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 13th, 2016
Introduction

Methods to classify ncRNAs

Methods to identify ncRNAs

Databases of ncRNAs

Final Remarks

References

Topics

Introduction
  About non-coding RNAs
  Classification × identification of ncRNAs
  Annotation of ncRNAs

Methods to classify ncRNAs
  Paradigms
  Prediction of IncRNAs
  Tools

Methods to identify ncRNAs
  Tools

Databases of ncRNAs

Final Remarks

Maria Emilia M. T. Walter  non-coding RNAs
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
Introduction

Methods to classify ncRNAs
Methods to identify ncRNAs
Databases of ncRNAs
Final Remarks
References

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)

<table>
<thead>
<tr>
<th>Name</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>miRNA (micro RNA)</td>
<td>family of genes with regulatory and post-translational functions</td>
</tr>
<tr>
<td>snoRNA (small nucleolar RNA)</td>
<td>modifications of rRNAs</td>
</tr>
<tr>
<td>siRNA (small interfering RNA)</td>
<td>active molecules in RNA <em>interference</em></td>
</tr>
<tr>
<td>snRNA (small nuclear RNA)</td>
<td>spliceosomal RNA</td>
</tr>
<tr>
<td>stRNA (small temporal RNA)</td>
<td>interrupts translation of mRNAs</td>
</tr>
<tr>
<td>piRNA (piwi-interacting RNAs)</td>
<td>regulation of translation and stabilization of mRNAs</td>
</tr>
<tr>
<td>rasiRNA (repeat-associated small interfering RNA)</td>
<td>silencing of gene transcription with chromatin remodeling</td>
</tr>
<tr>
<td>Y RNA humano</td>
<td>associated with replication of chromosomal DNA</td>
</tr>
</tbody>
</table>
Classification of some long ncRNAs (IncRNAs)

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

- many researches have shown that IncRNAs can regulate all steps of the gene expression process
- the majority of transcribed genes in human genome is composed of IncRNAs (Kapranov et al., 2007)
- number of IncRNAs 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
GGGCUAUUAGCUCAGUUUGGUAGAGCGCACCUCUGAUAAAGGGUGAGG
UCGCUGAUUCGAAAUCAGCAUAGCCCA

Introduction
Methods to classify ncRNAs
Methods to identify ncRNAs
Databases of ncRNAs
Final Remarks
References

About non-coding RNAs
Classification × identification of ncRNAs
Annotation of ncRNAs
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 IncRNAs: 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)
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)
ncRNAs can be classified in 6 categories:

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

Figure: Six categories of IncRNAs.
Homology

- **BLAST (Altschul et al., 1990, 1997)**
  - **Basic Local Alignment Search Tool, 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 associated 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
Introduction

Methods to classify ncRNAs
Methods to identify ncRNAs
Databases of ncRNAs
Final Remarks
References

Paradigms
Prediction of IncRNAs
Tools

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

Query: LAALLNKCKTPQGQRLVQNQWIKQPLMD

<table>
<thead>
<tr>
<th>position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>LAA</td>
<td>AAL</td>
<td>ALL</td>
<td>LLN</td>
<td></td>
</tr>
<tr>
<td>word</td>
<td>LAG</td>
<td>AAA</td>
<td>AAA</td>
<td>LVN</td>
<td></td>
</tr>
<tr>
<td>word</td>
<td>AAA</td>
<td>AGL</td>
<td>ALA</td>
<td>LLD</td>
<td></td>
</tr>
<tr>
<td>word</td>
<td>LGA</td>
<td>GAL</td>
<td>GLL</td>
<td>LLE</td>
<td></td>
</tr>
<tr>
<td>word</td>
<td>IAA</td>
<td>AAV</td>
<td>VVN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>word</td>
<td>AAI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>word</td>
<td>AGL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hash Table

<table>
<thead>
<tr>
<th>word</th>
<th>position</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA</td>
<td>1, 2, 15, 16...</td>
</tr>
<tr>
<td>AAL</td>
<td>2, 3, 10, 11...</td>
</tr>
<tr>
<td>AAA</td>
<td>2, 15, 43...</td>
</tr>
<tr>
<td>LAA</td>
<td>1, 5, 7, ...</td>
</tr>
<tr>
<td>GLL</td>
<td>3, 8, 34,...</td>
</tr>
<tr>
<td>VVN</td>
<td>4, 21, 25,...</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>
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
<table>
<thead>
<tr>
<th>Blast output</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{&gt;gi</td>
</tr>
<tr>
<td>Score = 255 bits (652), Expect = 4e-68</td>
</tr>
<tr>
<td>Identities = 140/322 (43%), Positives = 185/322 (56%), Gaps = 7/322 (2%)</td>
</tr>
<tr>
<td>Query: 221 SSATVSRNLNRFSTFVKSGGEAFVLGEASGFVKDGDKLCVLGYPYGPEWQENPYFPQCTI 280</td>
</tr>
<tr>
<td>Sbjct: 197 SSSSMKPLNKFPGFAKPGTEQYLL--AKQLAKPKPEKIPIIVGDYGPMWVYPTSTFDCVV 254</td>
</tr>
<tr>
<td>Query: 281 DDPTKQTTFKGMKSYISYKLVPTHTQPVPVHRYYKHFDWLYARLAEKF-PVISVPHLPEKQ 339</td>
</tr>
<tr>
<td>DP K +K G+KSYI Y+L PT+T V+ RYKHFDWLY RL KF I +P LP+KQ</td>
</tr>
<tr>
<td>Sbjct: 255 ADPRKGSKMYGLKSIEYQLTPTNTNRSVNHRYKHFDWLYERLLVFKFGSAIPISLPDKQ 314</td>
</tr>
<tr>
<td>Query: 340 ATGRFEEDFISKRRKGLIWWMNHMASHPVLAQCDVFQHFLTCPSTSTDEKAWKQGKRKAEK 399</td>
</tr>
<tr>
<td>TGRFEE+FI R + L WM M HPV+++ +VFQ FL + DEK WK GKRKA+</td>
</tr>
<tr>
<td>Sbjct: 315 VTGRFEEEFIKMRMRLEQLQAQMTRMCRHPCVESEVFQQFL---NFRDEKEWKTGKRKAER 371</td>
</tr>
</tbody>
</table>
Homology

- **Infernal** (web page, 2015c; Nawrocki and Eddy, 2013)
  - **Inference of RNA alignments**, 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)
- RNASnoopy (Tafer et al., 2010): predictor of targets for H/ACA snoRNAs:
  - computes interactions H/ACA - RNA thermodynamically 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)
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
ncRNA-Agents: architecture with 4 layers
ncRNA-Agents page

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

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

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

Maria Emilia M. T. Walter
non-coding RNAs
Prediction of IncRNAs: our SVM method

- **Data**
  - Ensembl transcripts with length more than 200 nucleotides
  - human (*Homo sapiens*)
    - IncRNAs: 29,020
    - protein coding: 92,425
  - mouse (*Mus musculus*)
    - IncRNAs: 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 IncRNA transcripts
  - test set
    - 7,333 protein coding transcripts
    - 7,333 IncRNA transcripts

- **mouse**
  - training set
    - 7,871 protein coding transcripts
    - 7,871 IncRNA transcripts
  - test set
    - 2,623 protein coding transcripts
    - 2,623 IncRNA 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 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:
  - IncRNAs of pig: 195 of 226 (86.2%)
  - IncRNAs of rat: 3,361 of 3,463 (97.0%)
  - IncRNAs of zebrafish: 3,881 of 3,940 (98.5%)
- Mouse Model:
  - IncRNAs of pig: 186 of 226 (82.3%)
  - IncRNAs of rat: 3,339 of 3,463 (96.4%)
  - IncRNAs 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)

Introduction
Methods to classify ncRNAs
Methods to identify ncRNAs
Databases of ncRNAs
Final Remarks
References

Many classes and families of ncRNAs (Rfam)

Family: mascRNA-menRNA (RF01684)

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

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 transcript called MALAT1 by an enzyme called RNase P. 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. Another small RNA that is homologous to mascRNA, called menRNA, is processed from another long ncRNA called MEN beta. 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.

See also[edit]
- Long noncoding RNA
- MALAT1
- MEN1

References[edit]

http://rfam.xfam.org/family/mascRNA-menRNA

Maria Emilia M. T. Walter

non-coding RNAs
Many classes and families of IncRNAs

- **NRED (ncRNA Expression Database) (Dinger et al., 2009)**
  - contains IncRNAs of the genomes of human and mouse
  - provides other information: known ncRNAs, evolutionary conservation, evidences for secondary structures and links to genomic contexts

- **DIANA-IncBase (Paraskevopoulou et al., 2013)**
  - provide interactions miRNA-IncRNA
  - contains two modules:
    - experimental: with detailed information about more than 5,000 interactions, among 2,958 IncRNAs and 120 miRNAs
    - prediction: with information of more than 10 millions of interaction, among 56,097 IncRNAs and 3,078 miRNAs

Maria Emilia M. T. Walter
Many classes and families of IncRNAs

- **IncRNADisease** (Chen et al., 2013)
  - contains IncRNAs associated to diseases
  - provides more than 600 input for IncRNAs related to diseases and 475 input for interactions of IncRNAs (251 IncRNAs and 217 diseases)

- **IncRNAdb** (web page, 2015d; Amaral et al., 2011; Quek et al., 2014)
  - contains a large volume of eukaryotic IncRNAs
  - 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 IncRNAs

- IncRNAtor (Park et al., 2014; web page, 2015e) provides (web interface) information for research of functional IncRNAs: annotation, sequence analysis, gene expression, protein interaction and phylogenetic conservation:
  - ncRNAs of six species (human, mouse, zebrafish, fruit fly, worm and yeast) collected from ENSEMBL, HGNC, MGI and IncRNAdb
  - 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-IncRNA, found by analyses of sequencing data through CLIP-seq and PAR-CLIP (interaction protein-RNA)
  - IncRNAs evolutionarily conserved among human and six other organisms to identify functional IncRNAs
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!

e-mail: mariaemilia@unb.br
Bibliography I


Bibliography X


Introduction

Methods to classify ncRNAs

Methods to identify ncRNAs

Databases of ncRNAs

Final Remarks

References

Bibliography XI


  http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snornera/home.

  http://bioinformatics.cenargen.embrapa.br/portrait/.


Wikipedia (2015). Non-coding RNA.