SMARTIV: a novel method for RNA sequence and structure motif discovery from in-vivo binding data

Maya Polishchuk, Inbal Paz, Yael Mandel-Gutfreund

RNA binding proteins (RBPs) are essential for cell processes. Many RBPs recognize specific RNA binding sites. These RNA binding sites are usually characterized by specific short sequences (frequently 4-8 nucleotides) known as binding motifs. In addition to primary RNA sequence, the structure of the RNA target is known to play a major role in guiding RBP-RNA recognition. It is well established that most RBPs prefer to bind their targets at single stranded regions [1]. However, more and more proteins are discovered to bind specifically to double stranded RNA [2,3,4].

In recent years, high-throughput binding techniques have been developed to identify the binding preferences of RBPs. These technologies can be roughly divided into methods that measure protein-RNA binding in vitro, based on High Throughput Systematic Evolution of Ligands by Exponential Enrichment (HT-SELEX) [5,6], and in vivo binding experiments, based on CrossLinking and ImmunoPrecipitation (CLIP), such as HITS-CLIP [7], PAR-CLIP [8], iCLIP [9], eCLIP and irCLIP [10]. These methods in combination with dedicated bioinformatics analysis tools generate a set of relatively short bound sequences (usually from several dozen nucleotides long) and quantify their binding signals [11]. However, these methods do not provide information regarding the structural preferences of the protein. Methods to obtain the secondary structure of RNA (mostly predictive) are becoming available (e.g. RNAsubopt [12], Sfold [13], RNAshapes [14]).

As more and more data is accumulating in the public databases from CLIP-seq binding experiments (e.g. DoRiNA [15], CLIPdb [16], ENCODE [17]), and while methods to obtain the secondary structure of RNA are becoming available, inferring both sequence and structure preferences of RBPs remains a challenge, and there is a strong need for computational tools that can be used for this task.

Here we present a novel method, named SMARTIV (Sequence and Structure Motif Enrichment Analysis for Ranked RNA daTa generated from In-Vivo binding experiments), for extracting enriched motifs from in vivo high-throughput RNA binding data combining sequence and secondary structure information [18,19,20].

SMARTIV uses ranked numerical binding scores obtained from CLIP results and predicted secondary structure of the sequences to generate motifs concisely represented in a graphical logo in a combined sequence and structure 8-letter alphabet A, C, G, U, a, c, g, u (upper case for unpaired and lower case for paired nucleotides).

The SMARTIV algorithm has several advantages over existing methods for extracting sequence and structure motifs from high-throughput RNA binding data. SMARTIV represents sequence and structure of the motifs in a highly informative and easy for visual perception single logo and is available both as a parameter-driven program and as a user-friendly web-server (http://smartiv.technion.ac.il). The method doesn't require splitting the input data artificially into bound and unbound datasets. SMARTIV provides data-driven p-value assessment for the detected motifs (represented as logo and corresponding PWM). Notably, the method extracts overrepresented secondary structures as structural motifs instead of adding structural information to enriched sequence motifs. Finally, SMARTIV is extremely fast. On average, we process one CLIP dataset in approximately 3–4 min on an Intel Core i7-2600 CPU @ 3.40 Ghz * 4 and 32 Gb memory.

We tested the method on CLIP-seq data for a variety of RBPs and show that our results are highly consistent with previously known sequence and structure binding preferences of the proteins.

We are grateful to Professor Zohar Yakhini for the advice on statistical estimation, Refael Kohen for the initial stage of approach development, Rona Mesika for help in data sources selection, Limor Leibovich and Leon Anavy for source code of DRIMust algorithm, and for financial support of Lady Davis Fellowship Trust (granted to Maya Polishchuk)

1. X.C. Li, H. Kazan, H.D. Lipshitz, Q.D. Morris (2014) Finding the target sites of RNA- binding proteins, Wiley Interdiscip. Rev. RNA 5 (1) 111–130.

2. A.Ramos, S.Grunert, J.Adams, D.R.Micklem, M.R.Proctor, S.Freund, M. Bycroft, D. St Johnston, G. Varani (2000) RNA recognition by a Staufen double- stranded RNA-binding domain, EMBO J. 19 (5) 997–1009.

3. G. Masliah, P. Barraud, F.H. Allain (2013) RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. Cell. Mol. Life Sci. 70 (11) 1875–1895.

4. T. Aviv, Z. Lin, G. Ben-Ari, C.A. Smibert, F. Sicheri (2006) Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p, Nat. Struct. Mol. Biol. 13 (2) 168–176.

5. D. Ray, H. Kazan, E.T. Chan, L. Pena Castillo, S. Chaudhry, S. Talukder, B.J. Blencowe, Q. Morris, T.R. Hughes (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins, Nat. Biotechnol. 27 (7) 667–670.

6. D. Ray, H. Kazan, K.B. Cook, M.T. Weirauch, H.S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L.H. Matzat, R.K. Dale, S.A. Smith, C. Yarosh, S.M. Kelly, B. Nabet, D. Mecenas, W. Li, R.S. Laishram, M. Qiao, H.D. Lipshitz, F. Piano, A.H. Corbett, R.P. Carstens, B.J. Frey, R.A. Anderson, K.W. Lynch, L.O. Penalva, E.P. Lei, A.G. Fraser, B.J. Blencowe, Q.D. Morris, T.R. Hughes (2013) A compendium of RNA-binding motifs for decoding gene regulation, Nature 499 (7457) 172–177.

7. D.D. Licatalosi, A. Mele, J.J. Fak, J. Ule, M. Kayikci, S.W. Chi, T.A. Clark, A.C. Schweitzer, J.E. Blume, X. Wang, J.C. Darnell, R.B. Darnell (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing, Nature 456 (7221) 464–469.

8. M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano Jr., A.C. Jungkamp, M. Munschauer, A. Ulrich, G.S. Wardle, S. Dewell, M. Zavolan, T. Tuschl (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP, Cell 141 (1) 129–141.

9. J. Konig, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D.J. Turner, N.M. Luscombe, J. Ule (2010) ICLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution, Nat. Struct. Mol. Biol. 17 (7) 909–915.

10. E.U. Van Nostrand, G.U. Pratt, A.A. Shishkin, C.U. Gelboin-Burkhart, M.U. Fang, B.U. Sundararaman, S.U. Blue, T.U. Nguyen, C. Surka, K.U. Elkins, R.U. Stanton, F. Rigo, M. Guttman, G.U. Yeo (2016) Robust transcriptome-wide discovery of RNA- binding protein binding sites with enhanced CLIP (eCLIP), Nat. Methods 13 (6) 508–514.

11. M.G. Uhl, T.G. Houwaart, G.I. Corrado, P.R.G. Wright, R. Backofen (2017) Computational analysis of CLIP-seq data, Methods, pii: S1046-2023(17)30082-8. doi: 10.1016/j.ymeth.2017.02.006. [Epub ahead of print].

12. Wuchty S, Fontana W, Hofacker IL, Schuster P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers. Feb, 49(2):145-65.

13. Ding Y., Lawrence C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res., 31:7280–7301.

14. Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics., 22:500-503.

15. K. Blin, C. Dieterich, R. Wurmus, N. Rajewsky, M. Landthaler, A. Akalin (2015) DoRiNA 2.0-upgrading the doRiNA database of RNA interactions in post-transcriptional regulation, Nucleic Acids Res. 43 (Database issue) D160–D167.

16. Y.C. Yang, C. Di, B. Hu, M. Zhou, Y. Liu, N. Song, Y. Li, J. Umetsu, Z.J. Lu (2015) CLIPdb: a CLIP-seq database for protein-RNA interactions, BMC Genomics 16 51.

17. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature. Sep 6, 489(7414):57-74.

18. M.Polishchuk, I.Paz, R.Kohen, R.Mesika, Z.Yakhini, Y.Mandel-Gutfreund. (2017) A combined sequence and structure based method for discovering enriched motifs in RNA from in vivo binding data, Methods, pii: S1046-2023(17)30100-7. doi: 10.1016/j.ymeth.2017.03.003. [Epub ahead of print].

19. L.Leibovich, I.Paz, Z.Yakhini, Y.Mandel-Gutfreund. (2013) DRIMust: A Web-Server for Discovering Rank InbalancesMotifs Using Suffix Trees, Nucleic Acid Res. 41, W174-W179.

20. D.Cohn-Alperovich, A.Rabner, I.Kifer, Y.Mandel-Gutfreund, Z.Yakhini. (2016) Mutual

Enrichment in Aggregated Ranked Lists with Application to Gene Expression Requlation, Bioinformatics 32(17), i464-i472