other

  - MI values are calculated for pairwise motif profiles of structural or linear motifs to detect interactions

  - Randomization-based statistical tests are applied to find the significant ones

TEISER

False- discovery rate

- To asses FDR, 30 runs with shuffled 5‘ and 3‘ UTR seq. are performed

- In all runs, not a single motif passed all statistical tests

  number of false positives in each run, on average, is smaller than $\frac{1}{30}$
  Corresponding to a FDR of <0.01

Hands-On: TEISER

TEISER (for Tool for Eliciting Informative Structural Elements in RNA)

- Download from https://tavazoielab.c2b2.columbia.edu/TEISER/
- Version 1.0
- Install via make
- Implemented in perl and C/C++

Hands-On: TEISER

TEISER (for Tool for Eliciting Informative Structural Elements in RNA)

- Initializing the structural seeds

  $TEISERDIR/Programs/seed_creator -min_stem_length 4 -max_stem_length 7 -min_loop_length 4 -max_loop_length 9 -min_inf_seq 4 -max_inf_seq 6 -max_inf 20 -min_inf 14 -outfile seeds.4-7.4-9.4-6.14-20

- Generates ~ 2.3 GB of seed data
- Runtime ~ 3 h

Hands-On: TEISER

TEISER (for Tool for Eliciting Informative Structural Elements in RNA)

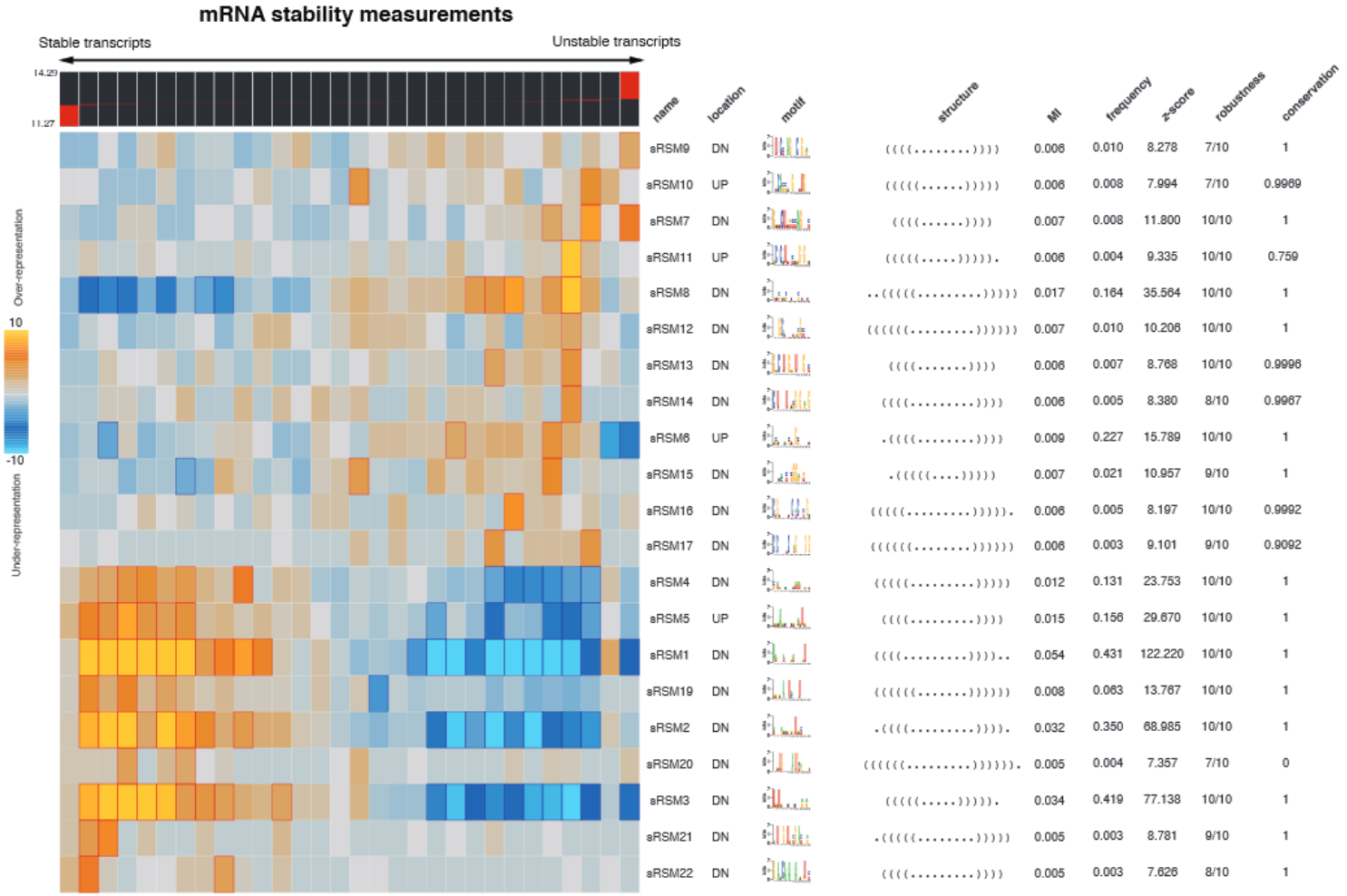
- Using TEISER

perl teiser_parallel.pl --expfile=<inp> --species=<sp> --exptype=<type>
--ebins=<int> --submit=<0/1>

perl teiser_parallel.pl --expfile=avg --species=human --exptype=continuous
--ebins=30 --submit=0

Hands-On: TEISER

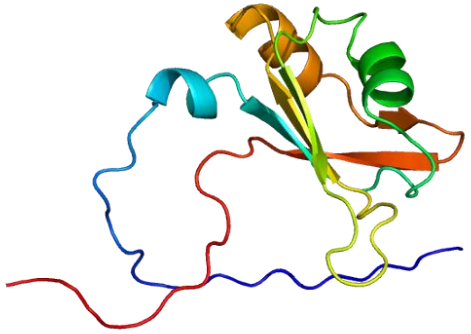
TEISER (for Tool for Eliciting Informative Structural Elements in RNA)



name	location	motif	structure	MI	frequency	z-score	robustness	conservation
sRSM9	DN		((((.....))))	0.006	0.010	8.278	7/10	1
sRSM10	UP		((((((.....))))))	0.006	0.008	7.994	7/10	0.9969
sRSM7	DN		((((.....))))	0.007	0.008	11.800	10/10	1
sRSM11	UP		((((((.....)))))).	0.006	0.004	9.335	10/10	0.759
sRSM8	DN		..(((((.....)))))	0.017	0.164	35.564	10/10	1
sRSM12	DN		((((((.....))))))	0.007	0.010	10.206	10/10	1
sRSM13	DN		((((.....))))	0.006	0.007	8.768	10/10	0.9998
sRSM14	DN		((((.....))))	0.006	0.005	8.380	8/10	0.9967
sRSM6	UP		.((((.....))))	0.009	0.227	15.789	10/10	1
sRSM15	DN		.(((((....)))))	0.007	0.021	10.957	9/10	1
sRSM16	DN		((((.....)))))).	0.006	0.005	8.197	10/10	0.9992
sRSM17	DN		((((((.....))))))	0.006	0.003	9.101	9/10	0.9092
sRSM4	DN		((((.....))))	0.012	0.131	23.753	10/10	1
sRSM5	UP		((((.....))))	0.015	0.156	29.670	10/10	1
sRSM1	DN		((((.....)))))..	0.054	0.431	122.220	10/10	1
sRSM19	DN		((((((.....))))))	0.008	0.063	13.767	10/10	1
sRSM2	DN		.((((.....)))).	0.032	0.350	68.985	10/10	1
sRSM20	DN		((((((.....)))))).	0.005	0.004	7.357	7/10	0
sRSM3	DN		((((.....)))).	0.034	0.419	77.138	10/10	1
sRSM21	DN		.(((((.....)))))	0.005	0.003	8.781	9/10	1
sRSM22	DN		((((.....))))	0.005	0.003	7.626	8/10	1

Experiment and Validation

- Detection of structural elements in mRNA in a genome-wide manner
- 8 highly significant elements (with structural information) are identified
  - strongest -> major role in global mRNA regulation
- Validation via biochemistry, mass-spectrometry, biochemistry and in-vivo binding studies
- HNRPA2B1 is identified to act as key regulator
  - binds this element
  - stabilizing target genes of this element



Heterogeneous nuclear ribonucleoprotein A2/B1

Summary

- Promising approach to decode post-transcriptional regulatory programs in RNA
- Can be used for whole genome experiments
- Predictions based rather on MI and stability measurements than on „free-energy“ – approaches
- Proposed method allows a de-novo motif discovery in RNA

Thank you for your attention !

Questions ?