EMBnet.journal
Bioinformatics in Action

Korfiati et al., EMBnet.journal 2017, 23, e882
dx.doi.org/10.14806/ej.23.0.882

Technical Notes

# InSyBio ncRNASeq: a Web tool for analysing non-coding RNAs

**Aigli Korfiati[1]✉, Konstantinos Theofilatos[1], Christos Alexakos[1], Seferina Mavroudi[1,2]**

[1] InSyBio Ltd, United Kingdom
[2] UK & Department of Social Work, School of Sciences of Health and Care, Technological Educational Institute of Western Greece, Koukouli, Patra, Greece

Competing interests: AK, KT, CA: employer of InSyBio Ltd. InSyBio Ltd participates in the NBG Business Seeds program by the National Bank of Greece and financed this study. However, InSyBio and NBG Business Seeds program were not involved in the study design, collection, analysis & interpretation of data and writing of the paper. They only participated in the process of selecting the suitable journal for submitting the paper. SM: external collaborator of InSyBio Ltd and an employer in the Department of Social Work, School of Sciences of Health and Care, Technological Educational Institute of Western Greece. InSyBio Ltd participates in the NBG Business Seeds program by the National Bank of Greece and financed this study. However, InSyBio and NBG Business Seeds program were not involved in the study design, collection, analysis & interpretation of data and writing of the paper. They only participated in the process of selecting the suitable journal for submitting the paper.

**Abstract**

Non-coding RNA (ncRNA) genes encode non-protein-coding RNAs, which are classified into infrastructural and regulatory ncRNAs, depending on their role. Regulatory ncRNAs are involved in a variety of key cellular processes, and are hence associated with several diseases. For this reason, their accurate and efficient identification, and the identification of their targets, is a promising research area and an open topic for the bioinformatics community. We present InSyBio ncRNASeq, a cloud-based Web platform that assists users to analyse ncRNA sequences, and predict whether they are miRNAs, pseudo-hairpins or whether they belong to another ncRNA category. Additionally, InSyBio ncRNASeq offers a unique miRNA target-prediction pipeline, which results in scored miRNA target sites in 3'-untranslated regions (3'-UTR) of messenger RNAs (mRNAs). We show this tool to have the highest accuracy metrics compared with other state-of-the-art methods and tools.

## Introduction

Traditionally, most RNA molecules were regarded as information-carrying intermediates from the gene to the translation machinery, with exceptions being transfer RNA (tRNA) and ribosomal RNA (rRNA), both of which are infrastructural RNAs that are central to the protein-synthesis process. However, since the late 1990s, evidence (Liu *et al.*, 2005) suggests that other types of non-protein-coding RNA molecules are present in organisms ranging from bacteria to mammals, and are involved in a variety of cellular processes, including plasmid replication, phage development, bacterial virulence, chromosome structure, DNA transcription, RNA processing and modification, development control, regulation of gene expression and others. In humans, it is estimated that about 98% of the genome can be transcribed, out of which only ~2% code for proteins, suggesting the possibility that a large percentage of the genome may encode ncRNAs. ncRNAs have been reported as biomarkers for disease and response to treatment in cancer, liver and cardiovascular diseases, and central nervous system disorders, among many others (Lopez *et al.*, 2015).

Although the importance of ncRNAs in cellular activities is well known, and new members and classes of ncRNA are continuously being reported, our current knowledge about the collection of all ncRNAs present in a particular genome is very limited because of the lack of effective computational or experimental methods. Indeed, their computational identification represents one of the most important and challenging issues in computational biology. Existing methods for ncRNA-gene prediction rely mostly on homology information, thus limiting their applications to ncRNA genes that have homologues.

Evidence linking ncRNAs with diseases seems to be attracting the attention of the pharmaceutical and biotechnology industries. Companies and institutions are developing ncRNA-based strategies against cancer, cardiovascular, neurological and muscular diseases. Thus, continued investigation of ncRNA biogenesis and function will provide a new framework for a more comprehensive understanding of human diseases.

InSyBio ncRNASeq enables users to analyse ncRNAs. Users can search and analyse both a specific

**Article history**
Received: 11 December 2016
Accepted: 7 April 2017
Published: XX XX 2017

Korfiati et al., EMBnet.journal 2017, 23, e882
http://dx.doi.org/10.14806/ej.23.0.882



RNA sequence of interest and a full-sequence data-set derived from available online databases or produced by experimental or computational procedures. With InSyBio ncRNASeq, users can predict and analyse ncRNAs and miRNA target genes. InSyBio ncRNASeq also allows result storage in its knowledge-base, equipped with information retrieval tools, to allow users to produce and extract their own data-sets.

# Methodologies

## RNA characterisation and miRNA prediction

InSyBio ncRNASeq enables users to characterise ncRNA sequences. A set of 58 informative features is calculated by supplying the sequence in FASTA format. The features include sequence, thermodynamic and structural properties of the RNA sequences; they were first introduced in (Kleftogiannis *et al.*, 2015), and are presented in Supplementary Table 1. Batch calculations of many sequences are allowed. This enables users to analyse data-sets derived from online databases, experimental sequencing or computational techniques. The results are presented in a browseable table in the Web interface, and can also be downloaded as a tab-delimited text file. For each ncRNA, its sequence and values of its 58 features are presented.

An additional functionality provided by InSyBio ncRNASeq is the prediction of pre-miRNAs, discriminating them from pseudo-hairpins and other molecules. The prediction of pre-miRNAs and pseudo-hairpins is accomplished through the application of a novel methodology, which combines a Genetic Algorithm (GA) (Holland, 1995) with epsilon-Support Vector Regression-SVR (Smola and Schölkopf, 2004) techniques. GAs, a state-of-the-art meta-heuristic optimisation technique, are used to optimise the feature subset to be used as input for the classification model and the regularisation parameters (C), the width of the Gaussian distribution (sigma) of the Radial Basis Function, and the epsilon threshold of e-SVR models. The accuracy of this technique in predicting pre-miRNAs is 95%. A sequence is predicted as 'other' if the minimum free energy is more than -15 kcal/mol, or the number of base pairs is less than 18. The supported input file comprises RNA sequences in FASTA format; batch calculations of many sequences are also allowed. For each ncRNA, the system provides the sequence, its calculated confidence score, the prediction whether it is an miRNA, a pseudo-hairpin or 'other', and its 58 features.

## miRNAs and transcripts knowledge-base

The InSyBio ncRNASeq knowledge-base includes stem-loop and mature miRNAs. For each stem-loop, InSyBio ncRNASeq provides the following information: accession ID, name, species, length, description and comments (if any). For the sequences, the FASTA format is downloadable, and the sequence, its description and secondary structure (in dot-bracket notation) are presented. Visualisation of the secondary structure is performed with FornaContainer (https://github.com/pkerpedjiev/fornac), which shows the Minimum Free Energy (MFE) structure. Stem-loops are linked to their corresponding mature miRNAs. For each mature miRNA, its accession ID, name and sequence are presented. The miRNA sequence is downloadable in FASTA format. Additional information about each mature miRNA is evidence, which may be experimental, similarity to another stem-loop structure, or found in the literature. References for the miRNA of interest are also available, as well as external links to other databases, such as MIRBASE (Griffiths-Jones *et al.*, 2006), ENTREZGENE (Maglott *et al.*, 2007), HGNC (Bruford *et al.*, 2008), RFAM (Griffiths-Jones *et al.*, 2005), and to publications.

The InSyBio ncRNASeq knowledge-base also includes genes and transcripts. For each gene, the information reported is the name, source, description, location, biotype, annotation method and version, as well as the list of its transcripts and links to each transcript. For each transcript, the information provided is the name, source, description, corresponding gene, corresponding protein, location, transcription start site, length, biotype, annotation method, version and 3'-UTR sequence. The 3'-UTR sequence is visualised with FornaContainer, and available for download in FASTA format.

## miRNA target analysis and prediction

InSyBio ncRNASeq enables users to calculate 124 features for every miRNA and potential target-site pair within an mRNA. These features include sequence, thermodynamic and structural properties of the miRNA:mRNA pair; a detailed description of them can be found at (Korfiati *et al.*, 2015). Users can supply a file of miRNA sequences and a file of the respective mRNA binding-site sequences, both in FASTA format. Feature calculation of many miRNA:mRNA pairs is allowed in batch. The results are presented in a browseable table in the Web interface, and can also be downloaded as a tab-delimited text file. For each miRNA:mRNA pair, the miRNA sequence, the mRNA binding-site sequence and the 124 miRNA:mRNA pair features are provided. A description of the supported features for the characterisation of miRNA:mRNA pairs is available in Supplementary Table 2.

With InSyBio ncRNASeq, users can computationally validate miRNA target sites with an intelligent computational technique (hybrid combination of GAs and e-SVRs) and 124 informative features, based on the method presented in (Korfiati *et al.*, 2015), by supplying a file of miRNA sequences and a file of the respective mRNA binding-site sequences in FASTA format. The ncRNASeq tool then creates the miRNA:mRNA pairs by combining all the mRNA target sites from the first

Korfiati et al., EMBnet.journal 2017, 23, e882
http://dx.doi.org/10.14806/ej.23.0.882



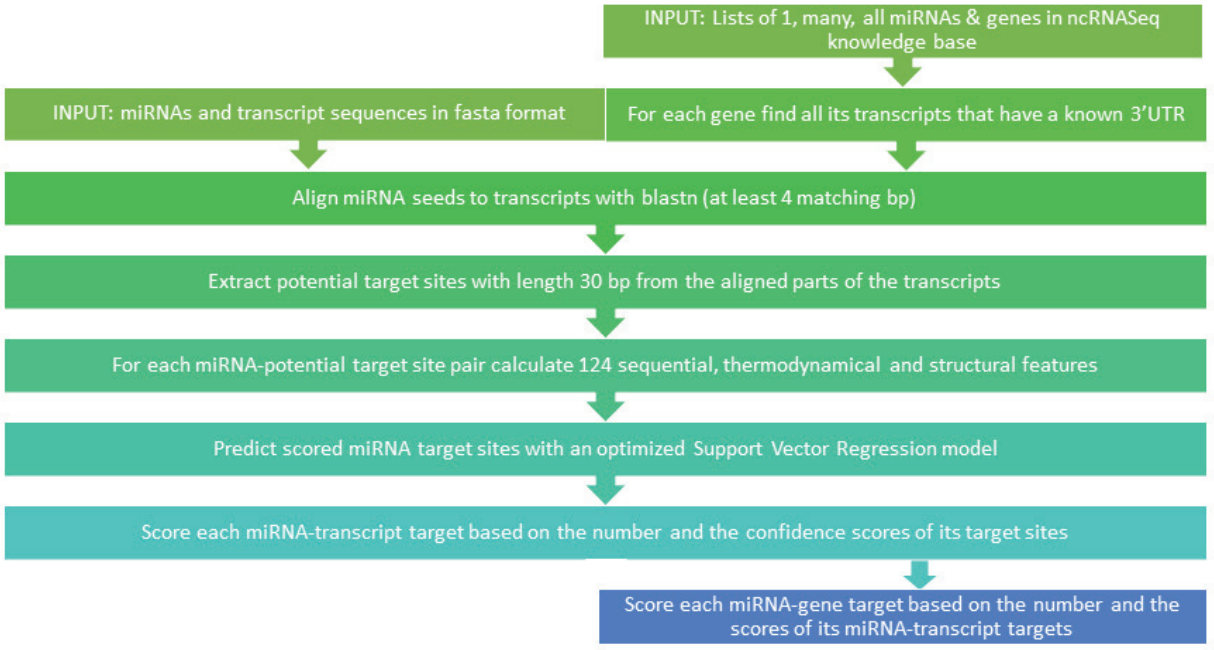


**Figure 1:** InSyBio ncRNASeq miRNA target-prediction workflow.

file with all the miRNAs from the second file. For each miRNA:mRNA pair, InSyBio ncRNASeq provides information about the miRNA sequence, the mRNA binding-site sequence, whether the miRNA:mRNA pair shares a targeting relation or not, the confidence score of the prediction, and the 124 miRNA:mRNA pair features grouped in categories (Thermodynamic, Positional, Motif and Structural). A link from the given miRNA to a protein in the InSyBio Interact database is also presented, indicating that the specific miRNA regulates the expression of the respective mRNA. InSyBio Interact contains information on protein functions, clusters, protein-protein interactions, etc.

Apart from miRNA target-site prediction, InSyBio ncRNASeq enables users to perform miRNA target prediction, either selecting genes or transcripts and miRNAs from the InSyBio ncRNASeq knowledge-base or providing their own miRNAs and transcript sequences in FASTA format. miRNA targets can be predicted in the 3'-UTR sequence of all transcripts of a gene for which a confidence score for the miRNA-gene interaction and additional scores for the miRNA-transcript interactions are provided. Figure 1 shows the miRNA target-prediction workflow.

## Results and Conclusions

As a case study, InSyBio ncRNASeq was used to predict the miRNAs that target the 24 biomarkers reported in (Theofilatos *et al.*, 2016) for Parkinson's Disease (PD). For these biomarkers, InSyBio ncRNASeq yielded 18,359 interactions from 1,594 miRNAs (Table 1). miRTarBase (Chou *et al.*, 2016), the experimentally validated miRNA-target interaction database, contains 585 interactions for the same genes with 489 miRNAs. The results of InSyBio ncRNASeq overlap those of MiRTarBase in 190 interactions (32.5%). The miRNA target prediction tool miRTar (Hsu *et al.*, 2011), which utilises four of the most well-known target-prediction algorithms ((TargetScanS (Lewis *et al.*, 2005), miRanda (Bino *et al.*, 2004), RNAhybrid (Krüger and Rehmsmeier, 2006) and PITA (Kertesz *et al.*, 2007), match only 16 interactions (2.7%) of the experimentally validated interactions in miRTarBase. Figure 2 shows the interaction network among the highest-scoring interactions (10%) predicted by InSyBio ncRNASeq.

These results demonstrate that InSyBio ncRNASeq outperforms other tools in predicting miRNA targets. The predicted interactions are scored, which helps biologists to better design their experiments. Additional

**Table 1:** miRNA target-prediction results for 24 biomarkers of Parkinson's Disease obtained using InSyBio ncRNASeq and miRTar, compared with experimentally validated interactions in miRTarBase

| | Interactions | miRNAs | Genes | Matching results in miRTarBase | Matching results in miRTarBase / miRTarBase total interactions | Matching results in miRTarBase / total predicted interactions |
|---|---|---|---|---|---|---|
| miRTarBase | 585 | 489 | 24 | _ | _ | _ |
| miRTar | 1,430 | 683 | 24 | 16 | 0.027 | ~0.011 |
| ncRNASeq | 18,359 | 1,594 | 24 | 190 | 0.325 | ~0.010 |

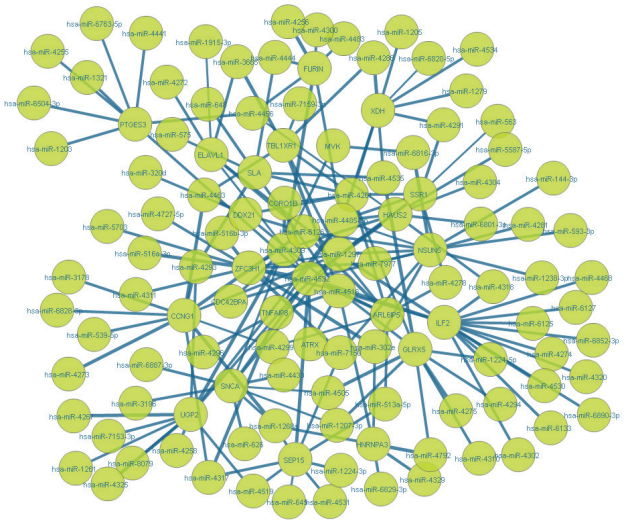

**Figure 2:** Interaction network of the highest-scoring miRNA target interactions (10%) predicted by InSyBio ncRNASeq for the 24 biomarkers of PD.

advantages of InSyBio ncRNASeq are that it predicts scored target sites in each mRNA, and computes sequence, thermodynamic and structural properties of the predicted miRNA:mRNA pair. Additionally, InSyBio ncRNASeq offers users the possibility to run batch searches with numerous miRNAs and mRNAs, with all the possible pairs being considered for potential miRNA:mRNA interactions scored according to the number of target sites and their confidence score. InSyBio ncRNASeq also enables users to interpret their results thanks to the integration of structural and functional information collected from external sources for miRNAs, genes and encoded proteins. Finally, the possibility offered to users for the upload of their own data-sets (transcript sequences) assists researchers searching for miRNA targets in altered transcripts, transcripts with mutations, etc.

## Availability

InSyBio ncRNASeq is one of the tools included in the InSyBio Suite. A demo version of InSyBio ncRNASeq is freely available at http://demo.insybio.com. A free evaluation version includes a one-month free licence that can be obtained by emailing info@insybio.com. To purchase the commercial version of InSyBio ncRNASeq, users can contact sales@insybio.com for a detailed quote and information. InSyBio is registered with the Information Commissioner's Office (ICO) under registration reference number ZA182885 for data-processing issues.

## Acknowledgements

InSyBio participates in the National Bank of Greece (NBG) Business Seeds programme. The NBG does not retain any copyright for this work.

## References

1. Bino J, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004) Human microRNA targets. PLoS Biol 2(11), e363. doi:10.1371/journal.pbio.0020363
2. Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E (2008) The HGNC Database in 2008: a resource for the human genome. Nucleic Acids Res. 36(suppl 1), D445-D448. doi:10.1093/nar/gkm881
3. Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, *et al.* (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res. 44(D1), D239-D247. doi:10.1093/nar/gkv1258
4. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR and Bateman A (2005) Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Research 33(suppl.1), D121-D124. doi:10.1093/nar/gki081
5. Griffiths-Jones S, Grocock RJ, Van Dongen S, Bateman A and Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 34(suppl.1), D140-D144. doi:10.1093/nar/gkj112
6. Holland J (1992) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. Cambridge, MA, USA:MIT Press.
7. Hsu JBK, Chiu CM, Hsu SD, Huang WY, Chien CH *et al.* (2011) miRTar: an integrated system for identifying miRNA-target interactions in human. BMC Bioinformatics 12:300 doi:10.1186/1471-2105-12-300
8. Kertesz M, Iovino N, Unnerstall U, Gaul U and Segal E (2007) The role of site accessibility in microRNA target recognition. Nature Genetics 39(10), 1278-1284. doi:10.1038/ng2135
9. Kleftogiannis D, Theofilatos K, Likothanassis S, Mavroudi S (2015) YamiPred: A novel evolutionary method for predicting pre-miRNAs and selecting relevant features. IEEE/ACM Trans Comput Biol Bioinform. 12(5):1183-92. doi:10.1109/TCBB.2014.2388227
10. Korfiati A, Theofilatos K, Kleftogiannis D, Alexakos C, Likothanassis S and Mavroudi S (2015) Predicting human miRNA target genes using a novel computational intelligent framework. Information Sciences 294, 576-585. doi:10.1016/j.ins.2014.09.016
11. Krüger J and Rehmsmeier M (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Research 34(Suppl. 2), W451-W454. doi:10.1093/nar/gkl243
12. Liu C, Bai B, Skogerbø G, Cai L, Deng W, Zhang Y *et al.* (2005) NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Res. 33(suppl.1), D112-D115. doi:10.1093/nar/gki041
13. Lewis BP, Burge CB and Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120(1), 15-20. doi:10.1016/j.cell.2004.12.035
14. Lopez JP. Cruceanu C, Fiori LM, Laboissiere S, Guillet I *et al.* (2015) Biomarker discovery: quantification of microRNAs and other small non-coding RNAs using next generation sequencing. BMC Med Genomics. 8:35. doi:10.1186/s12920-015-0109-x
15. Maglott D, Ostell J, Pruitt KD and Tatusova T (2007) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research 35 (suppl_1), D26-D31. https://doi.org/10.1093/nar/gkl993
16. Smola AJ and Schölkopf B (2004) A tutorial on support vector regression. Statistics and Computing 14:199. doi:10.1023/B:STCO.0000035301.49549.88
17. Theofilatos K, Dimitrakopoulos C, Alexakos C, Korfiati A, Likothanassis S and Mavroudi S. (2016) InSyBio BioNets: A new tool for analyzing biological networks and its application to biomarker discovery, EMBnet Journal 22, e871. doi:10.14806/ej.22.0.871