Computational methods to study non-coding RNAs

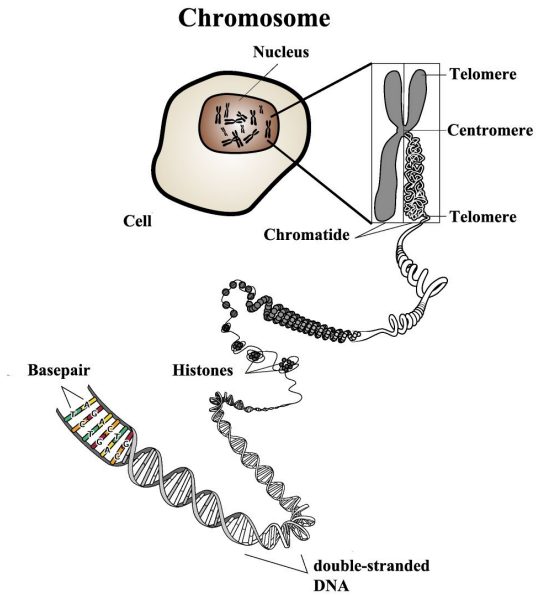
Maria Emilia M. T. Walter
Department of Computer Science
University of Brasilia

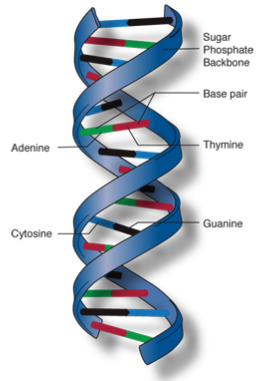
XII National Meeting and
XVIII Autumn School on Mathematical Biology
Universidad Nacional Autónoma de México
Morelia, October 13^{th}, 2016

Topics

Cell and DNA (Jacob, 2015)

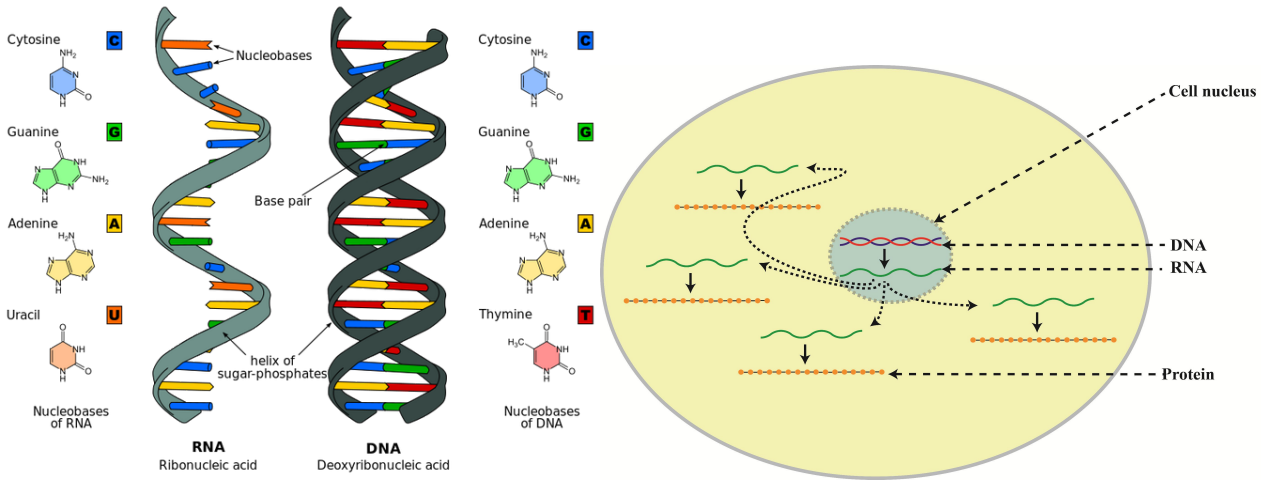
- each organism is consisted of cells
  - multicellular organisms have a cell and a cell nucleus
  - cell nucleus contains DNA: the hereditary material
  - DNA is packed into chromosomes (NHGRI, 2016)
    - human: 46, fruit flies: 8, *Pichia pastoris* (fungus): 4





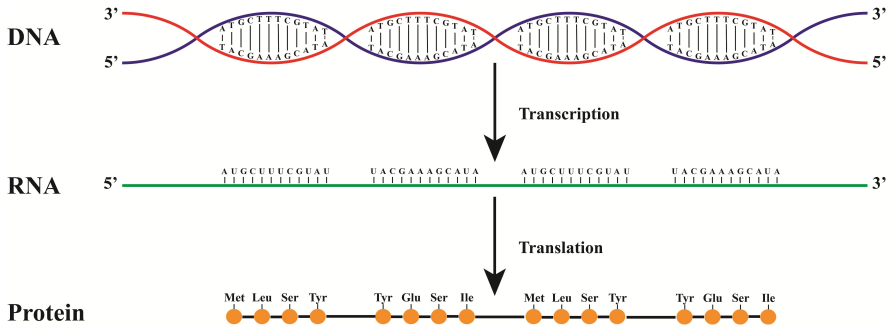
DNA, RNA and protein (NHGRI, 2016)

- chromosomes carry hereditary information (DNA double strand), used in growth, development, functioning and reproduction of all living organisms and many viruses
- DNA information produce proteins with RNA molecules

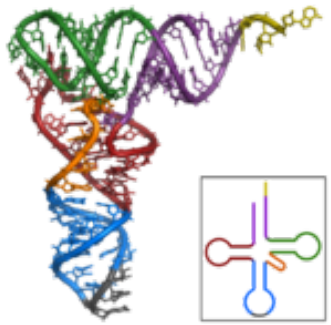


coding RNA and non-coding RNAs

- protein coding - messenger RNAs (mRNA) (Jacob, 2015)



- non-protein coding RNAs (ncRNAs) (Wikipedia, 2015)



Our focus: computational prediction of ncRNAs

- prediction: classification and identification
- computational prediction of proteins:
  - methods work well and are broadly used
  - example: BLAST tool to produce alignment of sequences
- ncRNAs can be “divided” in two sets:
  - small ncRNAs or short ncRNAs
  - long ncRNAs (lncRNAs)

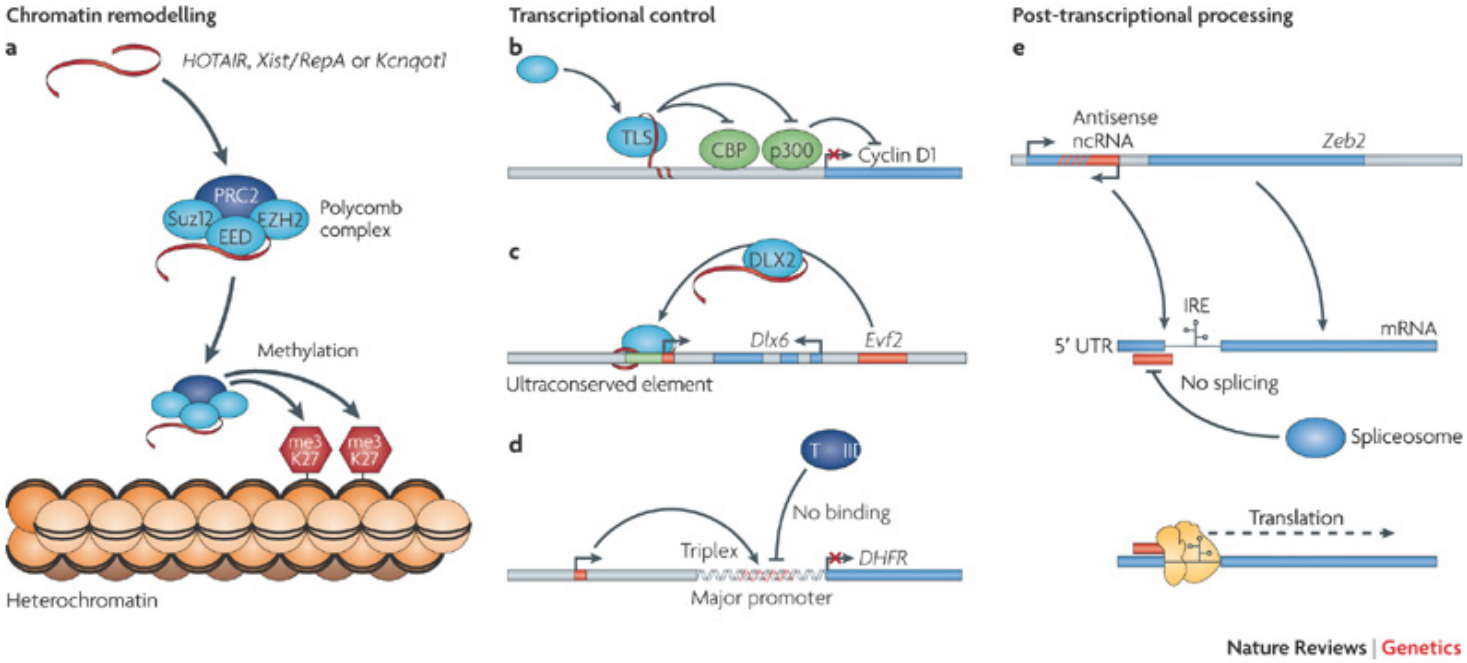
Classification of small ncRNAs

Table: Many known classes of small ncRNAs and corresponding functions (Eddy, 2001; Lakshmi and Agrawal, 2008; Kim et al., 2009; Stadler et al., 2009; Meiri et al., 2010; Christov et al., 2006)

Name	Function
miRNA (micro RNA)	family of genes with regulatory and post-translational functions
snoRNA (small nucleolar RNA)	modifications of rRNAs
siRNA (small interfering RNA)	active molecules in RNA *interference*
snRNA (small nuclear RNA)	spliceosomal RNA
stRNA (small temporal RNA)	interrupts translation of mRNAs
piRNA (piwi-interacting RNAs)	regulation of translation and stabilization of mRNAs
rasiRNA (repeat-associated small interfering RNA)	silencing of gene transcription with chromatin remodeling
Y RNA humano	associated with replication of chromosomal DNA

Classification of some long ncRNAs (lncRNAs)

- not yet well known, difficult to be predicted (Nature Review Genetics, 2015; Ma et al., 2012):



Classification of lncRNAs

- many researches have shown that lncRNAs can regulate all steps of the gene expression process
- the majority of transcribed genes in human genome is composed of lncRNAs (Kapranov et al., 2007)
- number of lncRNAs transcribed in mouse genome is $\sim$ 30.000 (Carninci and et al, 2005)

Identification of ncRNAs

- scanning genomes (chromosomes, scaffolds, super-contigs, contigs) to find ncRNAs:
  - RNA identification was strongly improved with the new sequencing technologies (RNA-seq) to characterize transcriptome outputs
  - transcriptome analysis, in particular, identification of ncRNAs, has been focus of laboratorial and computational techniques

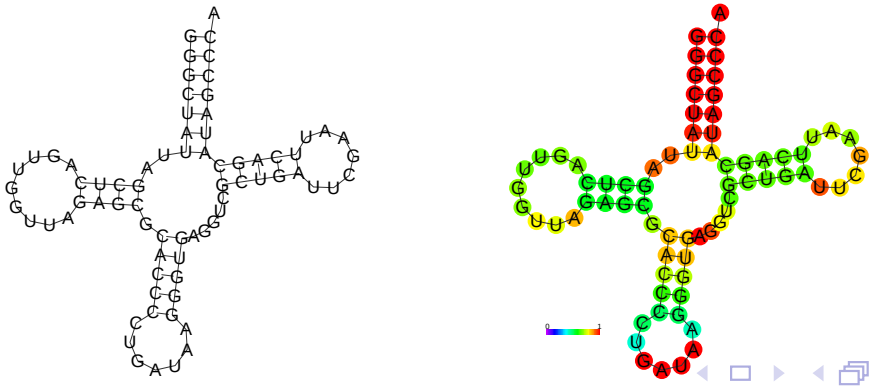
Characteristics and challenges

- annotation of (associating biological functions to) ncRNAs presents three main problems:
  - Prediction of secondary structures from primary structures

>test_sequence

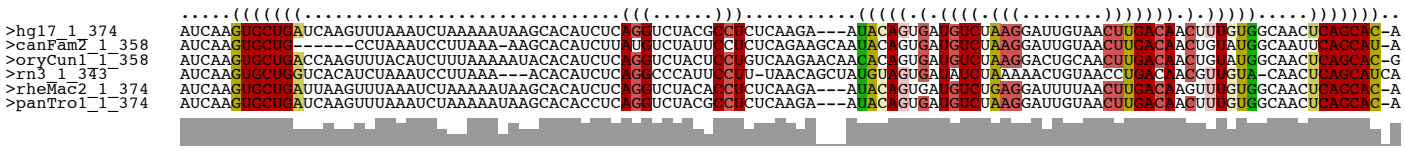
GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCCCUGAUAAGGGUGAGG

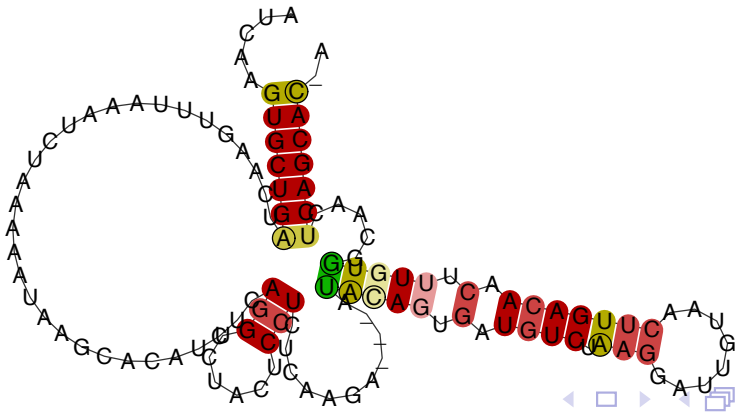
UCGCUGAUUCGAAUUCAGCAUAGCCCA



Characteristics and challenges

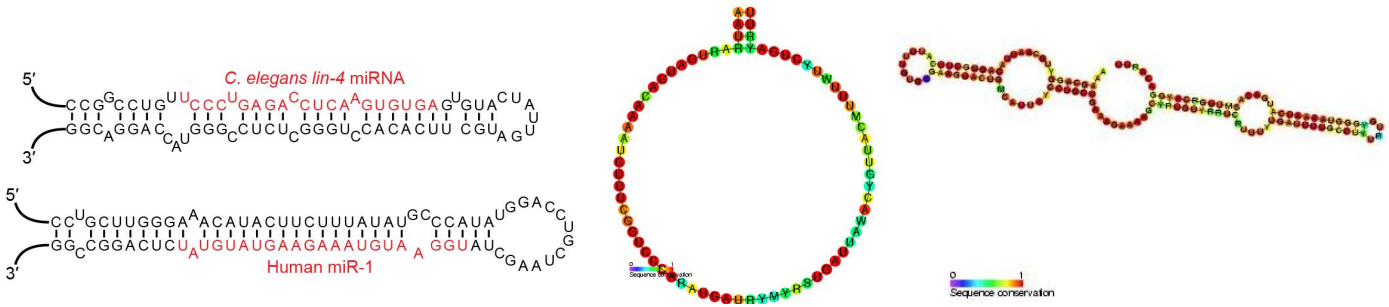
- annotation of ncRNAs involve three main problems:
  - Comparison of secondary structures





Characteristics and challenges

- annotation of ncRNAs involve three main problems:
  - classification and identification of small ncRNAs (miRNA, C/D box snoRNA, H/ACA box snoRNA)



  - classification and identification of lncRNAs: performed using the length of the sequence (more than 200 nucleotides), and the fact that they have small capacity of synthesizing proteins (Ponting et al., 2009; Orom and Shiekhattar, 2011; Mercer et al., 2009)

Homology

- homology: descending from a common ancestor
  - inferred by measuring how similar are the sequences
- prediction of ncRNAs by comparing genomes of two or more species
  - comparison depends on curated databases
    - query sequence of the studied organism
    - database of sequences with already known functions
    - “similar” sequences means that they probably have the same biological function (inherited from a common ancestor)
  - function of a similar sequence is transferred to the query sequence
- Examples:
  - Blast (Altschul et al., 1990, 1997)
  - Infernal (web page, 2015c; Nawrocki and Eddy, 2013)

Class prediction

- prediction of ncRNAs performed by methods of machine learning
- Supervised learning:
  - positive set and a negative set:
    - positive set: known set of ncRNAs
    - negative set: known set of proteins
  - uses characteristics *ab initio*
  - prediction can be done with more reliability
  - Examples:
    - CPC (web page, 2015a; Kong et al., 2007)
    - Portrait (web page, 2015h; Arrial et al., 2009)

De novo

- prediction of ncRNAs performed by models distinct from homology and class prediction
- example: method based in thermodynamics
  - Vienna package (web page, 2015k; Hofacker, 2003)

Multiagent systems (MAS)

- annotation of ncRNAs performed by:
  - multiagent system including many tools and databases
  - inference rules, simulating a human annotator reasoning
- Example:
  - ncRNA-Agents (Arruda et al., 2015; web page, 2015f)

Classes of lncRNAs

- ncRNAs can be classified in 6 categories:
  - intergenic (*long intergenic RNA* - lincRNA (Ponting et al., 2009)): lncRNA localized between two genes
  - intronic: lncRNA is derived from introns
  - bidirectional: start of the transcriptions of both lncRNA and another gene in the opposite strand are close
  - enhancer: enhancer-function can be mediated through a transcribed lncRNA
  - sense: lncRNA overlaps with one or more exons, in the transcription phase, in the sense strand
  - antisense: lncRNA overlaps with one or more exons, in the transcription phase, in the antisense strand

Classes of lncRNAs

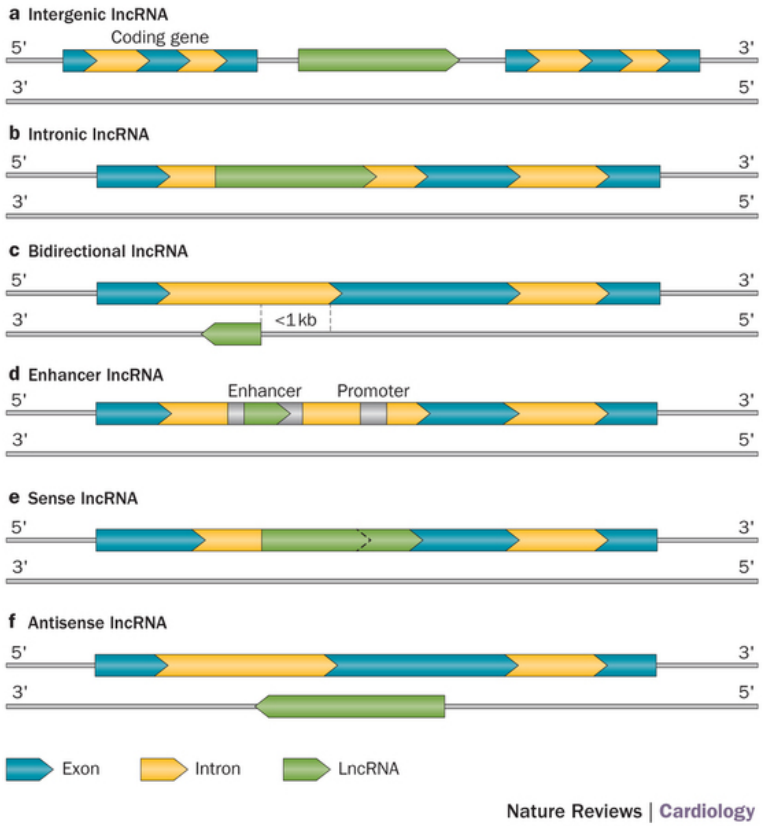


Figure: Six categories of lncRNAs.

Homology

- BLAST (Altschul et al., 1990, 1997)
  - **B**asic **L**ocal **A**lignment **S**earch **T**ool, 1990
  - method of local alignment
  - compares one sequence to other sequences, with already known functions, stored in a database

Alignment

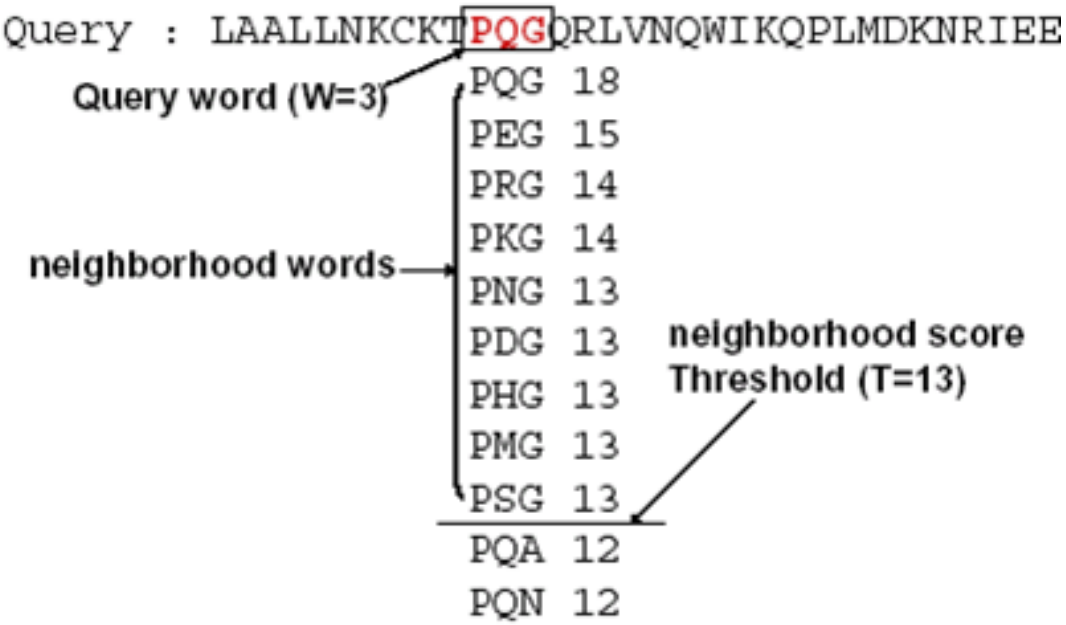
- example: take two sequences *GACGGATTAG* and *GATCGGAATAG*, a possible alignment:
  - G A C G G A T T A G (space: –)
  - G A T C G G A A T A G
  - G A – C G G A T T A G
  - G A T C G G A A T A G
- a **score** is associate to each alignment
- problem: take two sequences and determine the best alignment (highest score) between them
  - global alignment: compare the entire sequences
  - local alignment: compare parts of the sequences

Blast method

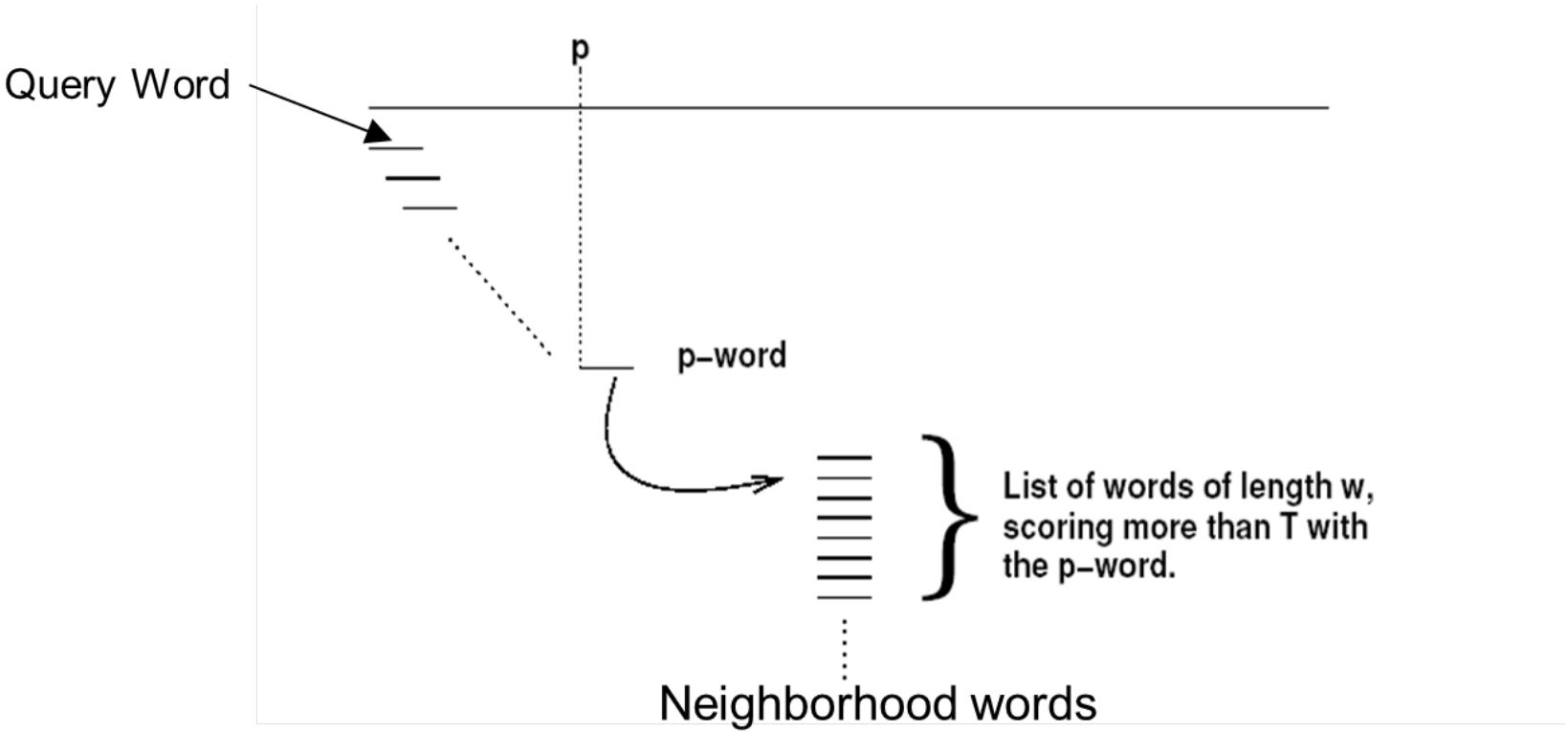
- determine local alignments with scores above a given threshold
- heuristic approach:
  - step 1: finds short words and generates a hash table
  - step 2: finds short words in a database and extends hits
  - step 3: computes statistics of each alignment

step 1

- compiles short strings (*words*) with high scores

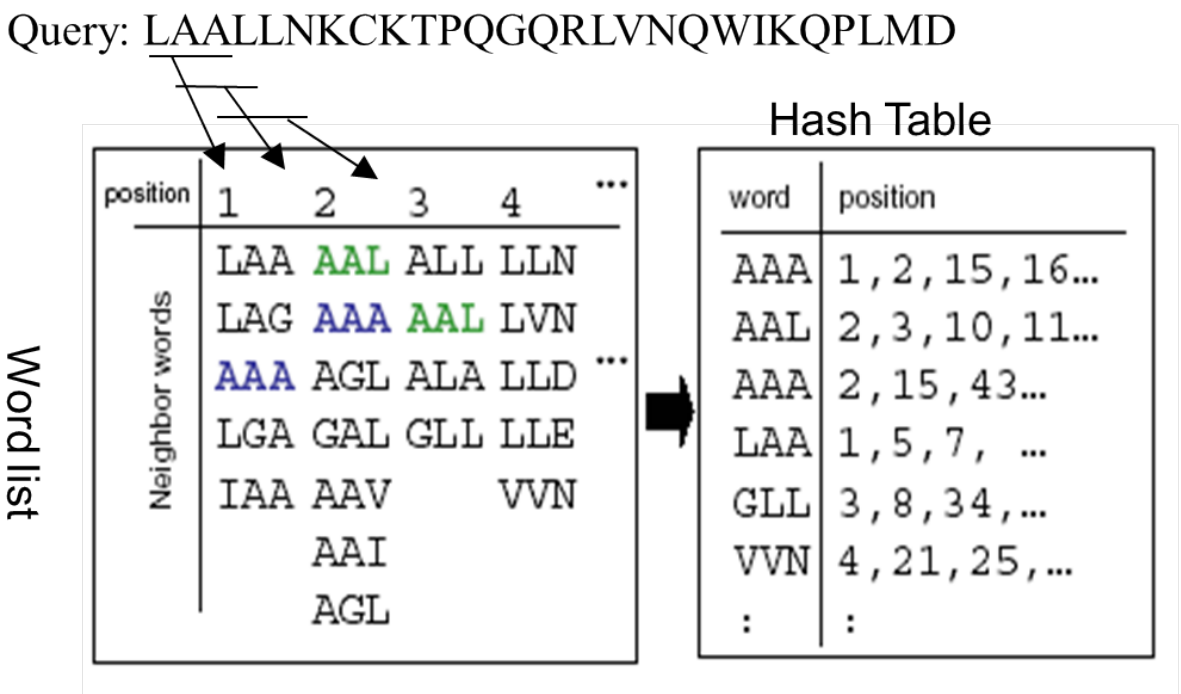


step 1


Query Word
p
p–word
List of words of length w, scoring more than T with the p–word.
Neighborhood words

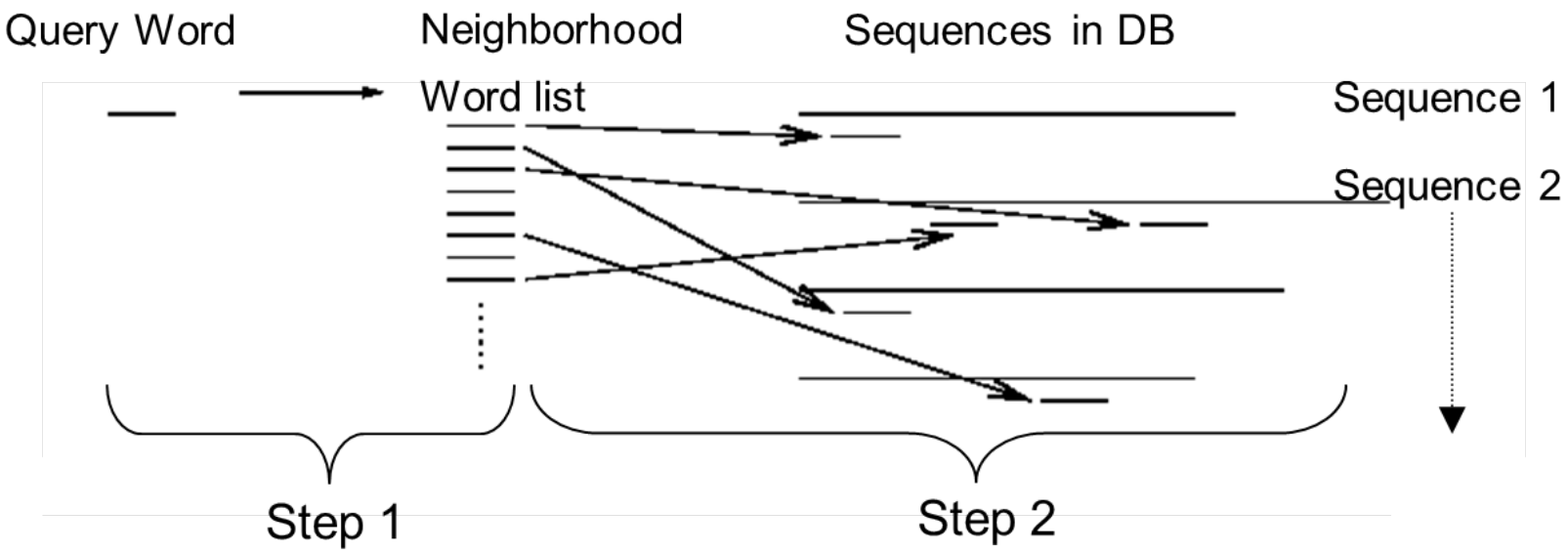
step 1

- generates a hash table



step 2

- finds hits: each hit produces a seed
- hit = *High Scoring Segment Pair* (HSP)
- seeds are extended



step 3

- extension is stopped when $e - value$ (depending on the score S) is lower than a given threshold
- some statistics are computed

Blast output

```
>gi|12643956|sp|Q9Y5X1|SNX9_HUMAN Sorting nexin 9 (SH3 and PX domain-
         containing protein 1) (SDP1 protein) Length = 595

 Score =  255 bits (652), Expect = 4e-68
 Identities = 140/322 (43%), Positives = 185/322 (56%), Gaps = 7/322 (2%)

Query: 221 SSATVSRNLNRFSTFVKSGGEAFVLGEASGFVKDGDKLCVVLGPYGPEWQENPYPFQCTI 280

Sbjct: 197 SSSSMKIPLNKFPGFAKPGTEQYLL--AKQLAKPKEKIPIIVGDYGPMWVYPTSTFDCVV 254

Query: 281 DDPTKQTKFKGMKSYISYKLVPTHTQVPVHRRYKHFDWLYARLAEKF-PVISVPHLPEKQ 339
            DP K +K  G+KSYI Y+L PT+T   V+ RYKHFDWLY RL  KF   I +P LP+KQ
Sbjct: 255 ADPRKGSKMYGLKSYIEYQLTPTNTNRSVNHRYKHFDWLYERLLVKFGSAIPIPSLPDKQ 314

Query: 340 ATGRFEEDFISKRRKGLIWWMNHMASHPVLAQCDVFQHFLTCPSSTDEKAWKQGKRKAEK 399
            TGRFEE+FI  R + L  WM  M  HPV+++ +VFQ FL   +  DEK WK GKRKAE+
Sbjct: 315 VTGRFEEEFIKMRMERLQAWMTRMCRHPVISESEVFQQFL---NFRDEKEWKTGKRKAER 371
```

Homology

- Infernal (web page, 2015c; Nawrocki and Eddy, 2013)
  - **Infer**ence of **RNA al**ignments, 2002
  - searches a database of families, each family containing a secondary structure of the consensus sequence of a multiple alignment of RNA sequences:
    - query sequence and families of RNAs are compared using their secondary structures
    - constructs alignments between the secondary structures of the query and the families of the database

Infernal method

- builds a **profile**, a secondary structure produced from an annotated multiple sequence alignment of an RNA family:
  - scoring system for substitutions, insertions, and deletions
- **profiles**: probabilistic model called **covariance model** (CM)
  - a specialized type of stochastic context-free grammar (SCFG)
  - CMs used to choose similarities between secondary structures of the query sequence and each of the RNA families of the database
- modeling secondary structure is computationally expensive: in recent years, the slow speed of CM homology searches was improved

Homology

- tRNAscan-SE (Lowe and Eddy, 1997)
  - considered one of the more precise predictors of tRNAs
  - also based on CMs
- snoStrip (Bartschat et al., 2014)
  - pipeline for automatic annotation of snoRNAs
  - method uses biological characteristics to predict sites of putative targets:
    - conservation of sequences
    - motifs of canonical boxes
    - secondary structures

Class prediction

- methods based in Support Vector Machine (SVM)
  - CPC (web page, 2015a; Kong et al., 2007)
    - evaluates the protein coding potential of transcripts using characteristics of primary structure of sequences
  - PSoL (Positive Sample only Learning) (Wang et al., 2006)
    - predicts small ncRNAs
  - Portrait (web page, 2015h; Arrial et al., 2009)
    - computes the probability that a transcript does not code for protein

Class prediction

- DARIO (Fasold et al., 2011)
  - web application to predict small ncRNAs from RNA-seq experiments
  - method based in the random forest classifier: identify patterns from characteristics previously identified in different classes of ncRNAs

De novo

- Vienna (web page, 2015k; Hofacker, 2003)
  - package with methods based on thermodynamics to produce or compare secondary structures of sequences:
    - RNAfold: folds a sequence of RNA in two dimensions, computing a spatial structure that presents *Minimum Free Energy* (MFE)
    - RNAz: *de novo* prediction of structured ncRNAs in many sequences (input: multiple alignment of these sequences)
    - RNAAlifold: computes MFE of multiple sequences of RNA (input: multiple alignment of these sequences)
- RNAsnoop (Tafer et al., 2010): predictor of targets for H/ACA snoRNAs:
  - computes interactions H/ACA - RNA termodynamically optimal with dynamic programming
  - uses SVM trained to identify true binding sites

Thermodynamics method (Lorenz et al., 2011)

- RNA secondary structure prediction: energy minimization, with three kinds of dynamic programming algorithms:
  - MFE algorithm of Zuker and Stiegler (1981), which yields a single optimal structure
  - partition function algorithm of McCaskill (1990), which calculates base pair probabilities in the thermodynamic ensemble
  - suboptimal folding algorithm of Wuchty et al. (1999), which generates all suboptimal structures within a given energy range of the optimal energy

Thermodynamics method (Lorenz et al., 2011)

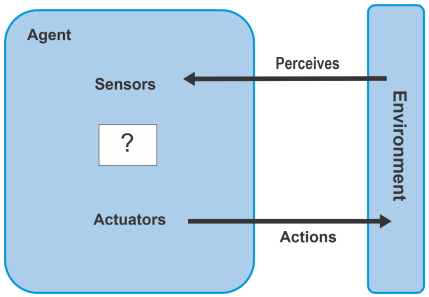
- secondary structure comparison:
  - measures of distance (dissimilarities) using either string alignment or tree-editing (Shapiro and Zhang, 1990)
- algorithm to design sequences with a predefined structure (inverse folding)

MAS

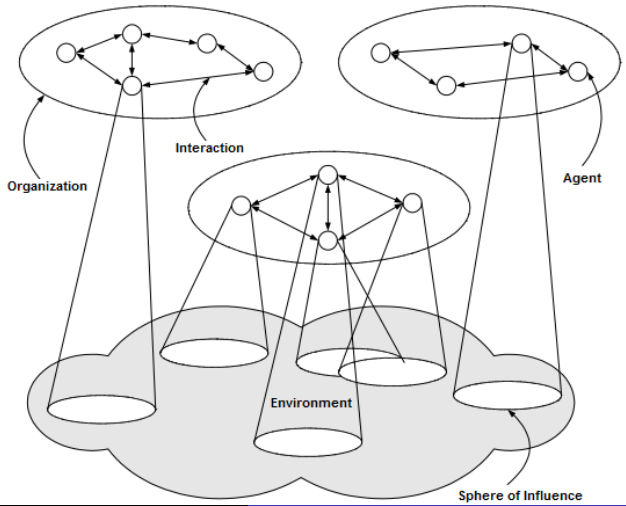
- ncRNA-Agents (web page, 2015f; Arruda et al., 2015)
  - system based on MAS, executes many tools and databases
  - simulates biological reasoning to annotate RNAs using inference rules

Multiagent systems (MAS)

- Agent

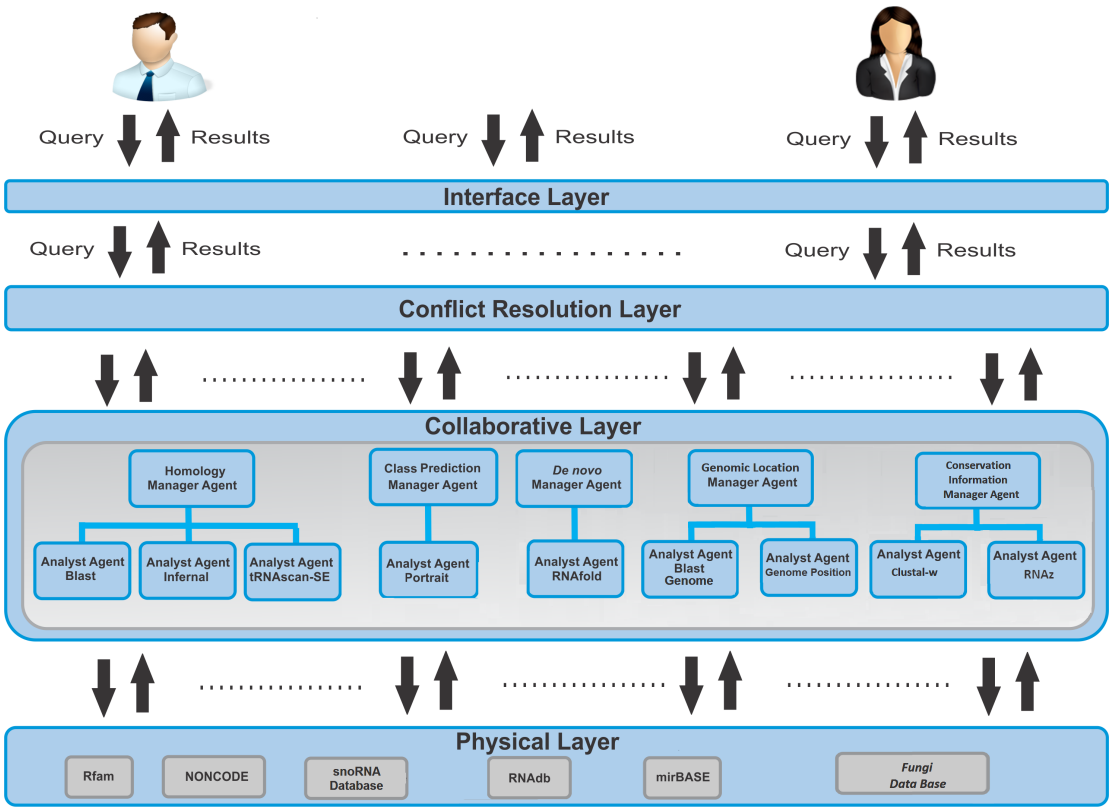


- SMA



SMA

- ncRNA-Agents: architecture with 4 layers



ncRNA-Agents page

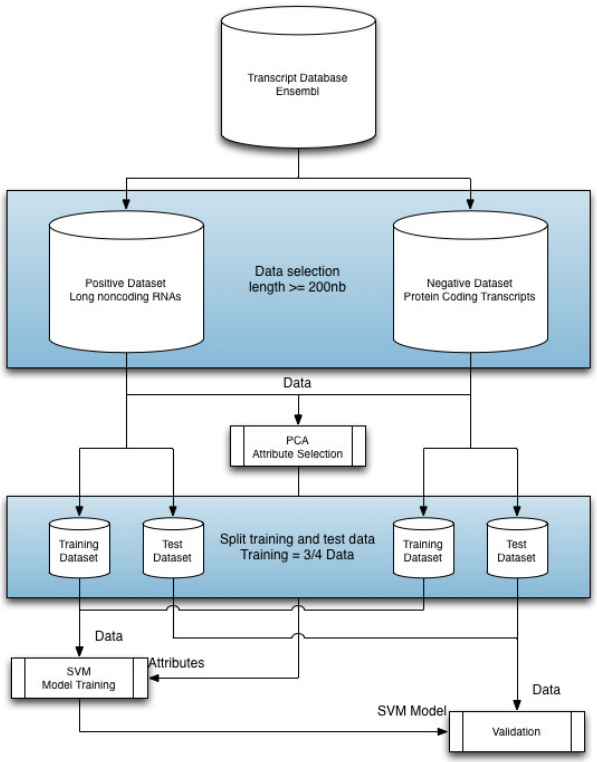
http://lbi.cenargen.embrapa.br:8080/ncrna-agents

Prediction of lncRNAs

- ncRScan-SVM (Sun et al., 2015)
  - predicts protein coding transcripts and lncRNAs using SVM
- iSeeRNA (Sun et al., 2013)
  - identifies lincRNAs in transcriptomes, using SVM
- lnc-GFP (long non-coding RNA global function predictor) (Guo et al., 2013)
  - predicts lncRNA functions
  - method based in a bi-colored network, integrates information of gene expression with information of protein interaction to predict putative functions for lncRNAs
- Workflows and pipelines (Jia et al., 2010; Xiao et al., 2015)

Prediction of lncRNAs: our SVM method

- method developed at University of Brasilia/Brasil, in collaboration with the University of Leipzig/Germany



Prediction of lncRNAs: our SVM method

- Data
  - Ensembl transcripts with length more than 200 nucleotides
  - human (*Homo sapiens*)
    - lncRNAs: 29,020
    - protein coding: 92,425
  - mouse (*Mus musculus*)
    - lncRNAs: 10,495
    - protein coding: 53,854
- Attributes
  - frequencies of patterns of nucleotides: 10, 20, 30, 40, 50 and 60
  - % of ORF length related to the transcript length
  - ORF length

Human and Mouse training and test sets

- human
  - training set
    - 21,999 protein coding transcripts
    - 21,999 lncRNA transcripts
  - test set
    - 7,333 protein coding transcripts
    - 7,333 lncRNA transcripts
- mouse
  - training set
    - 7,871 protein coding transcripts
    - 7,871 lncRNA transcripts
  - test set
    - 2,623 protein coding transcripts
    - 2,623 lncRNA transcripts

Results

- human
  - 97.7% of accuracy
  - iSeeRNA: 18.39% of accuracy (not fair, since another assembly was used in their method), 97.21% of sensitivity and 3.24% of specificity
- mouse
  - 97.0% of accuracy
  - iSeeRNA: 73.04% of of accuracy
- both, human and mouse
  - 96.9% of accuracy

Validation

- cross validation Human Model x Mouse Test: 96.9% of accuracy
- cross validation Mouse Model x Human Test: 96.5% of accuracy
- Human Model:
  - lncRNAs of pig: 195 of 226 (86.2%)
  - lncRNAs of rat: 3,361 of 3,463 (97.0%)
  - lncRNAs of zebrafish: 3,881 of 3,940 (98.5%)
- Mouse Model:
  - lncRNAs of pig: 186 of 226 (82.3%)
  - lncRNAs of rat: 3,339 of 3,463 (96.4%)
  - lncRNAs of zebrafish: 3,868 of 3,940 (98.1%)

Tools to identify ncRNAs

- SnoSeeker (Yang et al. (2006))
  - snoRNAs (small nucleolar RNAs): large group of ncRNAs in eukaryotes
  - divided in guide and orphan snoRNAs, according to the presence/absence of targets (antisense sequences for rRNAs or snRNAs)
  - programs CDseeker and ACAseeker: specific to identify genes of guide and orphan snoRNAs in mammal genomes
  - method identified many orphan snoRNAs in humans

Tools to identify ncRNAs

- snoReport (Hertel et al. (2008))
  - identify in sequences two classes: C/D box snoRNAs and H/ACA box snoRNAs
  - method combines prediction of secondary structure and SVM
  - like snoSeeker, it does not use information of sites of putative targets in rRNAs or spliceosomal RNAs
  - method found many orphan snoRNAs in many organisms
  - this year: an improved method developed at University of Brasilia and University of Leipzig was accepted for publication in BMC Bioinformatics

Databases

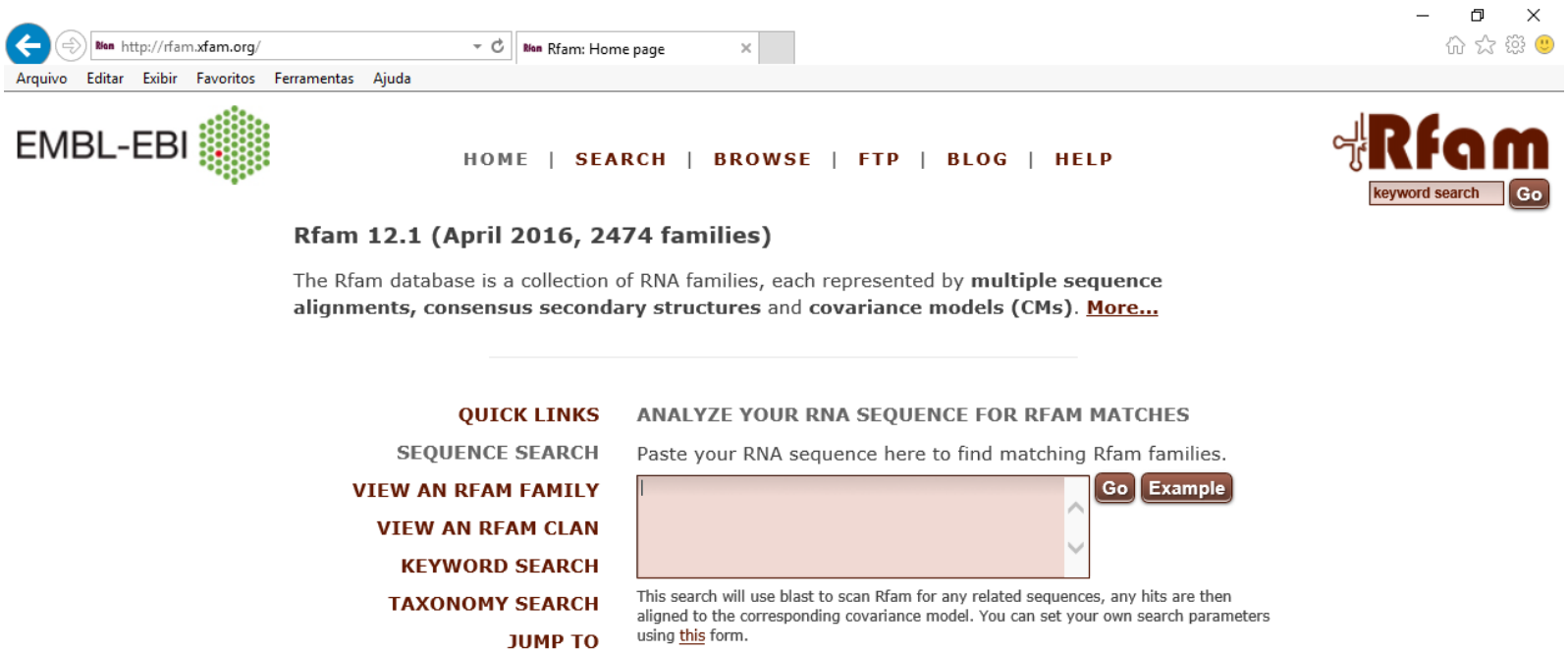
- different classes and families of ncRNAs
- NONCODE (Liu et al., 2005)
  - ncRNAs automatically found in the literature and GenBank, manually curated
- RNAdb (Pang et al., 2007)
  - ncRNAs of mammals
  - contains sequences and annotation of millions of ncRNAs
- miRBase (Kozomara and Griffiths-Jones, 2011)
  - contains microRNAs

Many classes and families of ncRNAs

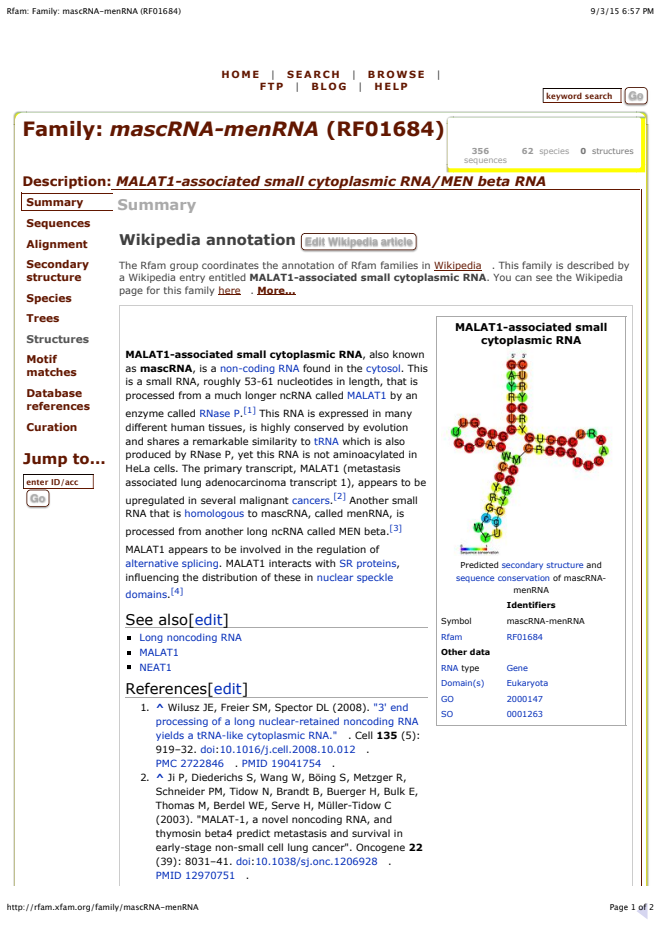
- snoRNAdb (web page, 2015j)
  - contains snoRNAs of many organisms: plants (*Arabidopsis thaliana*), Archea, fungi (*Saccharomyces cerevisiae*)
- Plant snoRNA database (web page, 2015g; Brown et al., 2003)
  - contains plant snoRNAs
- fRNAdb (Kin et al., 2007; web page, 2015b)
  - integrates databases of sequences, annotated and non-annotated, of the databases H-inv, NONCODE and RNAdb

Many classes and families of ncRNAs (Rfam)

- Rfam 12.1 (Griffiths-Jones et al., 2003; Burge et al., 2013; Nawrocki et al., 2014; web page, 2015i): 2.474 families



Many classes and families of ncRNAs (Rfam)



Rfam: Family: mascRNA-menRNA (RF01684) 9/3/15 6:57 PM

HOME | SEARCH | BROWSE | FTP | BLOG | HELP

keyword search Go

Family: *mascRNA-menRNA* (RF01684)

356 sequences 62 species 0 structures

Description: *MALAT1-associated small cytoplasmic RNA/MEN beta RNA*

Summary
Sequences
Alignment
Secondary structure
Species
Trees
Structures
Motif matches
Database references
Curation

Jump to... enter ID/acc Go

Summary

Wikipedia annotation Edit Wikipedia article

The Rfam group coordinates the annotation of Rfam families in Wikipedia . This family is described by a Wikipedia entry entitled **MALAT1-associated small cytoplasmic RNA**. You can see the Wikipedia page for this family here . **More...**

MALAT1-associated small cytoplasmic RNA, also known as **mascRNA**, is a non-coding RNA found in the cytosol. This is a small RNA, roughly 53-61 nucleotides in length, that is processed from a much longer ncRNA called MALAT1 by an enzyme called RNase P.[1] This RNA is expressed in many different human tissues, is highly conserved by evolution and shares a remarkable similarity to tRNA which is also produced by RNase P, yet this RNA is not aminoacylated in HeLa cells. The primary transcript, MALAT1 (metastasis associated lung adenocarcinoma transcript 1), appears to be upregulated in several malignant cancers.[2] Another small RNA that is homologous to mascRNA, called menRNA, is processed from another long ncRNA called MEN beta.[3]

MALAT1 appears to be involved in the regulation of alternative splicing. MALAT1 interacts with SR proteins, influencing the distribution of these in nuclear speckle domains.[4]

MALAT1-associated small cytoplasmic RNA

Predicted secondary structure and sequence conservation of mascRNA-menRNA

Identifiers	
Symbol	mascRNA-menRNA
Rfam	RF01684
Other data	
RNA type	Gene
Domain(s)	Eukaryota
GO	2000147
SO	0001263

See also[edit]

- Long noncoding RNA
- MALAT1
- NEAT1

References[edit]

1. ^ Wilusz JE, Freier SM, Spector DL (2008). "3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA." . Cell **135** (5): 919–32. doi:10.1016/j.cell.2008.10.012 . PMC 2722846 . PMID 19041754 .
2. ^ Ji P, Diederichs S, Wang W, Böing S, Metzger R, Schneider PM, Tidow N, Brandt B, Buerger H, Bulk E, Thomas M, Berdel WE, Serve H, Müller-Tidow C (2003). "MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer". Oncogene **22** (39): 8031–41. doi:10.1038/sj.onc.1206928 . PMID 12970751 .

http://rfam.xfam.org/family/mascRNA-menRNA Page 1 of 2

Many classes and families of lncRNAs

- NRED (ncRNA Expression Database) (Dinger et al., 2009)
  - contains lncRNAs of the genomes of human and mouse
  - provides other information: known ncRNAs, evolutionary conservation, evidences for secondary structures and links to genomic contexts
- DIANA-lncBase (Paraskevopoulou et al., 2013)
  - provide interactions miRNA-lncRNA
  - contains two modules:
    - experimental: with detailed information about more than 5.000 interactions, among 2.958 lncRNAs and 120 miRNAs
    - prediction: with information of more than 10 millions of interaction, among 56.097 lncRNAs and 3.078 miRNAs

Many classes and families of lncRNAs

- lncRNADisease (Chen et al., 2013)
  - contains lncRNAs associated to diseases
  - provides more than 600 input for lncRNAs related to diseases and 475 input for interactions of lncRNAs (251 lncRNAs and 217 diseases)
- lncRNAdb (web page, 2015d; Amaral et al., 2011; Quek et al., 2014)
  - contains a large volume of eukaryotic lncRNAs
  - each input is manually curated, from the literature
  - each input contains information about RNA, including:
    - nucleotide sequence and genomic context
    - information of gene expression
    - structural information and subcellular localization
    - conservation and function
    - bibliographic references

Many classes and families of lncRNAs

- lncRNAtor (Park et al., 2014; web page, 2015e) provides (web interface) information for research of functional lncRNAs: annotation, sequence analysis, gene expression, protein interaction and phylogenetic conservation:
  - ncRNAs of six species (human, mouse, zebrafish, fruit fly, worm and yeast) collcted from ENSEMBL, HGNC, MGI and lncRNAdb
  - information of gene expression of 208 researches of RNA-Seq (4.995 samples), collected from the databases GEO, ENCODE, modENCODE and TCGA, used to verify expression profile in many tissues, diseases and developing phases
  - provide interactions protein-lncRNA, found by analyses of sequencing data through CLIP-seq and PAR-CLIP (interaction protein-RNA)
  - lncRNAs evolutionarely conserved among human and six other organisms to identify functional lncRNAs

Reflections

- Technically speaking:
  - ncRNAs are important molecules, since they play important roles in cellular mechanisms
  - computational methods can support, very efficiently, to predict ncRNA functions and to design experiments in wet labs
  - techniques of distinct areas of computer science have been used:
    - algorithms: dynamic programming, special grammars involving probability - CM, thermodynamics - MFE, graphs
    - machine learning: supervised methods - SVM, semi-supervised methods, random forest
    - multiagent systems: agents executing annotation tools and rules simulating human reasoning

Reflections

- Speaking about Biologia Matematica (Computational Biology or Bioinformatics)
  - key word: collaboration
  - Brasil: many groups collaborating, mainly biologists and computer scientists (interaction is difficult!)
    - at the University of Brasilia: Computer Science and Molecular Biology, in transcriptome and genome projects
    - other areas of computer science: efficient storage for a large volume of data (databases, noSQL, big data), efficient computation (parallel computing - GPUs, cloud computing)
    - now: trying to attract statisticians

Reflections

- Multidisciplinary projects are essential in Science
  - large volumes of data
  - Exact Sciences: Mathematics, Statistics, Physics, Computer Science, ...
  - Life Sciences: Biology (Ecology), Chemistry, Medicine, ...
- How to train scientists?
- How to encourage interdisciplinary research?

Thanks to my collaborators and students

- Peter Stadler and Jana Hertel, University of Leipzig
- Marcelo Brigido, Laboratory of Molecular Biology - University of Brasilia
- Taina Raiol, Fiocruz Amazonia
- Celia Ghedini Ralha, Department of Computer Science - University of Brasilia
- Daniel Souza, Hugo Schneider, Joao Victor Araujo, Lucas Maciel and Bruno Kummel, Department of Computer Science - University of Brasilia

Thank you for your attention!

email: mariaemilia@unb.br

Bibliography I

Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D., et al. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Altschul, S. F., Madden, T. L., Schffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped Blast and PsiBlast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

Amaral, P. P., Clark, M. B., Gascoigne, D. K., Dinger, M. E., and Mattick, J. S. . (2011). lncRNAdb: a reference database for long noncoding RNAs. *Nucleic acids research*, 39:D146–D151.

Arrial, R., Togawa, R., and Brigido, M. (2009). Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC bioinformatics*, 10(1):239.

Arruda, W. C., Souza, D. S., Ralha, C. G., Walter, M. E. M. T., Raiol, T., Brigido, M. M., and Stadler, P. F. (2015). Knowledge-based reasoning to annotate noncoding RNA using multi-agent system. *Journal of Bioinformatics and Computational Biology*, page Online Ready.

Bibliography II

Bartschat, S., Kehr, S., Tafer, H., Stadler, P. F., and Hertel, J. (2014). snoStrip: A snoRNA annotation pipeline. *Bioinformatics*, 30(1):115–116.

Brown, J., Echeverria, M., Qu, L., Lowe, T., Bachellerie, J., Hüttenhofer, A., Kastenmayer, J., Green, P., Shaw, P., and Marshall, D. (2003). Plant snoRNA database. *Nucleic Acids Research*, 31(1):432–435.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P., and Bateman, A. (2013). Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research*, 41(D1):D226–D232.

Carninci, P. and et al (2005). The Transcriptional Landscape of the Mammalian Genome. *Science*, 309(5740):1559–1563.

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., and Cui, Q. (2013). LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Research*, 41(Database issue):D983–6.

Christov, C. P., Gardiner, T. J., Szüts, D., and Krude, T. (2006). Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. *Mol Cell Biol*, 26(18):6993–7004.

Bibliography III

Dinger, M. E., Pang, K. C., Mercer, T. R., Crowe, M. L., Grimmond, S. M., and Mattick, J. S. (2009). NRED: a database of long noncoding RNA expression. *Nucleic Acids Research*, 37(Database-Issue):122–126.

Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929.

Fasold, M., Langenberger, D., Binder, H., Stadler, P. F., and Hoffmann, S. (2011). DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic acids research*, 39(suppl 2):W112–W117.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. (2003). Rfam: an RNA family database. *Nucleic Acids Research*, 31(1):439–441.

Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., Luo, H., Zhao, G., Bu, D., Jiao, F., Shao, Q., Chen, R., and Zhao, Y. (2013). Long non-coding rnas function annotation: a global prediction method based on bi-colored networks. *Nucleic Acids Research*, 41(2).

Bibliography IV

Hertel, J., Hofacker, I. L., and Stadler, P. F. (2008). SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24(2):158–164. doi: 10.1093/bioinformatics/btm464.

Hofacker, I. (2003). Vienna RNA secondary structure server. *Nucleic acids research*, 31(13):3429–3431.

Jacob, C. (2015). DNA, RNA and protein - the Central Dogma. `http://www.science-explained.com/theory/dna-rna-and-protein/`. Online; accessed 01 September 2015.

Jia, H., Osak, M., Bogu, G. K., Stanton, L. W., Johnson, R., and Lipovich, L. (2010). Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA*, 16(8):1478–1487.

Bibliography V

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., and Gingeras, T. R. (2007). RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science*, 316(5830):1484–1488.

Kim, V. N., Han, J., and Siomi, M. C. (2009). Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology*, 10(2):126–139.

Kin, T., Yamada, K., Terai, G., Okida, H., Yoshinari, Y., Ono, Y., Kojima, A., Kimura, Y., Komori, T., and Asai, K. (2007). fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Research*, 35(Database issue):D145–148.

Kong, L., Y., Z., Ye, Z.-Q., X.-Q., L., Zhao, S.-Q., Wei, L., and G., G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, 35(Web Server issue):W345–9.

Bibliography VI

Kozomara, A. and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids research*, 39(Database issue):D152–7.

Lakshmi, S. S. and Agrawal, S. (2008). piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic acids research*, 36(suppl 1):D173–D177.

Liu, C., Bai, B., Skogerbø, G., Cai, L., Deng, W., Zhang, Y., Bu, D., Zhao, Y., and Chen, R. (2005). NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic acids research*, 33(Database issue):D112–5.

Lorenz, R., Bernhart, S. H., Hner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26.

Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research*, 25(5):O955–O964.

Bibliography VII

Ma, H., Hao, Y., Dong, X., Gong, Q., Chen, J., Zhang, J., and Tian, W. (2012). Molecular Mechanisms and Function Prediction of Long Noncoding RNA: Review Article. *The Scientific World Journal*, 2012(Article ID 541786):1–11.

Meiri, E., Levy, A., Benjamin, H., Ben-David, M., Cohen, L., Dov, A., Dromi, N., Elyakim, E., Yerushalmi, N., Zion, O., et al. (2010). Discovery of microRNAs and other small RNAs in solid tumors. *Nucleic acids research*, 38(18):6234–6246.

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nature reviews. Genetics*, 10(3):155–9.

Nature Review Genetics (2015). Long non-coding RNAs. `http://www.nature.com/nrg/journal/v10/n3/fig_tab/nrg2521_F2.html`. Online; accessed 01 September 2015.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. (2014). Rfam 12.0: updates to the RNA families database. *Nucleic Acids Research*, pages 1–8.

Nawrocki, E. P. and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29:2933–2935.

Bibliography VIII

NHGRI (2016). National Human Genome Research Institute: Structure of the Double Helix. https://geneed.nlm.nih.gov/topic_subtopic.php?tid=15&sid=16. Online; accessed 04 October 2016.

Orom, U. A. and Shiekhattar, R. (2011). Noncoding RNAs and enhancers: complications of a long-distance relationship. *Trends in genetics : TIG*, 27(10):433–439.

Pang, K., Stephen, S., Dinger, M., Engström, P., Lenhard, B., and Mattick, J. (2007). RNAdb 2.0: an expanded database of mammalian non-coding RNAs. *Nucleic Acids Research*, 35(suppl 1):D178–D182.

Paraskevopoulou, M. D., Georgakilas, G., Kostoulas, N., Reczko, M., Maragkakis, M., Dalamagas, T. M., and Hatzigeorgiou, A. G. (2013). DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Research*, 41(Database issue):D239–45.

Bibliography IX

Park, C., Yu, N., Choi, I., Kim, W., and Lee, S. (2014). lncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs. *Bioinformatics*, 30(17):2480–2485.

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell*, 136(4):629–641.

Quek, X. C., Thomson, D. W., Maag, J. L., Bartonicek, N., Signal, B., Clark, M. B., Gloss, B. S., and Dinger, M. E. (2014). lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Research*, 43:D168–D173.

Shapiro, B. A. and Zhang, K. Z. (1990). Comparing multiple RNA secondary structures using tree comparisons. *Comput Appl Biosci*, 6(4):309–318.

Stadler, P. F., Chen, J. J. L., Hackermüller, J., Hoffmann, S., Horn, F., Khaitovich, P., Kretzschmar, A. K., Mosig, A., Prohaska, S. J., Qi, X., et al. (2009). Evolution of vault RNAs. *Molecular biology and evolution*, 26(9):1975–1991.

Bibliography X

Sun, K., Chen, X., Jiang, P., Song, X., Wang, H., and Sun, H. (2013). iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC genomics*, 14 Suppl 2:S7.

Sun, L., Liu, H., Zhang, L., and Meng, J. (2015). lncRScan-SVM: A Tool for Predicting Long Non- Coding RNAs Using Support Vector Machine. *PLoS ONE*, 10(10):e0139654.

Tafer, H., Kehr, S., Hertel, J., Hofacker, I. L., and Stadler, P. F. (2010). RNAsnoop: efficient target prediction for H/ACA snoRNAs. *Bioinformatics (Oxford, England)*, 26(5):610–616.

Wang, C., Ding, C., Meraz, R. F., and Holbrook, S. R. (2006). PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. *Bioinformatics*, 22(21):2590–2596.

web page (2015a). CPC. `http://cpc.cbi.pku.edu.cn`.

web page (2015b). fRNAdb. `http://www.ncrna.org/`.

web page (2015c). Infernal user's guide. `http://infernal.janelia.org`.

Bibliography XI

web page (2015d). lncRNAdb. http://www.lncrnadb.org/.

web page (2015e). lncRNAtor. http://lncrnator.ewha.ac.kr/index.htm.

web page (2015f). ncRNA-Agents. http://www.biomol.unb.br/ncrna-agents.

web page (2015g). Plant snoRNA database.
http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/home.

web page (2015h). Portrait.
http://bioinformatics.cenargen.embrapa.br/portrait/.

web page (2015i). Rfam database.
http://ftp.sanger.ac.uk/pub/databases/Rfam.

web page (2015j). snoRNAdb. http://lowelab.ucsc.edu/snoRNAdb/.

web page (2015k). Vienna. http://rna.tbi.univie.ac.at/.

Wikipedia (2015). Non-coding RNA.
https://en.wikipedia.org/wiki/Non-coding_RNA. Online; accessed 01 September 2015.

Bibliography XII

Xiao, Y., Lv, Y., Zhao, H., Gong, Y., Hu, J., Li, F., Xu, J., Bai, J., Yu, F., and Li, X. (2015). Predicting the Functions of Long Noncoding RNAs Using RNA-Seq Based on Bayesian Network. *BioMed Research International*, 2015(Article ID 839590):1–14.

Yang, J.-H., Zhang, X.-C., Huang, Z.-P., Zhou, H., Huang, M.-B.and Zhang, S., Chen, Y.-Q., and Qu, L.-H. (2006). snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Research*, 34(18):5112–5123. http://doi.org/10.1093/nar/gkl672.