

Translational Bioinformatics


BIOST 2055

The world of small, non-coding RNAs

Takis Benos
Department of Computational & Systems Biology



Apr 4, 2012

Reading: handouts



Overview

- RNAi key players: siRNA and microRNA
- RNA folding
- microRNAs



© Benos BIOST2055 4-APR-2012 2

Translational Bioinformatics


BIOINF 2016

RNAi(nterference): some history

Takis Benos
Department of Computational & Systems Biology

March 21, 2012

Reading: handouts & papers



What is RNAi ?

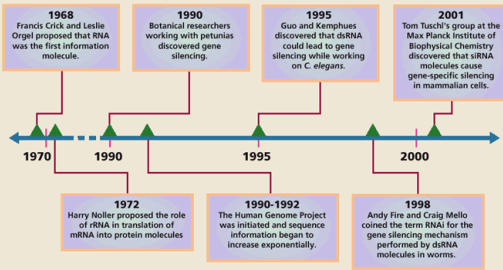
- RNAi is a cellular process by which the expression of genes is regulated at the mRNA level
- RNAi appeared under different names, until people realized it was the same process:
 - Co-suppression
 - Post-transcriptional gene silencing (PTGS)
 - Quelling



© Benos BIOS2055 4-APR-2012

4

Timeline for RNAi Discoveries




© Benos BIOS2055 4-APR-2012

Nature Biotechnology 21, 1441 - 1446 (2003)

BioCrim

From petunias to worms

- In the early 90's scientists tried to darken petunia's color by overexpressing the *chalcone synthetase* gene.
 - The result: 
- In 1995, Guo and Kemphues used anti-sense RNA to *C. elegans* par-1 gene to show they have cloned the correct gene.
 - Both sense and anti-sense par-1 gene produced the same (mutant) phenotype. (Hmml Hmml Hmml)
- Similar phenomena observed in fungus *N. crassa* and plant viruses
 - The phenomenon was shown to be post-transcriptional, but the mechanism remained unknown



© Benos BIOS2055 4-APR-2012

6


Acknowledgements

Some of the slides used in this lecture are adapted or modified slides from lectures of:

- [Gian Garriga](#), UC Berkeley

Other sources of information:

- Wikipedia
- WWW




© Benos BIOST2055 4-APR-2012 10

Translational Bioinformatics

BIOINF 2016


RNA Folding

Takis Benos
Department of Computational & Systems Biology
March 21, 2012
Reading: handouts & papers



Overview

- About the RNA and its structure
- RNA structure prediction
 - [Nussinov and Zucker algorithms](#)
 - [CONTRAFold](#)
 - [Prediction from multiple alignments](#)
- Case study: how RNA folding affects influenza adaptation?



© Benos BIOST2055 4-APR-2012 12

RNA structure

- RNA is a polymer of A, C, G, U
- Base pairs:
 -
 - Each base can only pair with one other base at a time

© Benos BEOST2055 4-APR-2012 13

RNA secondary structure

© Benos BEOST2055 4-APR-2012 14

RNA secondary structure prediction

- What makes RNA to fold?
- Problem definition: given an RNA sequence, find the set of base pairs that is "correct" or "optimal"
 - Maximum number of base pairs (Ruth Nussinov)
 - Minimum energy (Michael Zucker)
- Search problem: number of possible structures
 - 200 bases RNA: $>10^{50}$ possible base-paired structures
- Algorithm: dynamic programming
 - None of the above two can predict pseudoknots... (although they are really important)

© Benos BEOST2055 4-APR-2012 15

Base Pair Maximization - Dynamic Programming Algorithm

$S(i,j)$ is the folding of the subsequence of the RNA strand from index i to index j which results in the highest number of base pairs

Simple Example:
Maximizing Base Pair

$$S(i,j) = \max \begin{cases} S(i+1, j-1) + 1 & \text{[if } i,j \text{ base pair]} \\ S(i+1, j) \\ S(i, j-1) \\ \max_{i < k < j} S(i, k) + S(k+1, j) \end{cases}$$

Images: Sean Eddy

© Benos BEOST2055 4-APR-2012 16

Base Pair Maximization - Drawbacks

- Base pair maximization will not necessarily lead to the most stable structure
 - May create structure with many interior loops or hairpins which are energetically unfavorable
- Comparable to aligning sequences with scattered matches - not biologically reasonable

© Benos BEOST2055 4-APR-2012 17

Energy Minimization

- Thermodynamic Stability
 - Estimated using experimental techniques
 - Theory : Most Stable is the Most likely
- No Pseudoknots due to algorithm limitations
- Uses Dynamic Programming alignment technique
- Attempts to maximize the score taking into account thermodynamics
- MFOLD and ViennaRNA

© Benos BEOST2055 4-APR-2012 18

GENE EXPRESSION REGULATION: DATA AND ALGORITHMS

MSCBIO2020

microRNA genes and their targets

Takis Benos
Department of Computational & Systems Biology

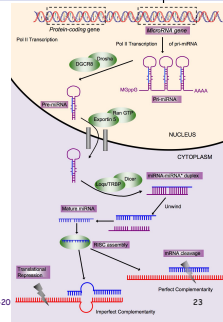
February 23, 2011

Reading: handouts & papers



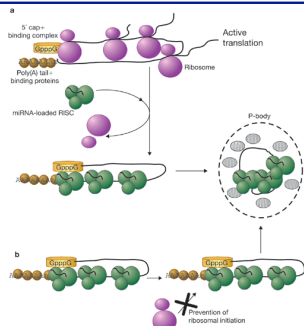
miRNA genes: a couple of things we know about them

- Size
 - 60-80bp pre-miRNA
 - 20-24 nucleotides mature miRNA
- Role: translation regulation, cancer diagnosis
- Location: intergenic or intronic
- Regulation: pol II (mostly)



© Benos BEOST2055 4-APR-20 23

miRNA method of action



24

Summary of Players

- Drosha and Pasha are part of the "Microprocessor" protein complex (~600-650kDa)
- Drosha and Dicer are RNase III enzymes
- Pasha is a dsRNA binding protein
- Exportin 5 is a member of the karyopherin nucleocytoplasmic transport factors that requires Ran and GTP
- Argonautes are RNase H enzymes

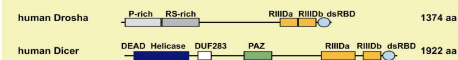


© Benoit BEOST2055 4-APR-2012

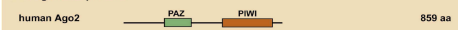
25

Players

A. RNase III type proteins



B. Argonaute proteins



C. dsRNA-binding proteins



D. DEAD-box helicases



© Benoit BEOST2055 4-APR-2012

26

miRNA function: some examples

	miRNA	Target genes	Function
<i>C. elegans</i>	<i>lin-4</i>	<i>lin-14</i> , <i>lin-28</i>	Early Developmental timing
	<i>let-7</i>	<i>lin-41</i> , <i>hbl-1</i> , <i>daf-12</i> , ...	Late Developmental timing
	<i>lgy-6</i>	<i>cog1</i>	L/R neuronal symmetry
	<i>miR-273</i>	<i>die-1</i>	
<i>Drosophila</i>	<i>Bantam</i>	<i>hid</i>	Programmed cell death
Mouse	<i>miR-196</i>	<i>Hoxb8</i>	Developmental patterning
	<i>miR-1</i>	<i>Hand2</i>	Cardiomyocyte differentiation & proliferation



27


microRNAs: some on-line resources

Databases

- mirBase: <http://mirbase.org>
- TarBase: <http://diana.cslab.ece.ntua.gr/tarbase>
- microRNA.org: <http://microRNA.org>

Target predictions

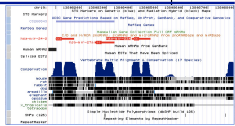
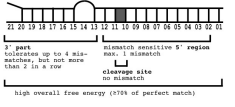

- TargetScan: <http://targetscan.org>
- PicTar: <http://pictar.mdc-berlin.de/>
- miRanda: <http://www.microrna.org/microrna/home.do>
- miRDB: <http://mirdb.org>



© Benos BEOST2055 4-APR-2012 28

miRNA computational predictions


- miRNA gene prediction
 - miRNA features
 - Gene prediction methods
- miRNA target prediction
 - Physical characteristics
 - Target prediction methods

© Benos BEOST2055 4-APR-2012 29

In the beginning, miRNA genes were identified...

- In the lab
 - Forward genetics: start from the mutant phenotype and look for the responsible gene
 - Very slow, inefficient (can only be applied to certain cases)
 - cDNA sequencing: size-fractionate RNA, clone, sequence
 - Slow, expensive
 - Deep sequencing of small RNAs (e.g., 454, Solexa)
 - Expensive, we do not know how many small RNA flavors exist
- In silico methods
 - Conservation-based
 - Clustering
 - SVMs



© Benos BEOST2055 4-APR-2012 30

1. miRNA gene prediction

- Computational prediction
 - Structural features (e.g., hairpin length, thermodynamic stability, etc)
 - Sequence features (e.g., nucleotide content, location, etc)
 - Evolutionary conservation
- Methodologies
 - Neighbor stem loop searches (*identify closely located stem loops*)
 - Gene-finding (*identify conserved genomic regions, then run MFold*)
 - Homology search (*direct BLAST searches*)



© Benos BIOST2055 4-APR-2012

31

1. miRNA gene prediction (cntd)

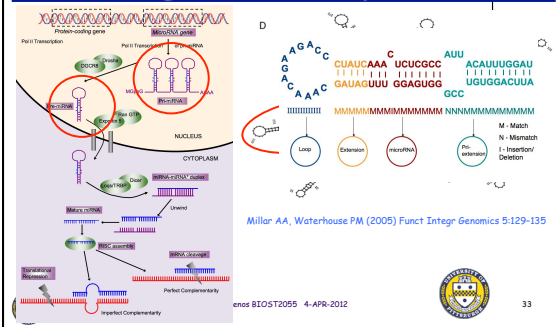
- Programs
 - miRseeker (Lai *et al.* 2003): assesses folding patterns of RNA sequences conserved between two *Drosophila* species
 - MiRscan (Lim *et al.* 2003): uses *RNAFold* to find hairpin structures in evolutionary conserved sequences (in worms)
 - Berezikov *et al.* (2005): uses *phylogenetic shadowing* together with other properties to identify miRNA genes
 - Kadri *et al.* (2009): uses *hierarchical HMM* with no evolutionary information



© Benos BIOST2055 4-APR-2012

32

miRNA biogenesis: stemloops



© Benos BIOST2055 4-APR-2012

33

Stemloop characteristics (species)



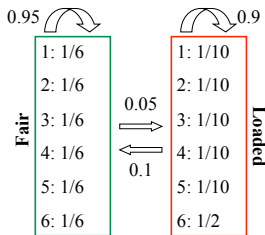
	Hairpin (bases)	Loop (bases)	Extension (bp mostly)	miRNA (bp mostly)	Pri-miR ext (bp mostly)
Mean (SD)					
Vertebrates	86.7 (13.8)	7.3 (3.5)	5.0 (3.4)	22.0 (0.9)	12.6 (7.0)
Invertebrates	91.8 (13.1)	7.9 (3.9)	5.8 (4.5)	22.2 (1.3)	13.8 (5.9)
Plants	119.5 (43.2)	6.8 (3.7)	22.8 (18.5)	21.3 (1.0)	12.5 (9.9)
Min - Max					
Vertebrates	55 - 153	3 - 22	0 - 34	16 - 26	0 - 50
Invertebrates	54 - 215	3 - 30	0 - 55	18 - 28	0 - 32
Plants	57 - 337	3 - 35	0 - 102	16 - 24	0 - 78



© Benos BEOST2055 4-APR-2012

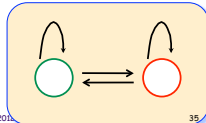
34

HMM example: the dishonest casino



Classification Problem
Given the model, parameters and a set of observations can we determine if they come from the fair or the loaded dice?

Q: what is hidden?

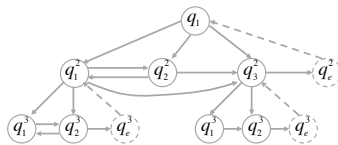


© Benos BEOST2055 4-APR-2012

35

Hierarchical hidden Markov models

- Internal States
- Production States
- End States
- Parameter Set λ



Fine et al., 1998; Machine Learning, 32, 41-62



© Benos BEOST2055 4-APR-2012

36

HHMMiR model based on miRNA stemloop characteristics

© Benos BIOS2055 4-APR-2012 37

Parameter estimation: Baum-Welch vs. MLE

	Sn (SD)	FDR (SD)
Baum-Welch	0.84 (0.19)	0.12 (0.06)
MLE	0.74 (0.14)	0.16 (0.08)

© Benos 38

Performance of HHMMiR across species (trained on human data)

Organism	Known hairpins	% predicted
<i>M. musculus</i>	422	74.7
<i>G. gallus</i>	147	89.1
<i>D. rerio</i>	334	88.3
<i>C. elegans</i>	131	85.5
<i>D. melanogaster</i>	143	93.0
<i>A. thaliana</i>	114	97.4
<i>O. sativa</i>	188	85.7
Total	1,479	85.1

© Benos BIOS2055 4-APR-2012 39

rna22: results (cntd)

Table 2. rna22's Estimates of the Number of MicroRNA Precursors for the Worm, Fruit Fly, Mouse, and Human Genomes

Genome	Number of MicroRNA Precursors Contained in the Seed Training Set	Number of MicroRNA Precursors that Are in the Training Set and Can Be Detected by mi22	Total Number of MicroRNA Precursors		Estimated Error when Predicting MicroRNA Precursors (≤ -15 kcal/mol) (≤ -18 kcal/mol)
			Detected by mi22 Including Newly Known Ones (≤ -15 kcal/mol)	Estimated Error when Predicting MicroRNA Precursors (≤ -15 kcal/mol)	
<i>C. elegans</i>	100	78 (78.0%)	350 (342)	$\leq 1%$ ($\leq 2%$)	
<i>D. melanogaster</i>	78	62 (79.5%)	604 (7,236)	$\leq 1%$ ($\leq 2%$)	
<i>M. musculus</i>	202	165 (81.7%)	>25,000 (>44,000)	$\leq 1%$ ($\leq 2%$)	
<i>H. sapiens</i>	178	150 (84.3%)	>25,000 (>50,000)	$\leq 1%$ ($\leq 2%$)	

Results are reported for two folding energy cutoffs: -15 kcal/mol and -18 kcal/mol.



rna22: evaluation

- **Advantages**
 - Predicts miRNA target genes w/o knowledge of the miRNA gene
 - No need for evolutionary conservation
 - Performs better when miRNA genes have multiple targets in the same mRNA
- **Disadvantages**
 - No consideration of the miRNA constrains *per se* (e.g., 5' "seed")
 - May miss target genes with one or few target sequences in their 3' UTR
 - Number of false positives cannot be estimated
 - Heuristics



Acknowledgements

Some of the slides used in this lecture are adapted or modified slides from lectures of:

- Brian Reinert, University of New Mexico