Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 2 April 2015
Applications Note

OXFORD

RNA Biology

Supplementary Methods and Discussion

Alexander R. Gawronski [1,*], Michael Uhl [3], S. Cenk Sahinalp [2,4,*], Rolf Backofen [3,*]

[1] Computing Science, Simon Fraser University, Burnaby, Canada
[2] Vancouver Prostate Centre, Vancouver, BC, Canada
[3] Institut für Informatik, University of Freiburg, Freiburg im Breisgau, Germany
[4] Department of Computer Science, Indiana University, Bloomington, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

1 Methodological Details of IntaRNA2

The next stage is the prediction of RNA-RNA interactions using IntaRNA2 (Mann *et al.*, 2017) with the modifications outlined above. IntaRNA is a popular accessibility-based tool known for its highly competitive performance (Lai and Meyer, 2016). The hybridization calculation follows that of RNAHybrid (Rehmsmeier *et al.*, 2004) with a time and space complexity of $O(nm)$. The accessibility is calculated in $O(nL^2)$ using RNAplfold (Bernhart *et al.*, 2006), an algorithm that computes accessibility in a locally folded region of length L. Both energy contributions are calculated for every combination of intervals on both sequences requiring a time and space complexity of $O(n^2m^2)$. Using the same restriction on interaction length w as RNAup (Muckstein *et al.*, 2006), the time and space complexity is $O(nmw^2)$. By using sparsification (Figure 1), this complexity is further reduced to $O(nm)$ space and $O(n\bar{m})$ time where $\bar{m} = max(m, L^3)$.

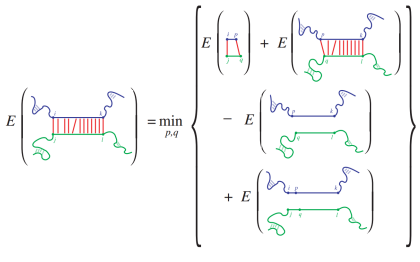

Fig. 1. Heuristic for reducing time complexity of IntaRNA (figure taken from (Busch et al., 2008)). The top energies are of the hybridization and the two bottom energies are for the accessibilities. The accessibilities are not additive so the contribution needs to be subtracted and then added back with the extended region.

2 P-value Computation for Predicted Interactions

2.1 RNA-RNA Interactions

In a previous work, RNA-RNA interaction energies were fitted to a generalized extreme value (GEV) distribution in order to compute interaction p-values (Wright *et al.*, 2014). From our recent experience we found that a gamma distribution fits the data better(data not shown), so it was used for all experiments. Regardless, we support a CopraRNA-style GEV approach through a user-specified parameter. We first compute a background gamma cumulative distribution function (CDF), which has two parameters: shape (α) and rate (β) (Equation 1- 3). The background values are obtained by assuming that a top percent (default 3%) are true interactions and the rest are background. The parameters of the function are estimated using maximum-likelihood fitting. This is done using the "fit" function in the python "stats" package from the scipy library (Jones *et al.*, 01). With these estimated α and β parameters, the p-values for each energy value (x) can be computed using the survival function ($1 - cdf(x)$).

$$F(x;\alpha,\beta) = \int_0^x f(u;\alpha,\beta)du \tag{1}$$

$$f(x;\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1}e^{-x\beta}}{\Gamma(\alpha)} \tag{2}$$

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx \tag{3}$$

2.2 RNA-Protien Interactions

P-values for each peak score were calculated based on position-wise score data from 5000 randomly selected transcripts, using R's empirical cumulative distribution function (ECDF). The function returns the p-value of a given score based on the constructed ECDF and the ecdf() object can be stored on disk for subsequent recalculations (found together with models in Supplementary file 1). We chose this non-parametric approach since the scores did not show a clear unimodal distribution for most models,

Table 1. Used GraphProt models with model parameters, training set information and filter p-values. GraphProt model parameters are: epochs, lambda, R, D, bitsize, abstraction. PMID: data source pubmed ID, method: CLIP-seq protocol, filter_p: p-value used for filtering predicted sites, pos_tr: number of positive training sites, neg_tr: number of negative training sites

RBP	PMID	method	model_type	filter_p	epochs	lambda	R	D	bitsize	abstraction	pos_tr	neg_tr	ROC	APR
AGO1-4	20371350	PARCLIP	structure	0.02376089	20	0.000001	4	1	14	3	36802	31310	0.85584	0.86766
ELAVL1	21723170	PARCLIP	sequence	0.001640548	10	0.001	3	5	14	-	7747	7750	0.92887	0.94365
EWSR1	20371350	PARCLIP	sequence	0.005115178	50	0.001	1	2	14	-	16292	14720	0.94345	0.9496
FMR1	27018577	eCLIP	structure	0.04819012	40	0.0001	4	5	14	3	2587	2587	0.88109	0.87115
FUS	22081015	PARCLIP	sequence	0.003591709	40	0.0001	1	1	14	-	34581	31480	0.96988	0.97034
HNRNPC	27018577	eCLIP	sequence	0.0006383588	50	0.001	3	6	14	-	2511	2511	0.95636	0.95178
HNRNPK	27018577	eCLIP	sequence	0.0011904	10	0.001	2	1	14	-	2674	2673	0.9823	0.98059
IGF2BP1-3	20371350	PARCLIP	structure	0.01519445	50	0.0001	4	0	14	3	8539	6838	0.88223	0.89533
KHDRBS1	27018577	eCLIP	structure	0.003200621	40	0.001	3	2	14	3	2552	2552	0.9234	0.92122
MOV10	22844102	PARCLIP	sequence	0.02331425	20	0.001	4	2	14	-	13793	12987	0.79824	0.7715
PUM2	27018577	PARCLIP	sequence	0.002040983	40	0.001	4	4	14	-	9116	8227	0.94144	0.95158
QKI	27018577	eCLIP	structure	0.0006862552	40	0.000001	4	2	14	3	2650	2650	0.94722	0.95187
SND1	27018577	eCLIP	structure	0.04999487	50	0.0001	3	4	14	3	2413	2413	0.89622	0.88589
TAF15	22081015	PARCLIP	sequence	0.003209317	50	0.001	3	2	14	-	7298	6606	0.96794	0.964
TARDBP	27018577	eCLIP	sequence	0.0003341065	30	0.001	4	5	14	-	2752	2752	0.98524	0.98712
TIA1	27018577	eCLIP	sequence	0.009658455	30	0.001	2	5	14	-	3073	3073	0.84148	0.86061
TNRC6A	27018577	eCLIP	structure	0.04627634	50	0.001	3	0	14	3	2653	2653	0.83569	0.85761

which prevented the use of conventional fitting procedures for unimodal distributions. For each model, we then calculated the top position-wise score of each positive training site to construct a second ECDF. To get a threshold for filtering the peak score p-values, the score at 50 % of the distribution was taken and inserted into the first ECDF to get its p-value. This way we obtain an individual p-value threshold for each RBP model, allowing us to select binding sites with scores comparable to the scores found in the respective positive training sites. The obtained filter p-values for each model can be found in Supplementary Table 1.

3 Challenges and Limitations

Predicting combined interactions between lncRNAs, RBPs and target RNAs on a transcriptome-wide scale is an inherently difficult task, due to several reasons: firstly, the limited number of known lncRNA mechanism cases makes it difficult to tune the model. Specifically, the selection of various parameters in terms of distances between interactions and various cutoffs becomes nearly *ad hoc*. Moreover, it is unknown to what extent the studied cases occur in the cell or whether they are typical representatives of a certain class of interactions. Secondly, even with the careful filtering applied in this work, RNA-RNA and RNA-protein predictions are fairly non-specific. With thousands of predicted targets, it is likely that many are false positives. Given that only the most significant interaction combinations are included, it is difficult to determine which are true predictions since they are all plausible. Despite these difficulties, the presented work provides a solid starting point for further experimental investigation.

One way to improve the current approach would be the development of more realistic interaction models. As for the RBP-target prediction, information on RBP affinities for a range of target RNAs as well as the relative importance of target sequence, structure and context should help to design more accurate models. So far, detailed affinity distributions have only been reported for the *E. coli* C6 protein, utilizing the high-throughput sequencing kinetics (HiTS-KIN) protocol (Lin *et al.*, 2016). Lately, a more simple affinity approach was combined with estimating the sequential and structural binding properties of 78 human RBPs, using an RNA Bind-n-Seq variant with 5 different protein concentrations (Dominguez *et al.*, 2017). In order to improve prediction specificity, it is also possible to use CLIP data to cluster RBPs with common binding sites and to learn properties from these sites, as shown by Li et al. (Li *et al.*, 2017). As for the lncRNA-target prediction, integrating protein binding information directly into the RNA-RNA interaction calculation might lead to the prediction of more realistic hybrids. Moreover, incorporating RNA structure probing data of the involved RNAs, e.g. determined by selective 2-hydroxyl acylation and profiling (SHAPE), could improve the hybrid prediction. As the number of studied lncRNA mechanisms gradually increases, machine learning approaches could further help to improve model performance by learning optimal parameter combinations from the data.

Another more immediate extension of this work would be the incorporation of additional data, such as new RBP predictions or miRNA interaction information. It is conceivable to assume that lncRNAs might block or sequester miRNAs, just as they do RBPs. Inclusion of miRNA target sites would therefore broaden the scope of mechanisms MechRNA can predict. The modular nature of MechRNA makes such extensions possible, which might open exciting new avenues for lncRNA research.

References

Bernhart, S. H., Hofacker, I. L., and Stadler, P. F. (2006). Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**(5), 614–615.

Busch, A., Richter, A. S., and Backofen, R. (2008). IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**(24), 2849–2856.

Dominguez, D., *et al.* (2017). Sequence, structure and context preferences of human rna binding proteins. *bioRxiv*, page 201996.

Jones, E., *et al.* (2001–). SciPy: Open source scientific tools for Python. [Online; accessed <today>].

Lai, D. and Meyer, I. M. (2016). A comprehensive comparison of general rna–rna interaction prediction methods. *Nucleic acids research*, **44**(7), e61–e61.

Li, Y. E., *et al.* (2017). Identification of high-confidence rna regulatory elements by combinatorial classification of rna–protein binding sites. *Genome biology*, **18**(1), 169.

Lin, H.-C., *et al.* (2016). Analysis of the rna binding specificity landscape of c5 protein reveals structure and sequence preferences that direct rnase p specificity. *Cell chemical biology*, **23**(10), 1271–1281.

Mann, M., Wright, P. R., and Backofen, R. (2017). IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Res.*

Muckstein, U., *et al.* (2006). Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**(10), 1177–1182.

Rehmsmeier, M., *et al.* (2004). Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**(10), 1507–1517.

Wright, P. R., *et al.* (2014). CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains. *Nucleic Acids Res.*, **42**(Web Server issue), W119–123.

Table 2. Selected LncRNAs for MechRNA analysis. The lncRNAs vary in terms of what is known about their mechanisms, allowing MechRNA to be tested with various amounts of a priori data. PCAT1 has a question mark indicating that competitive binding is the hypothesis not been validated yet.

	Protein Binding		**RNA-RNA Interaction**				
TP53 Transcript	**HuR S**	**HuR E**	**TP53 S**	**TP53 E**	**7SL S**	**7SL E**	**FE**
ENST00000618944	1950	1971	1980	2022	256	298	-51.563
ENST00000504937	1817	1838	1847	1889	256	298	-51.563
ENST00000445888	2071	2092	2101	2143	256	298	-51.563
ENST00000420246	2201	2222	2231	2273	256	298	-51.563
ENST00000269305	**2125**	**2146**	**2155**	**2197**	**256**	**298**	**-51.563**
ENST00000610292	2185	2206	2215	2257	256	298	-51.563
ENST00000620739	2125	2146	2155	2197	256	298	-51.563
ENST00000455263	2128	2149	2158	2200	256	298	-51.563
ENST00000610623	1877	1898	1907	1949	256	298	-51.563
ENST00000504290	1877	1898	1907	1949	256	298	-51.563
ENST00000610538	2128	2149	2158	2200	256	298	-51.563
ENST00000619485	2071	2092	2101	2143	256	298	-51.563
ENST00000510385	1950	1971	1980	2022	256	298	-51.563
ENST00000622645	2201	2222	2231	2273	256	298	-51.563
ENST00000619186	1817	1838	1847	1889	256	298	-51.563
ENST00000617185	2270	2291	2300	2342	256	298	-51.563