Theoretical Physics Methods for Computational Biology.

Fourth lecture

M. Caselle
Dip di Fisica Teorica, Univ. di Torino

Berlin, 08/04/2006

Plan of the lecture

- Identification of T.F. and miRNA binding sites in eukaryotes

  1] Introduction
  2] Our strategy for the TF binding site identification
  3] Example: Transcription factor binding sites in yeast
  4] Example: Transcription factor binding sites in human
  5] MiRNA target sites

- Graph theory approach to fragile sites characterization

References

- M Caselle, F. Di Cunto and P. Provero
  "Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes."
  BMC Bioinformatics 2002, 3:7,
- D. Corá, F. Di Cunto, P. Provero, L.Silengo and M. Caselle
  "Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs"
  BMC Bioinformatics 2004, 5:57,
- D. Cora', C. Herrmann, C. Dieterich, F. Di Cunto, P. Provero and M. Caselle
  "Ab initio identification of putative human transcription factor binding sites by comparative genomics"
  BMC Bioinformatics 2005, 6:110

1. Introduction.

Let use recall a few important features of the human genome and of the process of gene regulation that we have seen in the first lecture

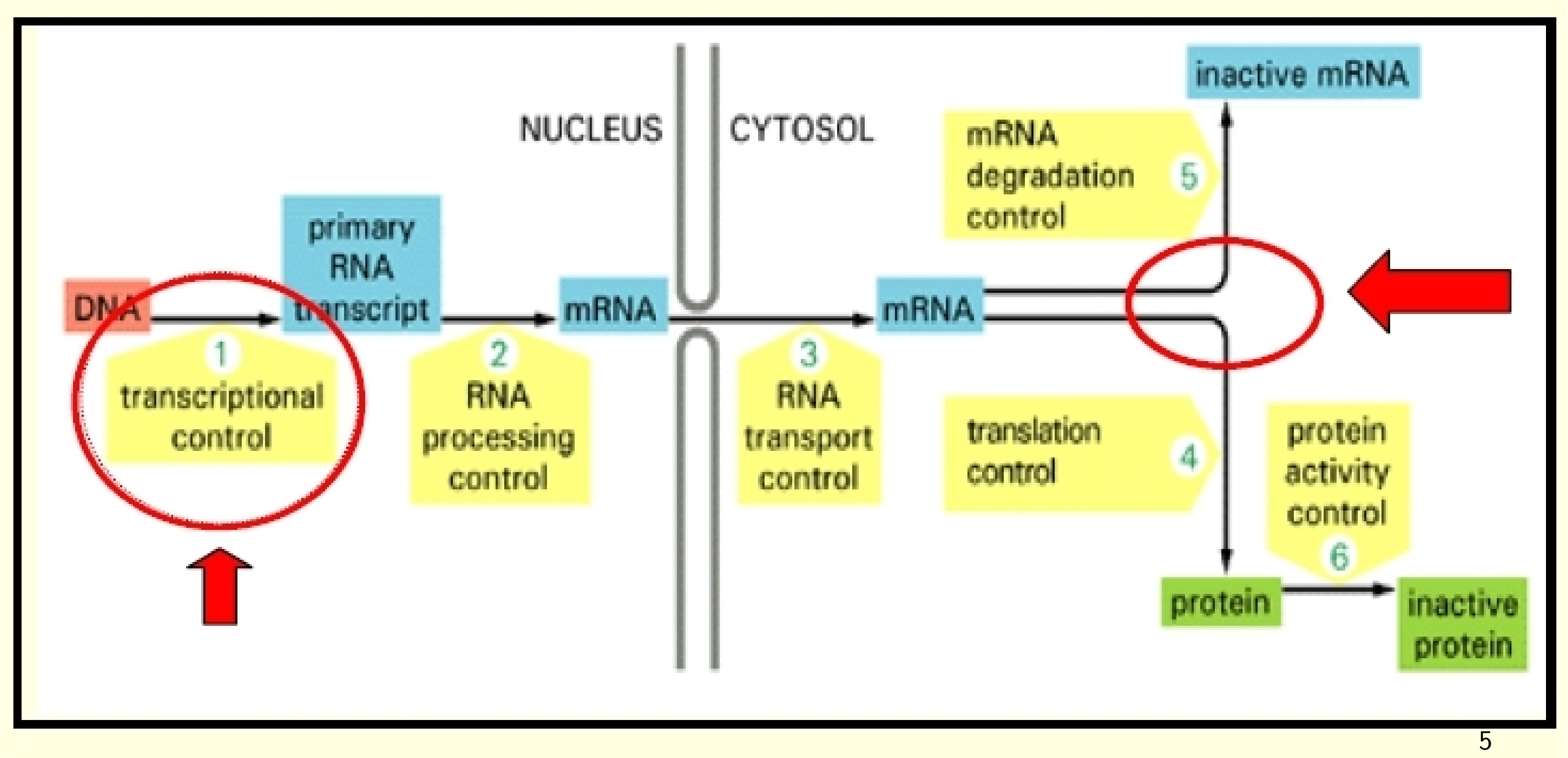
- The density of protein-coding and RNA-coding sequences becomes lower and lower as the complexity of the organism increases. It is rather high in Prokaryotes, low in S. Cerevisiae, very low in the human genome: most of DNA in the human genome is not coding and is expected to be involved in the regulation of gene expression

- Gene expression is tightly controlled and regulated:
  - All cells in the body carry the full set of genes, but only express about 20% of them at any particular time

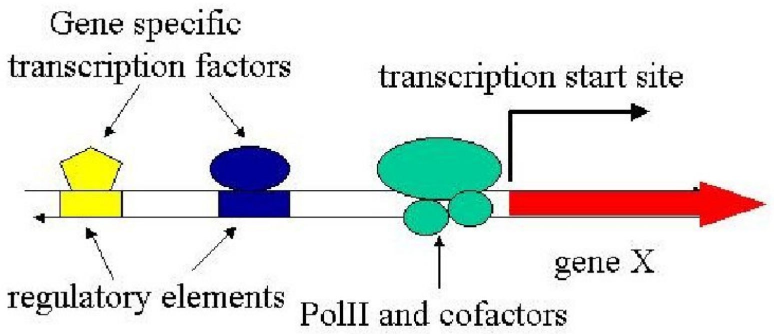
– Different proteins are expressed in different cells (neurons, muscle cells....) according to the different functions of the cell.

The most important example of such interactions is the transcriptional regulation of protein coding genes. Even if this is not the only regulatory mechanism of gene expression in eukaryotes it is certainly the most widespread one.

The goal of our research project (as of many others in the world) is to reconstruct these interactions by comparing existing biological information (like the coregulation of sets of genes) with the statistical properties of the sequence data.


NUCLEUS
CYTOSOL
DNA
primary RNA transcript
mRNA
mRNA
inactive mRNA
1 transcriptional control
2 RNA processing control
3 RNA transport control
4 translation control
5 mRNA degradation control
6 protein activity control
protein
inactive protein

Transcriptional regulation



TFs act by binding to specific, often short (5-10 bp) DNA sequences in the upstream noncoding region of genes.

T.F.'s themselves are proteins produced by other genes.

The Genome is a complex network of interactions between genes and their products This network pattern is ubiquitous in Postgenomic biology

The problem.

However, computational detection of regulatory sites is a difficult task, in particular in eukaryotes:

- the consensus sequences recognized by transcriptional factors are generally rather short (5-20 bp)
- they can be quite variable
- they are in general dispersed over large distances
- they are generally active in both orientations

A simple study of relative frequencies of sequences can be meaningless

Luckily we have a few tools to attack the problem:

- Binding sites are often overrepresented. One can use this to separate the signal (binding site) from the noise (background upstream sequence)

- Binding sites are often evolutionary conserved. One can use comparative genomics to recognize them.

- Genes which share the same functions may also share the same regulatory mechanisms. One may use microarray experiments or functional annotations to identify binding sites.

The standard strategy :

Coregulated genes.

step 1 Identify a set of genes experimentally known or presumed to be coregulated (for example because they are involved in the same biological process or because they show similar expression profiles in microarray experiments).

step 2 Find which short motifs are overrepresented in their upstream region, compared to suitably defined background motif frequencies that take into account the basic features of non-coding DNA of the organism under study. These motifs are likely to be involved in the coregulation of the genes in the set.

See for instance:

- van Helden J, André B, Collado-Vides J,

  Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 1998, **281**, 827-842.

- Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM:

  Systematic determination of genetic network architecture. *Nature Genetics* 1999, **22**, 281-285.

Main problems:

- Unsupervised clustering
- Motif's variability
- Same expression without coregulation

Our Proposal

Reverse the procedure!

M Caselle, F. Di Cunto and P. Provero,

Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes.

BMC Bioinformatics 2002, **3**:*7*

first step Grouping of genes based on the motifs that are overrepresented in their upstream regions. To each possible word w we associate the set S_w of all the genes in whose upstream region the word w is overrepresented

second step Select those sets which show some kind of functional characterization using microarray experiments or Gene Ontology annotations.

- Microarray: For each set S_w we compare the expression distribution within the set with the genome wide one (using for example Kolmogorv-Smirnov test).

- Gene Ontology: For each set S_w we compute the prevalence of all GO terms among the annotated genes in the set, and the probability that such prevalence would occur in a randomly chosen set of the same size:
  - hypergeometric distribution to assess the significance of the intersection
  - evaluation of false discovery rate through comparison with randomly generated gene sets (using only the best p-value for each set as criterion for the comparison)

The words which survive this analysis are candidates to be binding sites.

The Gene Ontology Consortium "Gene Ontology: tool for the unification of Biology" Nature Genetics **25** (2000) 25.

Overrepresented words
in the upstream regions

Many binding sites are effective only when repeated many times in the upstream region of the gene they regulate.

Example: the word GATAAG—CTTATC is a known binding factor for nitrogen-regulated genes: Examine the 500 bp's upstream of two of them.

>YPR138C upstream sequence, from -500 to -1

```
TCCACCTTATCTCGGCGCCAAATCCTTATC
TCTCGTAGCTGGTTTGCCCGCGATAAGGCG
GGCGAGTTATTTTGAAGTTTTCCATAAACT
GGTTTTCCATCTCGAGGTTTTTCCTCGCTT
TCCACGCTATGACCCTTTTTAGTTAAGGTA
CCCGATGGCATACTTTATATATTATATATA
TATGTTAAGTTAATATGTTTTAGCAGATTT
GATATGCTGATATGCAGCACGGACTTTCCC
TCTCCTTGTCTTATCGCATCTTATCGCAAC
AATTTGATAGATATCTTCTCCCTTTCCTAT
CTTGTAGAATAAGGTTGTGTGCTTTGAGTC
TGATAGCCGTCTTCTTTCGGTCGCTTCTTC
TCTCTTTTGGTTCTTTGATTGTCTATTACA
```

>YIR028W upstream sequence, from -500 to -1

```
ATTCTCGGGTCTAATGTGGCTCGAGGGTAT
CTCTTATCGGTATTACTTTCTTATCAATGA
AAAATTTCTGCCAGGGAAAATGCGCCCGCT
TTTTTTCCGGCCATCCTTACTCGCTGTCGC
ATACAAAATAGCGCCTCTAATCTAGTTGCG
ATAAGGAATGTGTATGTGTAATTGAAGATC
CAGGATGTTTTCCTTTTCAGGGAGATGAGA
AGGAATAATAGGATGGATTGACCGCTTTGC
TGTCACGTCGATAAGGTTCCTTTAAAAATT
GTGTCCAATGATTAGCATAGAGAGGTAGAG
TATCAGAGAAACAAGTTTGTAATCGAGAAA
CTTGATCTGCTAGTGTTGAGCATAGAAGGC
TAGGAAAACATGGGGAAGAAAAAAAAAGTA
```

The sets $S(\text{word})$

- For each word (5 to 8 bp's) compute the frequency in the upstream sequences of the whole genome considered as a single sample: these will be our reference frequencies.

- Then count occurrences of the word in the upstream region of each gene separately.

- If the number of occurencies of the word in the upstream region of gene G is statistically significant (compared to a binomial distribution based on the above reference frequencies), then the gene G belongs to the set $S(\text{word})$.

Choices in our study on yeast:

- *upstream sequences length: 500 bp*
- *probability cutoff* $P = 0.01$

The Gene–Ontology filter.

For each set $S(m)$ we computed the prevalence of all Gene Ontology (GO) terms among the annotated genes in the set, and the probability that such prevalence would occur in a randomly chosen set of genes of the same size.

For a given GO term t let $K(t)$ be the total number of ORFs annotated to it in the genome, and $k(m,t)$ the number of ORFs annotated to it in the set $S(m)$. If J and $j(m)$ denote the number of ORFs in the genome and in $S(m)$ respectively, such probability is given by the right tail of the appropriate hypergeometric distribution:

$$P(J, K(t), j(m), k(m,t)) = \sum_{h=k(m,t)}^{\min(j(m),K(t))} F(J, K(t), j(m), h)$$

where

$$F(M, m, N, n) = \frac{\binom{m}{n} \binom{M-m}{N-n}}{\binom{M}{N}}$$

In this way a P-value can be associated to each pair made of a motif and a Gene Ontology term.

False discovery rate

Problem:

Given the huge number of P-values that we compute (in principle equal to the number of GO terms multiplied by the number of words analysed) it is clear that very low P-values could appear simply by chance.

The usual way of dealing with this issue, that is the Bonferroni correction, is not appropriate, because due to the hierarchical nature of the Gene Ontology annotation scheme, the P-values we compute are very far from being independent from each other.

Our proposal

We randomly generated a large number N_R of sets of genes comparable in size to the typical size of the sets associated to the motifs and ranked the random sets based on their **best** P-values.

In this way we can determine a false discovery probability $p_f(C)$ as a function of the cutoff on P-values C

Warning:

The lower is the FDR required, the higher is the precision required in determining the function $p_f(C)$ and hence the number N_R of sets to be generated randomly. For instance a FDR of 0.01 requires the generation of 3.5×10^6 randomly chosen sets.

3. Example: Yeast

Identification of TF binding sites in yeast using Gene–Ontology

Output of the analysis:

- With the false discovery rate set at 0.01 we find a total of 108 associations between 80 different words (of 5-8 letters) and 41 Gene Ontology terms.

- The words can be organized in 12 different groups. Within each group the motifs are very similar to each other and are associated to the same or to very similar Gene Ontology terms. For each group we construct a consensus sequence ("motifs") by aligning the words.

Validation:

- Comparison with known TF's and binding sites (Transfac + literature survey)
- Comparison with the genome wide ChIP experiment of: T.I. Lee et al., Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science 298, (2002) 799.*

Results:

- All the motifs we find correspond to known binding sites. (No false positive!)

- For some of the motifs we are able to
  - refine the putative binding sequences.
  - identify candidates for combinatorial regulation (example: PAC and RRPE))
  - Refine the functional annotation of already known TF's
  - identify new potential targets of known TF's (example: Hcm1p)

D. Corá, F. Di Cunto, P. Provero, L.Silengo and M. Caselle, *BMC Bioinformatics 2004*, **5**:*57*

motif	C	F	P	TF
TGAAAC	-	-	sexual reproduction	DIG1 STE12
TGAAACA	-	-	sexual reproduction	DIG1 STE12
TGAAACA				
ACTGTG	-	-	sulfur amino acid transport	MET4
TGTGGC	-	-	sulfur metabolism	MET4 MET31
ACTGTGGC				

Table 1: *Two examples of motifs with significant intersection with ChIP data*

D. Cora', C. Herrmann, C. Dieterich, F. Di Cunto, P. Provero and M. Caselle, *BMC Bioinformatics 2005*, **6**:*110*

4. Example: Human

The extension of our algorithm to the human genome is not straightforward. At least 15.000 bp long upstream regions must be taken into account leading to a very small signal to noise ratio.

It is mandatory to perform a comparative analysis selecting only those parts of the upstream regions which are conserved between men and mouse.

This can be done using the CORG database:

C. Dieterich et al., CORG: a database for comparative regulatory genomics. *Nucleic Acid Res.*, **31**, *(2003) 374.*

The CORG database.

CORG is a collection of conserved sequence blocks in the non-coding, upstream regions of orthologous genes from man and mouse.

These blocks are obtained by searching statistically significant local suboptimal alignments of 15kb regions upstream of the translation start site.

The database contains more than 10,000 pairs of orthologous genes. The alignments were obtained using the Waterman-Eggert algorithm. We used two different choices of the PAM matrix: PAM1 and PAM10 to test the robustness of the results.

The two releases are very different:

- PAM1
  - number of genes in the database: 10999
  - mean number of conserved blocks for gene: ~ 20
  - mean length of the union of conserved blocks: ~ 500
  - number of genes with a GO annotation 6187
- PAM10
  - number of genes in the database: 12943
  - mean number of conserved blocks for gene: ~ 40
  - mean length of the union of conserved blocks: ~ 900
  - number of genes with a GO annotation 7260

Results.

In the PAM10 case, out of the 43250 possible words of 5,6,7 and 8 letters

- 154 different words survive the G–O filter
- 331 words survive the Microarray filter
- the intersection between the two sets is 109 words which corresponds to a p–value e^{-201}
- similar results are obtained with PAM1. Despite the fact that the PAM1 and PAM10 CORG databases are very different our results seems to be very robust: most of the words are present in both releases.

Clustering of words.

Due to the larger amount of words and to the higher motif's variability, clustering of words is more delicate than in the yeast case. To decide if two words belong to the same motif we make a two steps analysis.

- First step: we check if at least one of the following conditions is met:
  - at least one GO term is significant for both motifs
  - there is at least one time point in the cell cycle MA experiment in which both motifs are simultaneously significant.
  - the intersection of the two sets of genes (labeled by the two words that we are testing) is statistically significant.

- Second step: we check if an alignment can be found between the two words with no gaps, at least 4 bases correctly aligned and at most 1 mismatch.

Validation.

Comparing our finding with the data collected in the Transfac database we were able to recognize some well known TF's.

Example: NF–kB

motif	C	F	P
GGAAATTC	-	chemoattractant	-
GGRAAKTCCC	Transfac consensus		

Table 2: *The putative NF–kB motif.*

Example: E2F

motif	C	F	P
TTTCGCGC	-	-	DNA replication initiation
TTTSGCGC	Transfac consensus		

Table 3: *The putative E2F motif.*

Example: A putative new motif

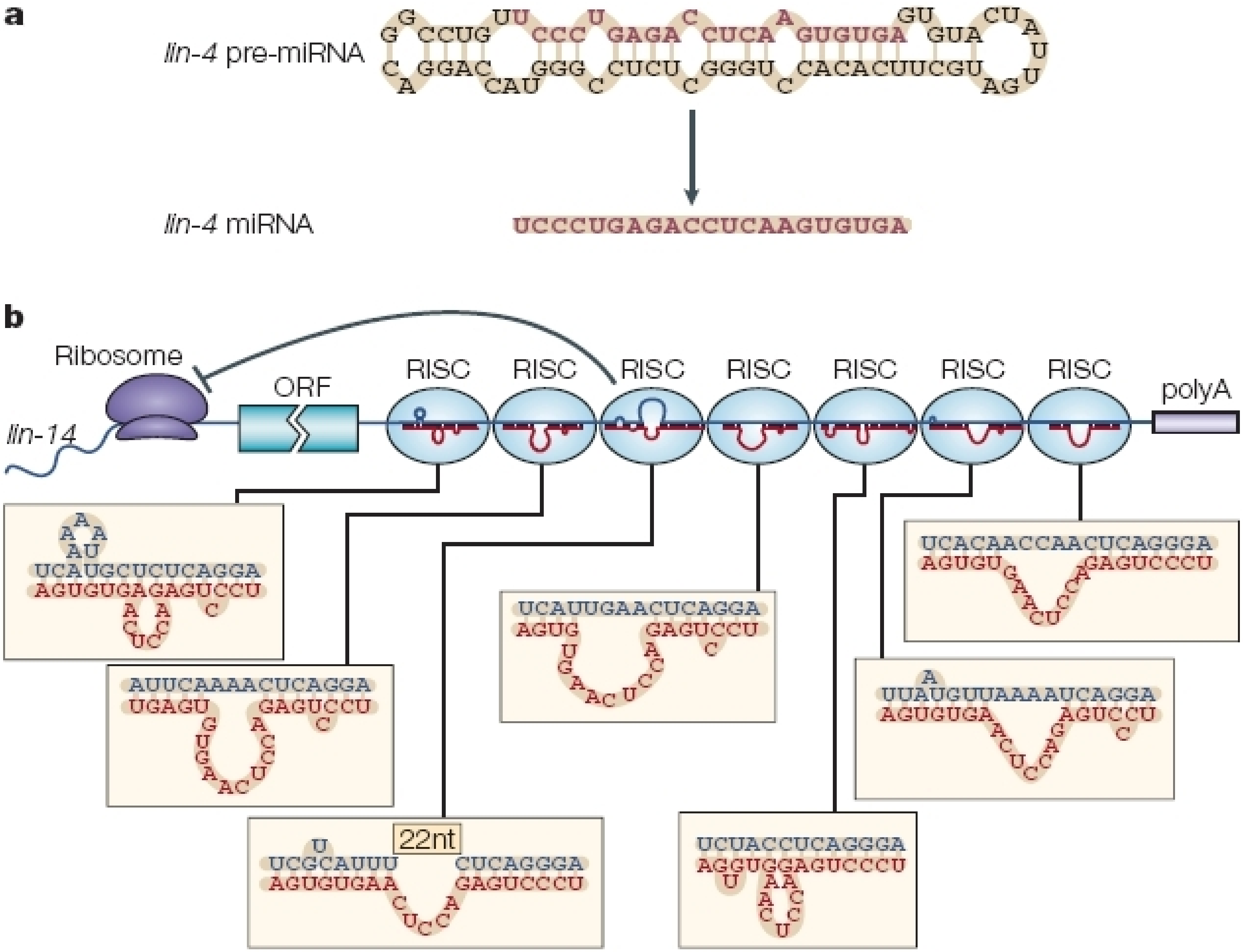
motif	C	F	P
AATGTTG	Golgi lumen	-	-
TGTTGA	Golgi lumen	-	-
ATGTTGA	Golgi lumen	-	-
TTATGTA	Golgi lumen	-	-
TWATGTTGA			

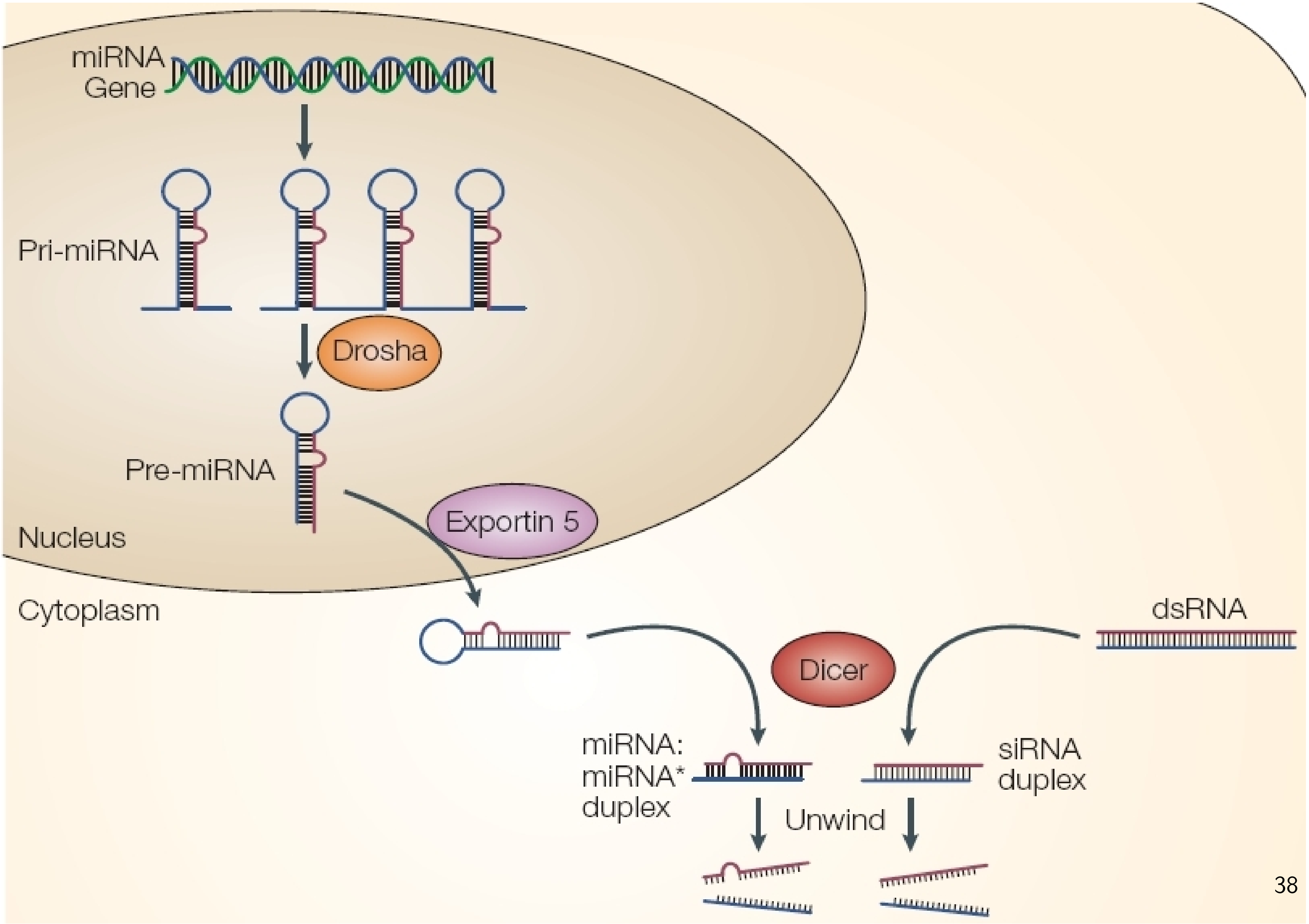
Table 4: *A putative motif with no reference in Transfac.*

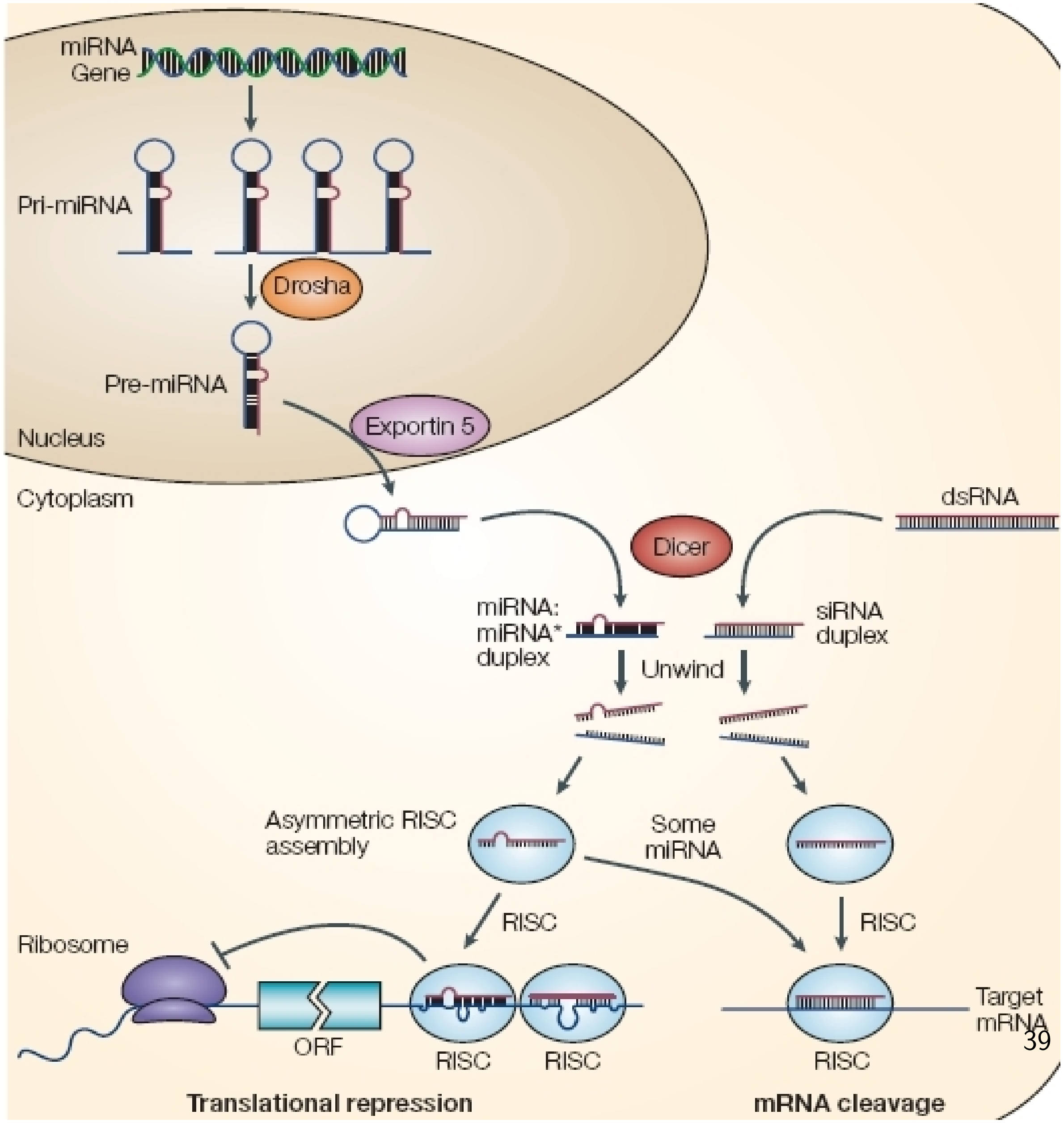
5. miRNA.

Gene expression can be regulated at many of the steps in the pathway from DNA to RNA and protein.

MicroRNAs(miRNAs) are a family of 21 - 25 nucleotide small RNAs that negatively regulate gene expression at the post-transcriptional level.


a
lin-4 pre-miRNA
G UU U C A GU CU
G CCUG CCC GAGA CUCA GUGUGA GUA A
C GGAC GGG CUCU GGGU CACACUUCGU U
A CAU C C C AG U
lin-4 miRNA
UCCCUGAGACCUCAAGUGUGA
b
Ribosome
ORF
RISC
RISC
RISC
RISC
RISC
RISC
RISC
polyA
lin-14
A
A A
AU
UCAUGCUCUCAGGA
AGUGUGAGAGUCCU
A A C
C C
UC
AUUCAAAACUCAGGA
UGAGU GAGUCCU
G A C
U C
G C
A C
AC U
22nt
U
UCGCAUUU CUCAGGGA
AGUGUGAA GAGUCCCU
C A
UCC
UCAUUGAACUCAGGA
AGUG GAGUCCU
U A C
G C
A C
AC U
UCUACCUCAGGGA
AGGUGGAGUCCCU
U AA
A C
C C
U
UCACAACCAACUCAGGGA
AGUGU GAGUCCCU
G A
A C
A C
CU
A
UUAUGUUAAAAUCAGGA
AGUGUGA AGUCCU
A G C
C A
U C
C


miRNA
Gene
Pri-miRNA
Drosha
Pre-miRNA
Nucleus
Exportin 5
Cytoplasm
dsRNA
Dicer
miRNA:
miRNA*
duplex
siRNA
duplex
Unwind


miRNA Gene
Pri-miRNA
Drosha
Pre-miRNA
Exportin 5
Nucleus
Cytoplasm
dsRNA
Dicer
miRNA:
miRNA*
duplex
siRNA
duplex
Unwind
Asymmetric RISC
assembly
Some
miRNA
RISC
RISC
Ribosome
ORF
RISC
RISC
Target
mRNA
RISC
Translational repression
mRNA cleavage

The main ingredients of the problem

- miRNAs inhibit translation by pairing on suitable binding sites in the 3'-UTRs.

- Perfect complementarity or G-U pairing between the target 3-UTR and the first nucleotides 1-7 or 2-8 of miRNA is needed.

- Additivity of the inhibitor function.

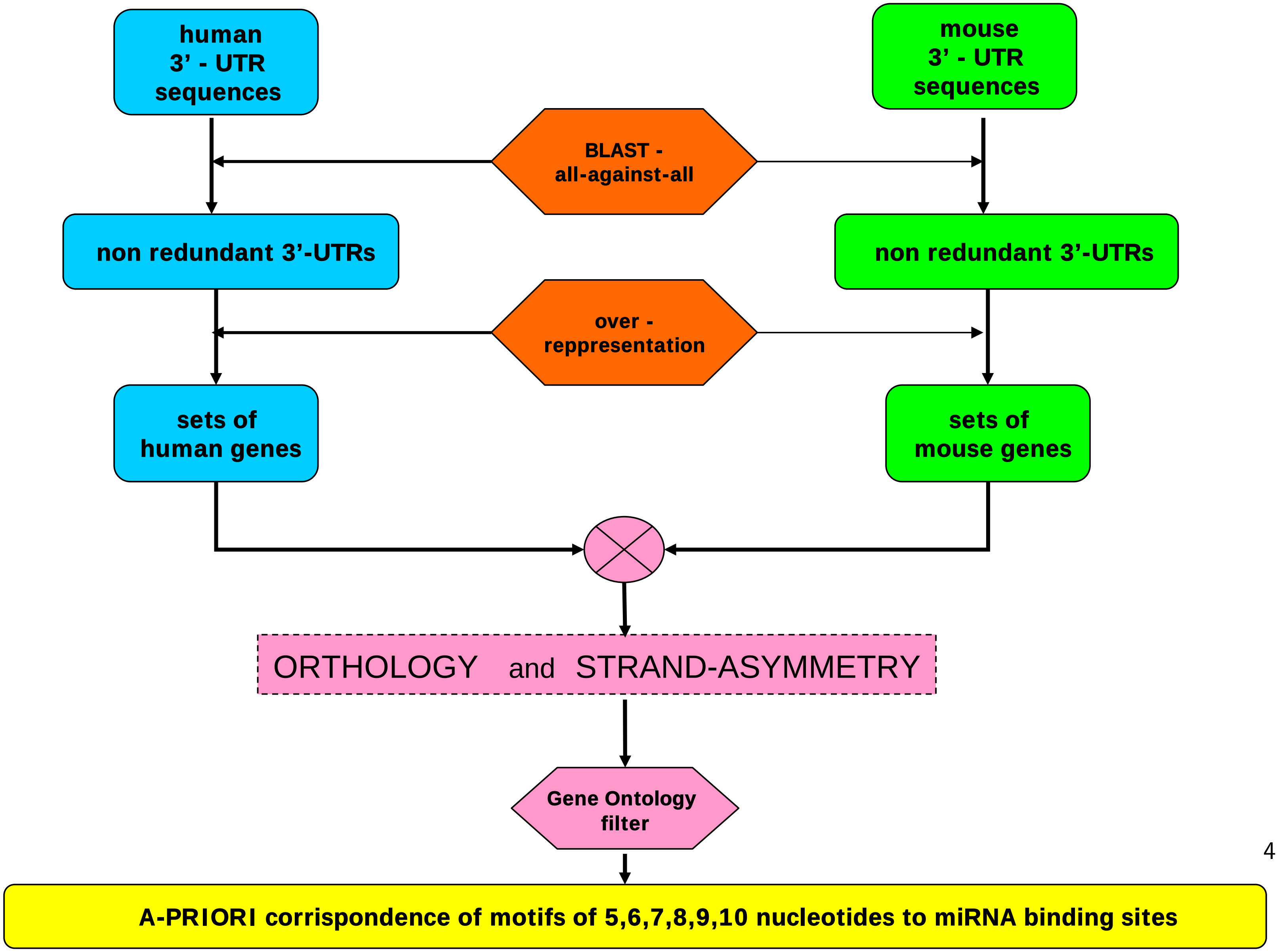
- Evolutionary conservation of miRNA target sites.

- Remarkable tissue specificity.

our proposal

Use a mixture of the following ingredients

- word overrepresentation
- mouse-human conservation of overrepresented words
- strand asymmetry
- Gene-ontology filter



human 3' - UTR sequences

mouse 3' - UTR sequences

BLAST - all-against-all

non redundant 3'-UTRs

non redundant 3'-UTRs

over - reppresentation

sets of human genes

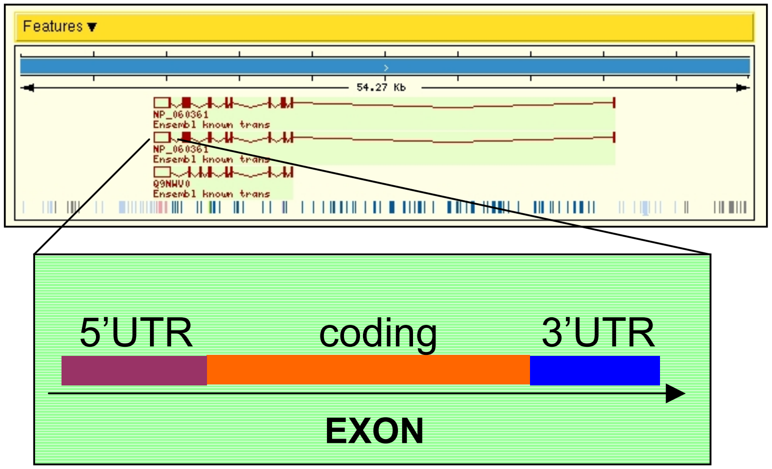
sets of mouse genes

ORTHOLOGY and STRAND-ASYMMETRY

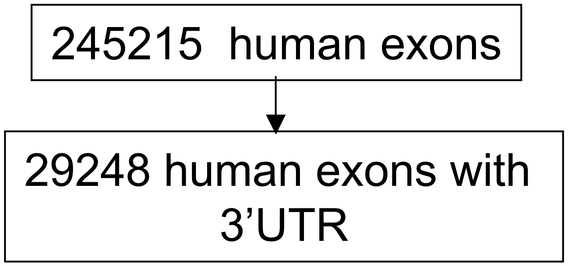
Gene Ontology filter

A-PRIORI corrispondence of motifs of 5,6,7,8,9,10 nucleotides to miRNA binding sites

We need an advanced retrival system to automatically download all the 3’-UTR region of a given genome (human / mouse ...), and the corresponding annotations.

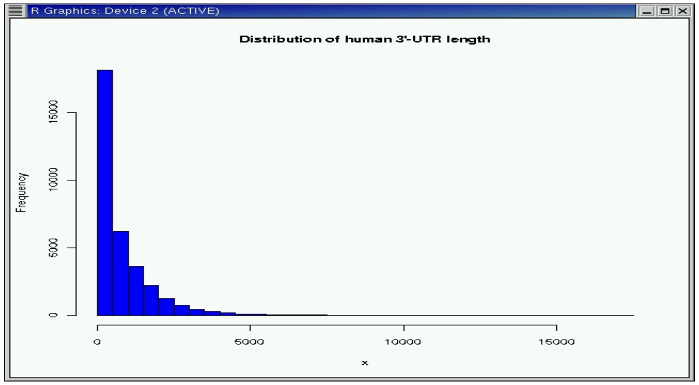


Direct query the EnsEMBL mySQL via perl API.



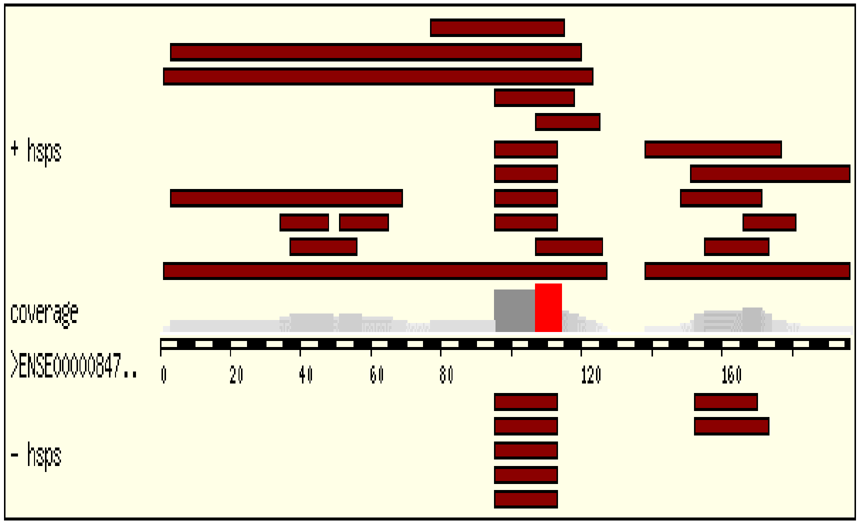
245215 human exons

29248 human exons with 3’UTR



We expect a lot of redundancy into human / mouse DNA sequences.

We use BLAST to list all the pairs of nearly identical 3’UTR sequences (Blast P-value less than 10e-40). Then we used these results to form clusters of nearly Identical 3’UTR regions. Finally, we retained for further analysis only one gene per group, chosen at random.



24283 non-redundant 3’UTR sequences, mapped into 15988 unique ENSG gene ids.

For each word *S* of 5, 6, 7, 8, 9, 10 nucleotides we construct the set of all genes in whose 3’UTR region the word *S* is <u>overrepresented</u>.

- null hypothesis: a random binomial distribution always separating a word from its reverse complement.
- discard overlapping motifs (ATTTT vs TTTTG)
- separated strands

> 3’ UTR sequence
ACTTTTTTACCCTCGTGTGTT
GCAGACTTTTTGCCACTTTTA
AAACGCTGACAATTCGACCC
TTTCCAATCTCTCAAAAGTTT
CGACGAGCTGTACAACCCCC
CCCCC ……………………..

$$b_g(w) = \sum_{n=n_g(w)}^{L_g(w)} \binom{L_g(w)}{n} p(w)^n \left[1 - p(w)\right]^{L_g(w)-n}$$

Binomial P-value

Human sets were then compared to their mouse orthologs with respect to two properties: we selected the DNA motifs showing in the 3'-utr region:

- conserved overrepresentation

- preferential strand asymmetry

The DNA motifs thus selected were further investigated through the analysis of the sets of genes in whose 3'-utr region they are overrepresented, using the Gene Ontology and the mouse phenotype annotation system.

conserved overrepresentation

We select for further analysis only the sets that shown a number of common orthologues paired genes greater than expected by chance according an hypergeometric pvalue evaluation.

The hypergeometric pvalue for each set $F(M, m, N, n)$ was defined, focusing on a certain sets:

M $=$ num tot of human genes with an orthologue

m $=$ num of human genes in the considered set with an orthologue

N $=$ num of human genes in the considered set with an orthologue in the corresponding mouse set

n $=$ num the intersection of m and N

preferential strand asymmetry

We studied the distribution of the variable:

$$\delta N = N_+(w) - N_-(w)$$

being $N_+(w)$ the number of occurency of the word w on the $+$ strand and $N_-(w)$ the number of occurency of the word w on the - strand.

In a pure random case, the variable $\frac{\delta N}{\sqrt{N}}$ is a gaussian with mean 1 and variance 1.

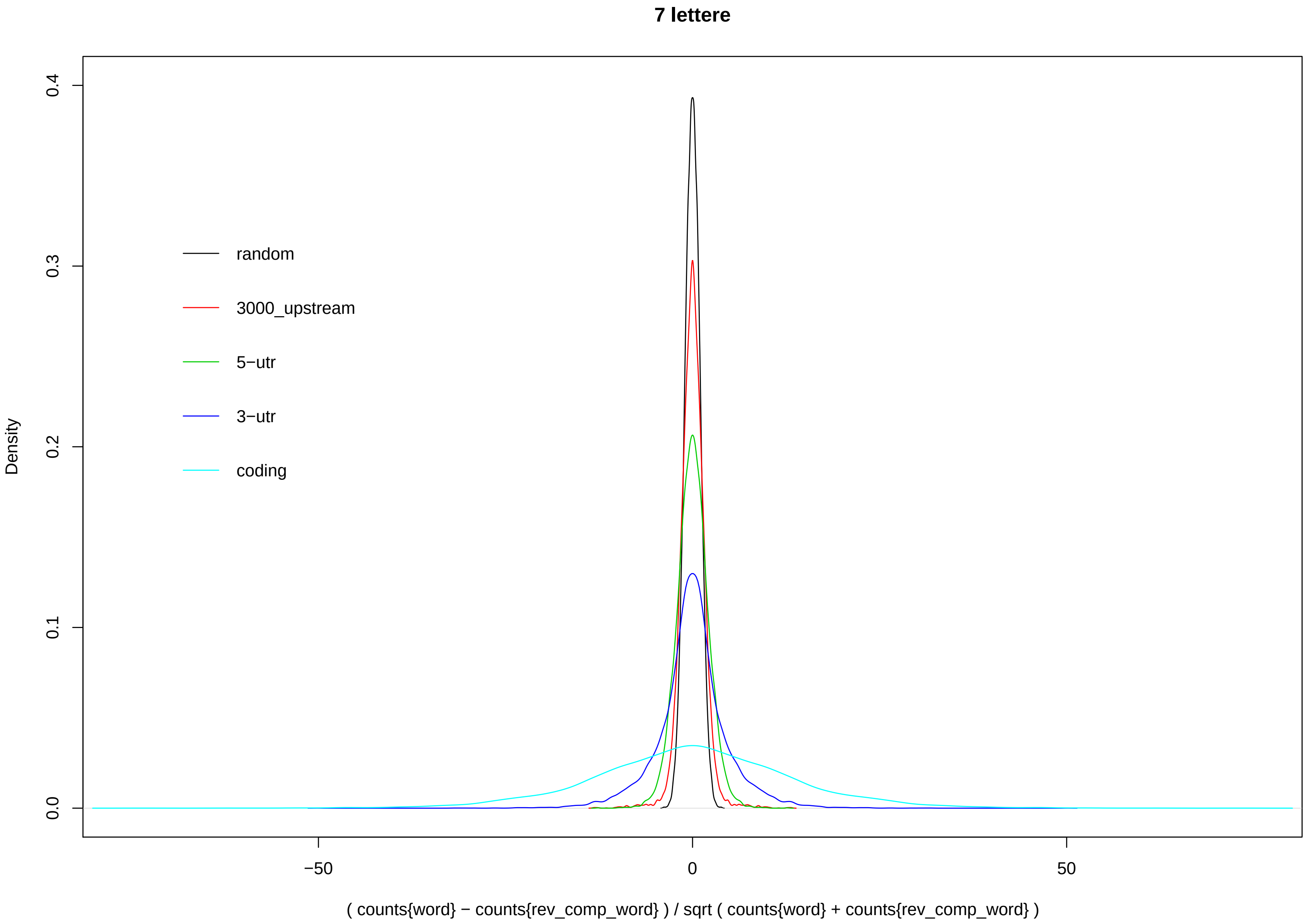
example of results

	num. tot. words (sets dim)	num. words cons. overr.	num. words asymm. $+9\ \sigma$
L=5	1024	362 / 173	197 / 100
L=6	4096	512 / 119	374 / 85
L=7	16384	604 / 75	369 / 30
L=8	65533	1964 / 83	199 / 4
L=9	260542	6285 / 90	37 / 0
L=10	973139	12382 / 69	9 / 0

In red: number of words validated with a known miRNA present in the microRNA-Registry (total 328 known miRNA).

	num. tot. words (sets dim)	num. words cons. overr. and asymm. $+9\ \sigma$
L=5	1024	362 / 121
L=6	4096	512 / 155
L=7	16384	604 / 98
L=8	65533	1964 / 64
L=9	260542	6285 / 14
L=10	973139	12382 / 2

In red intersection between words conserved-overrepresented and with strand asymmetry $+9\ \sigma$.


7 lettere
Density
0.0
0.1
0.2
0.3
0.4
random
3000_upstream
5−utr
3−utr
coding
−50
0
50
(counts{word} − counts{rev_comp_word}) / sqrt (counts{word} + counts{rev_comp_word})

miRNA References

- He and Hannon “MicroRNA: small RNAs with a big role in gene regulation.”
  Nat Rev. Genet. 2004 Jul;5(7):522-31.
- Bartel, D “MicroRNAs: Genomics, Biogenesis, Mechanism, and Function.”
  Cell. 2004 Jan 23;116(2):281-97.
- John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS “Human MicroRNA targets.”
  PLoS Biol. 2004 Nov;2(11):e363. Epub 2004 Oct 05.
- Griffiths-Jones S. “The microRNA registry.”
  Nucleic Acids Res. 2004 Jan 1;32 Database issue:D109-11.
- http://www.microrna.org/

Graph theory approach to fragile sites characterization

Fragile sites are regions of the human chromosomes which are particularly prone to genomic instability: breakage, sister chromatide exchange and recombination...

Common fragile sites are present in all individuals, they are conserved between man and mice thus they are expected to have a functional role.

In normal situations they are not dangerous for the organism but they seem to be particularly expressed in cancer cells.

Their functional role and their biology is poorly understood. also their connection with cancerogenesis is still an open issue.

Our proposal

Look for anomalous correlations in the pattern of CFS's expression.

Organize these correlations in a network and extract (if they exist) connected components and/or communities.

Analyze the gene content of correlated CFS using Gene-Ontology looking for functional signatures

A.Re, D. Cora', A. Puliti, M.Caselle and I.Sbrana

Correlated fragile site expression allows the identification of candidate fragile genes involved in immunity and associated with carcinogenesis

Submitted to BMC Bioinformatics

Main results

- Impressive correlations indeed exist in the pattern of individual expression of CFS: if a CFS is expressed in a given patient one could predict which other CFS are simulatneously expressed witha good degree of confidence.

- The network of these correlated CFS has various connecetd components, among them a "percolating" one. These components have a very small probability to appear in a random graph and thus may denote some hidden functional relationship among CFS.

- The GO analysis of the gene content of these connected CFS shows a remarkable enrichment of terms related to the immune response and of genes involved in the replication process.

A possible interpretation of this result is that breakage at fragile sites could be protective against cancer.

According to this picture breaks would represent a signature of replication stress and would activate the DNA damage check points, leading to cell-cycle arrest or apoptosis to ensure genomic integrity

At the same time it has been recently realized that there is a strong connection between the immune response and processes that regulate genome integrity. DNA damage rensponse, besides arresting the cell cycle and triggering apoptosis may partecipate in alerting the immune system to the presence of potentially dangerous cells, thus triggering an immune response against them.