Fragment based detection of ncRNAs

Marcus Lechner

Philipps-University

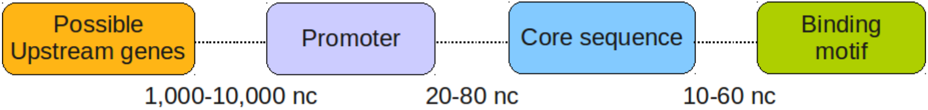
Bled, 2011-02-14

Table of contents

1. Introduction
2. Fragment based approach
3. Outlook

Introduction

- detection of ncRNAs is an important task in genome annotation
- many ncRNA classes and families known
- finding members in other species can be rather complex
- in most cases a combination of tools is necessary
- candidates have to be scored



Complexity of ncRNA detection - Features

- sequence similarity
- secondary structure
- promoter data
- terminator data
- protein interaction sites
- RNA interaction
- synteny
- specific distances

Complexity of ncRNA detection - Tools

- sequence: blast, GotohScan
- motifs: RNABOB, fragrep, RNAmotif
- secondary structure: RNA Vienna Package, RNAshape
- RNA-interaction: Petcofold, RIPalign
- terminators: TranstermHP
- covariance models: Infernal

Complexity of ncRNA detection - Result

Bioinformatician

- build a basically new pipeline for each RNA class
  1. evaluate features
  2. find suitable tools
  3. plug them together
  4. combine and assess results
  5. score the results
- time intensive
- manual labor

Complexity of ncRNA detection - Result

Bioinformatician

- build a basically new pipeline for each RNA class
  1. evaluate features
  2. find suitable tools
  3. plug them together
  4. combine and assess results
  5. score the results
- time intensive
- manual labor

Biologist

Fragment based approach

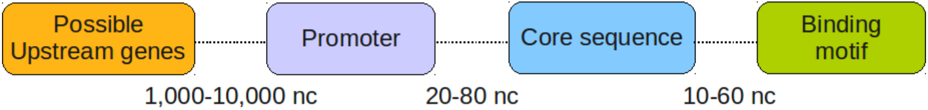
Deals with

1. evaluate features
2. (find suitable tools)
3. plug them together
4. combine and assess results
5. score the results

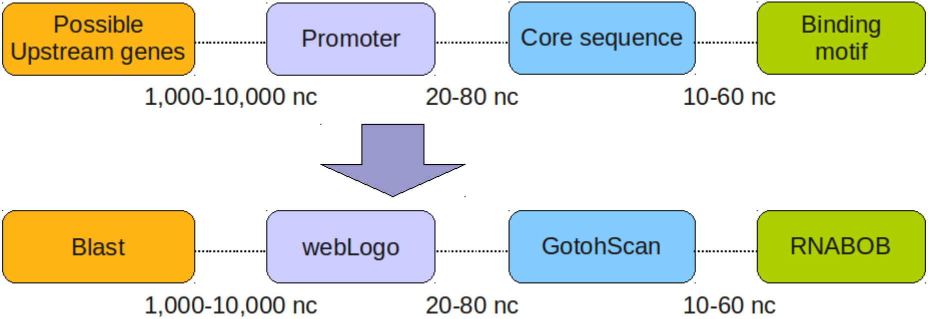
Applicable for

- bioinformatician with expert knowledge about the tools
  - command line tool
  - all-in-one online tool
- scientists with limited background in computer science
  - all-in-one online tool

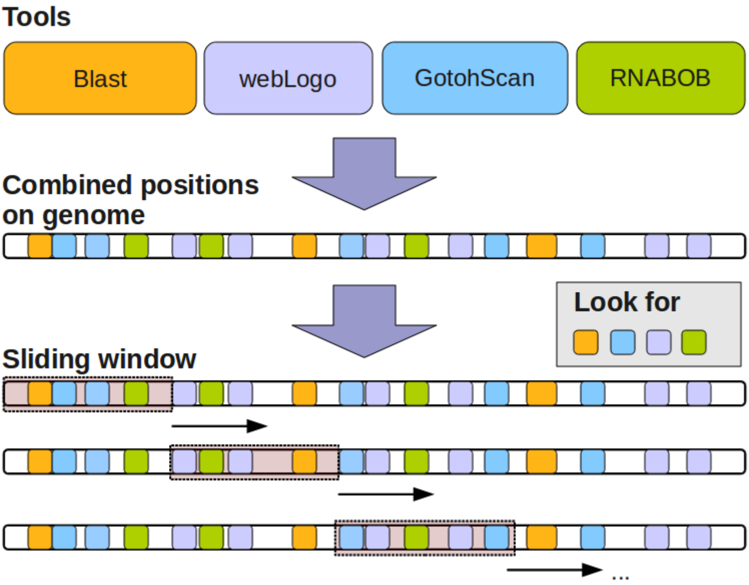
Starting point



Starting point



Core problem



Solution

Evaluations step

How often does which object occur?

🟧 8 🟪 4
🟩 6 🟦 3 ➡ Rarest: 🟦 Motif order: 🟪🟩🟧

Look for

🟧 🟦 🟪 🟩

Recursion step (for each window)

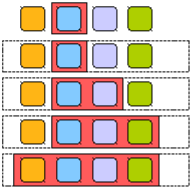
```
For each rarest object
  extend the window to match maximal spread
    findNextMotif()
```

Example

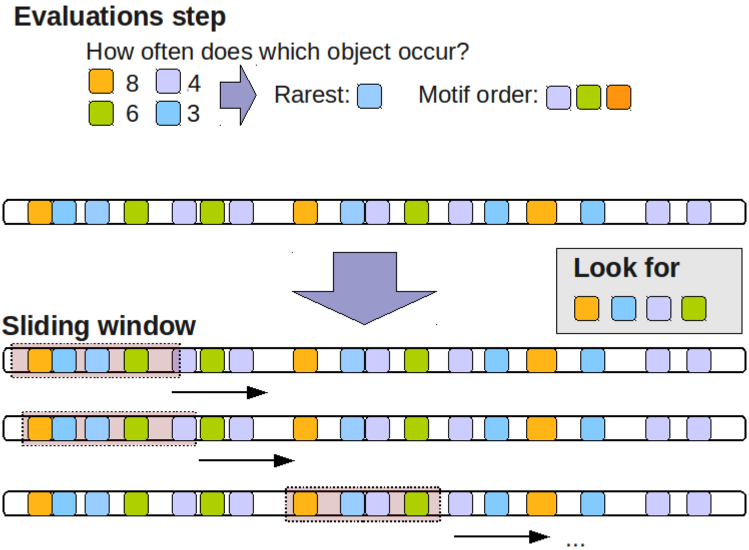
Find rarest

Extend window

Find next...



Result



Evaluations step

How often does which object occur?

🟧 8 🟪 4

🟩 6 🟦 3

Rarest: 🟦

Motif order: 🟪 🟩 🟧

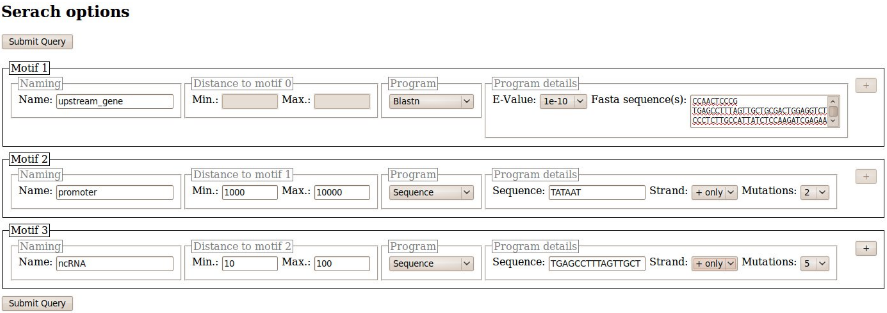
Look for

🟧 🟦 🟪 🟩

Sliding window

...

Online tool



Serach options

Submit Query

Motif 1

Naming — Name: upstream_gene

Distance to motif 0 — Min.: Max.:

Program: Blastn

Program details — E-Value: 1e-10 Fasta sequence(s):

```
CCAACTCCCG
TGAGCCTTTAGTTGCTGCGACTGGAGGTCT
CCCTCTTGCCATTATCTCCAAGATCGAGAA
```

Motif 2

Naming — Name: promoter

Distance to motif 1 — Min.: 1000 Max.: 10000

Program: Sequence

Program details — Sequence: TATAAT Strand: + only Mutations: 2

Motif 3

Naming — Name: ncRNA

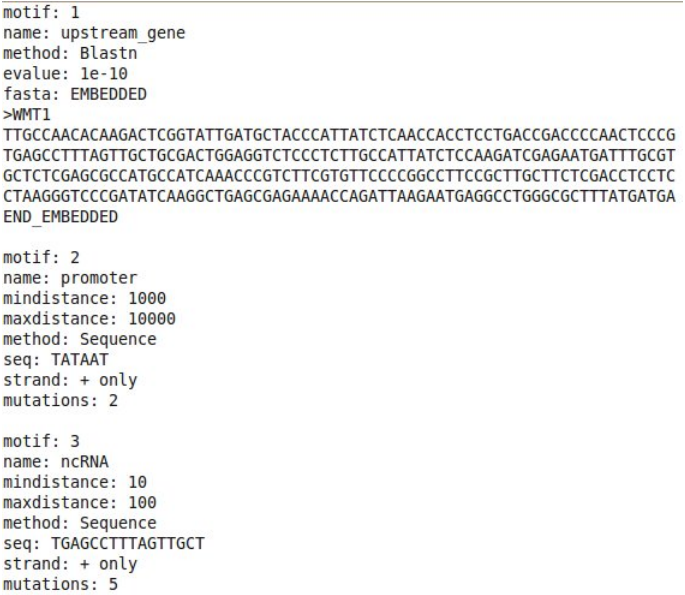
Distance to motif 2 — Min.: 10 Max.: 100

Program: Sequence

Program details — Sequence: TGAGCCTTTAGTTGCT Strand: + only Mutations: 5

Submit Query

Data format

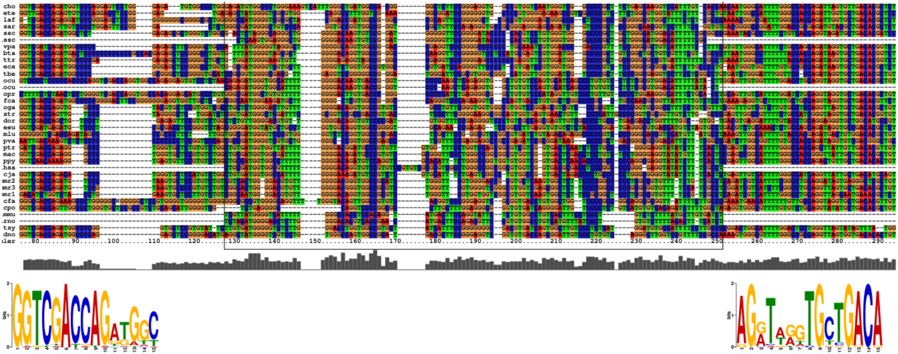


```
motif: 1
name: upstream_gene
method: Blastn
evalue: 1e-10
fasta: EMBEDDED
>WMT1
TTGCCAACACAAGACTCGGTATTGATGCTACCCATTATCTCAACCACCTCCTGACCGACCCCAACTCCCG
TGAGCCTTTAGTTGCTGCGACTGGAGGTCTCCCTCTTGCCATTATCTCCAAGATCGAGAATGATTTGCGT
GCTCTCGAGCGCCATGCCATCAAACCCGTCTTCGTGTTCCCCGGCCTTCCGCTTGCTTCTCGACCTCCTC
CTAAGGGTCCCGATATCAAGGCTGAGCGAGAAAACCAGATTAAGAATGAGGCCTGGGCGCTTTATGATGA
END_EMBEDDED

motif: 2
name: promoter
mindistance: 1000
maxdistance: 10000
method: Sequence
seq: TATAAT
strand: + only
mutations: 2

motif: 3
name: ncRNA
mindistance: 10
maxdistance: 100
method: Sequence
seq: TGAGCCTTTAGTTGCT
strand: + only
mutations: 5
```

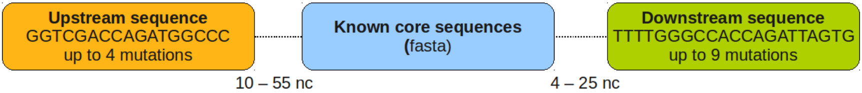
Does it work? - A test with paRNAs

paRNA

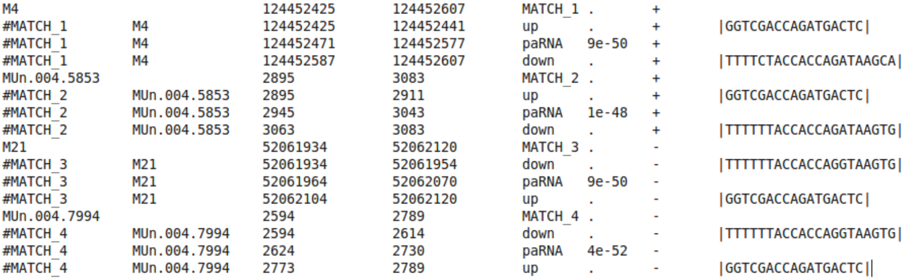
- promoter associated RNAs
- complementary in sequence to the rDNA promoter
- needed for rDNA methylation an silencing



Does it work? - A test with paRNAs



Run with genome of Bos taurus



```
M4                                  124452425     124452607     MATCH_1 .       +
#MATCH_1       M4                   124452425     124452441     up      .       +       |GGTCGACCAGATGACTC|
#MATCH_1       M4                   124452471     124452577     paRNA   9e-50   +
#MATCH_1       M4                   124452587     124452607     down    .       +       |TTTTCTACCACCAGATAAGCA|
MUn.004.5853                        2895          3083          MATCH_2 .       +
#MATCH_2       MUn.004.5853         2895          2911          up      .       +       |GGTCGACCAGATGACTC|
#MATCH_2       MUn.004.5853         2945          3043          paRNA   1e-48   +
#MATCH_2       MUn.004.5853         3063          3083          down    .       +       |TTTTTTACCACCAGATAAGTG|
M21                                 52061934      52062120      MATCH_3 .       -
#MATCH_3       M21                  52061934      52061954      down    .       -       |TTTTTTACCACCAGGTAAGTG|
#MATCH_3       M21                  52061964      52062070      paRNA   9e-50   -
#MATCH_3       M21                  52062104      52062120      up      .       -       |GGTCGACCAGATGACTC|
MUn.004.7994                        2594          2789          MATCH_4 .       -
#MATCH_4       MUn.004.7994         2594          2614          down    .       -       |TTTTTTACCACCAGGTAAGTG|
#MATCH_4       MUn.004.7994         2624          2730          paRNA   4e-52   -
#MATCH_4       MUn.004.7994         2773          2789          up      .       -       |GGTCGACCAGATGACTC||
```

Does it work? - A test with paRNAs

bta	4
cfa	1
cho	6
cja	7
cpo	3
dno	6
dor	33
eca	8
eeu	1
ete	10
fca	7
hsa	0
laf	27
mac	15
mlu	4
mmr	13
mmu	1
ocu	37
oga	1
opr	30
pca	2
ppy	1
ptr	1
pva	4
rno	2
sar	10
ssc	1
str	6
tbe	20
tsy	1
ttr	12

Have

- all-in-one tool
- web interface
- command line interface
- export and import function
- replicability
- extendable
- parallelization
- .bed output

Outlook

- add more tools
- some applications
- add chose from a list of genomes instead of upload
- predefined search patterns
- flexible scoring scheme
- .bed input
- more fancy output
- output of necessary citations
- small tutorial scientists with limited background in computer science

Thank you