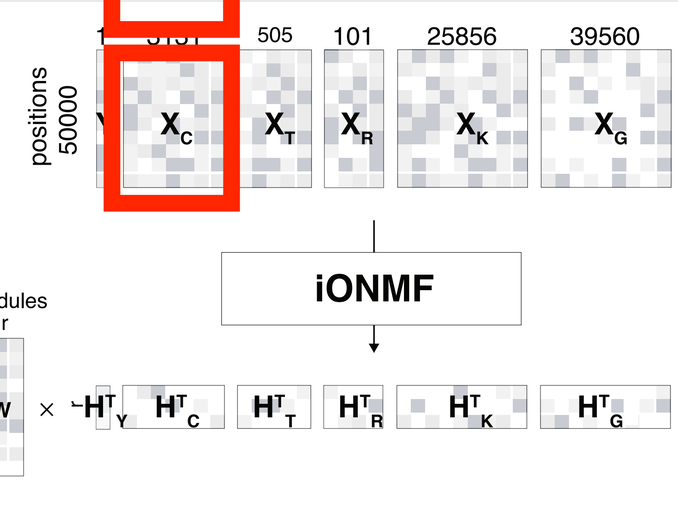
Orthogonal nonnegative factorization-based analysis of nineteen protein-RNA interaction CLIP data sets

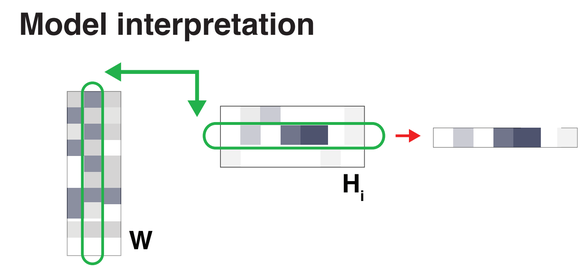
Martin Stražar[1], Marinka Žitnik[1], Blaž Zupan[1], Jernej Ule[1,2] and Tomaž Curk[1]

[1]Bioinformatics Laboratory, University of Ljubljana, Faculty of computer and information science, Večna pot 113, 1000 Ljubljana
[2]Department of Molecular Neuroscience, UCL Institute of Neurology, Queen Square, London, WC1N 3BG, UK

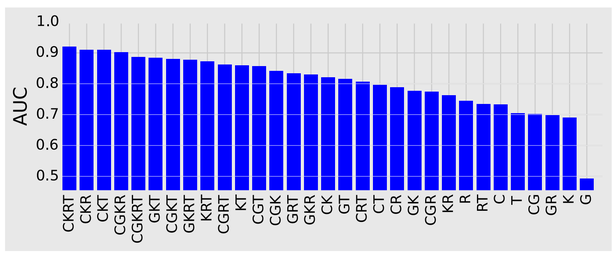
Abstract

RNA binding proteins (RBPs) play a major role in post-transcriptional processes: splicing, transport and polyadenylation[1]. Computational modelling of protein-RNA interactions depends on precise structural knowledge, which is often difficult to obtain[2]. We propose an approach to model protein-RNA interactions using publicly available data. We combine data sources to achieve accurate prediction and a comprehensive model based on integrative, orthogonal nonnegative matrix factorization (iONMF).





Combinations of data sources improve prediction



Data sources are matrices describing features in the interval [-50, 50] nt positions around crosslink sites.

Ranking of individual data sources:

R: RNA structure, probability of double-stranded RNA, which affects nucleotide accessibility.

C: co-binding; combinatorial protein-RNA interactions, which compete or cooperate for the same RNA.

T: region type, **K:** RNA k-mers, **G:** Gene Ontology.

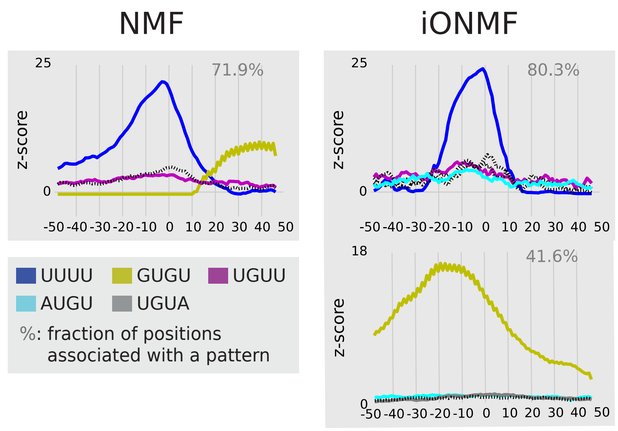
$$J = \sum_{i=1}^{N} (\|\mathbf{X}_i - \mathbf{W}\mathbf{H}_i^T\|_2^2 + \alpha\|\mathbf{H}_i^T\mathbf{H}_i - \mathbf{I}\|_2^2) \qquad \text{loss function}$$

iONMF: integrative orthogonal nonnegative matrix factorization

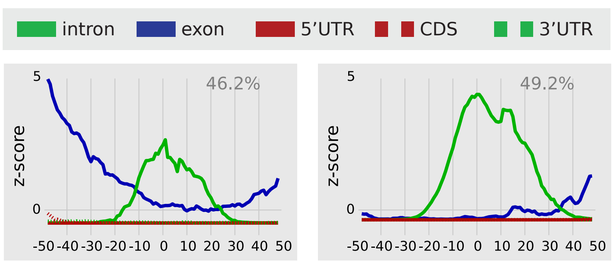
The original data matrices $\mathbf{X}_i$ are approximated **simultaneously** as a product of:

- Common matrix **W**: similarity of positions based on all data sources.
- Matrices $\mathbf{H}_i$: commonly occuring data source-specific patterns.

Orthogonality constraints induce non-overlapping patterns

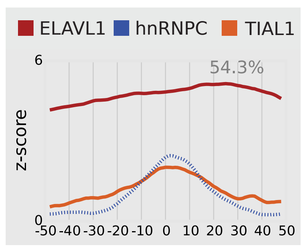


With NMF, UG-repeats associated with TDP-43 are occluded by U-rich patterns.
iONMF successfully separates the two signals.



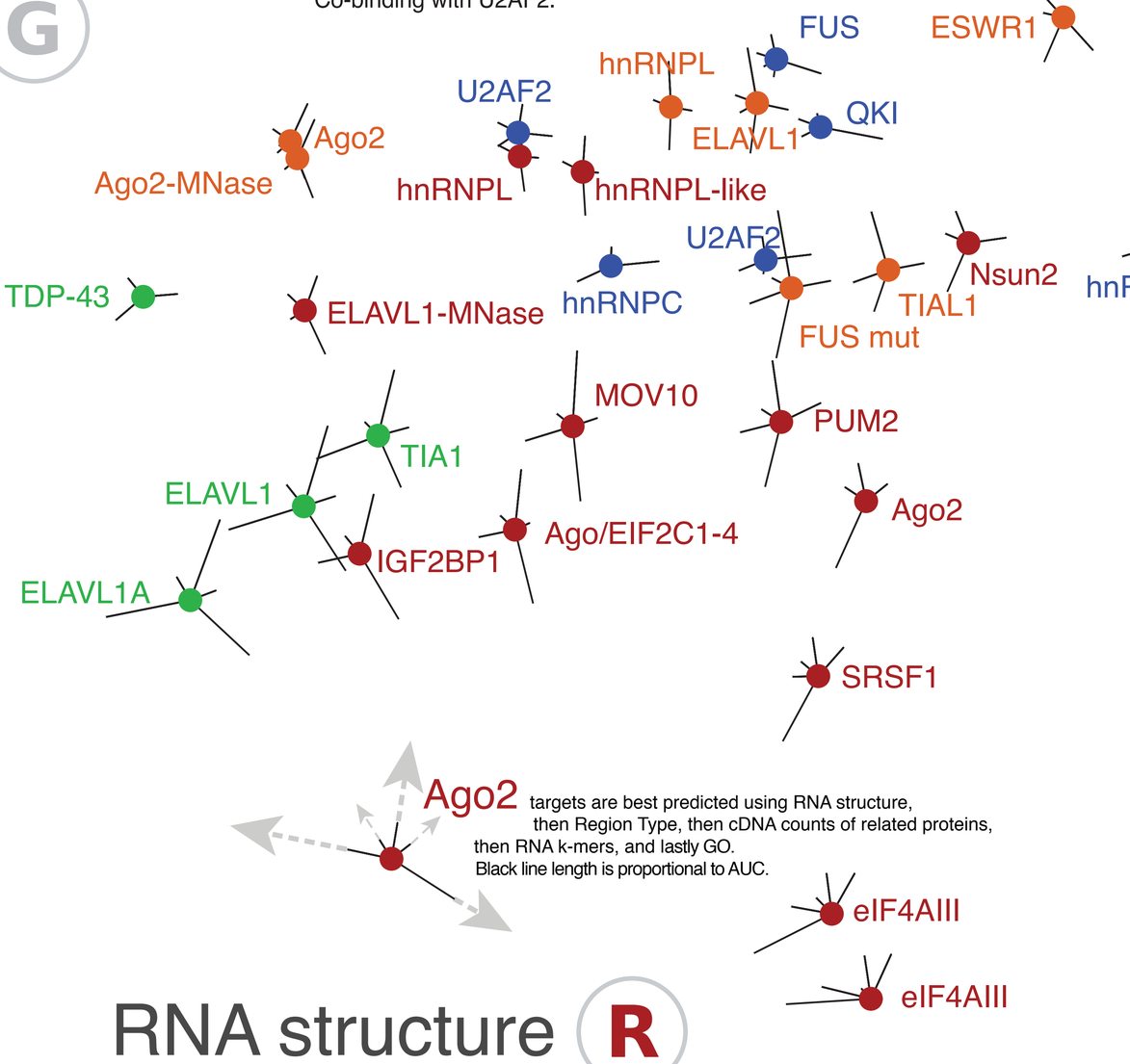
Positions relative to crosslink sites of Ago2 (left) and Ago/EIF2C1-4 (right) and frequency of the five region types. All experiments on Argonaute proteins report binding to short-introns. This agrees with known Ago binding to miRNA[3].

Gene Ontology G



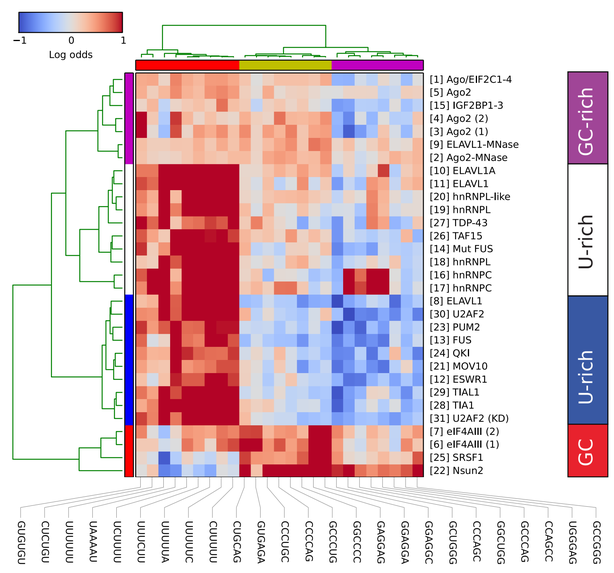
Co-binding with U2AF2.

Co-binding C



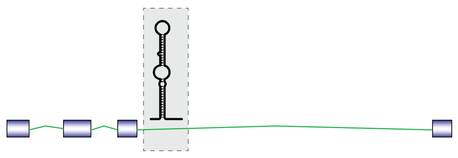
Ago2 targets are best predicted using RNA structure, then Region Type, then cDNA counts of related proteins, then RNA k-mers, and lastly GO.
Black line length is proportional to AUC.

RNA k-mers K



Splicing regulators bind U-rich (intronic) motifs.

Region type T



Example of an Ago2 target: intronic miRNA MIR7-3.

RNA structure R

Splicing regulators (hnRNPs, U2AF2, ELAVL1, FUS, QKI) bind intronic regions.

Clustering based on affinity for double stranded RNA.

Double stranded RNA

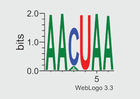
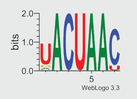
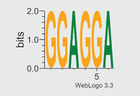
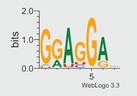
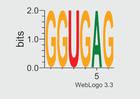
ELAVL1 (HuR) binds **unstructured** (s. strand. RNA)[5].

Argonaute binding sites are known to contain and unstructured region **upstream** of the crosslink sites[6].

hnRNPC binding site hnRNPC motifs
CAPS regulated exon

```
ttttctttctttttttttttattttattttttttttgagacagGTTCTCGTCCTGTCA
CCCAGGCTGGTATGCGGTGGTGTGATCGTAGCTCACTGCAGTCTCGAACTCCTGGGTTCAA
GCGATCCTTCCACTTCAGCCTCCCAAGTAGCTGGTACTACAGgtgtgtgccacgacacccg
gctaagttttgaaatttatttttttgtagagacaggattttcctatgttgcccaggctggt
tttcaaactcctggcc
```

Common motifs at antisense Alu elements (GGCUGG, GCCCAG, CCUGCC, GCCGGG) are associated with hnRNPC.

Protein	iONMF	RNACompete
QKI		
SRSF1		
FUS		

+ 12 more

Discovered motifs are in accordance with *in vitro* assay RNAcompete[8].

References

Data was obtained from servers iCount (http://icount.biolab.si) and DoRiNA (dorina.mdc-berlin.de).

[1] König et. al, Nature review Genetics (2013)
[2] Puton et al., Journal of structural biology (2013)
[3] G. Hutvagner and M. J. Simard, Nat. Rev. Mol. Cell Biol., Jan. 2008.
[4] I.L. Hofacker et.al Monatshefte f. Chemie (1994).
[5] H. Kazan and Q. Morris, PLoS Comput. Biol, Jan. 2010.
[6] Li et al., Genome Res. (2014.)
[7] D. Ray, T. R. Hughes, et al. Nature, Jul. 2013.

This work was supported by grants from the Slovenian Research Agency (P2-0209, J7-5460) and the European Research Council (206726-CLIP).