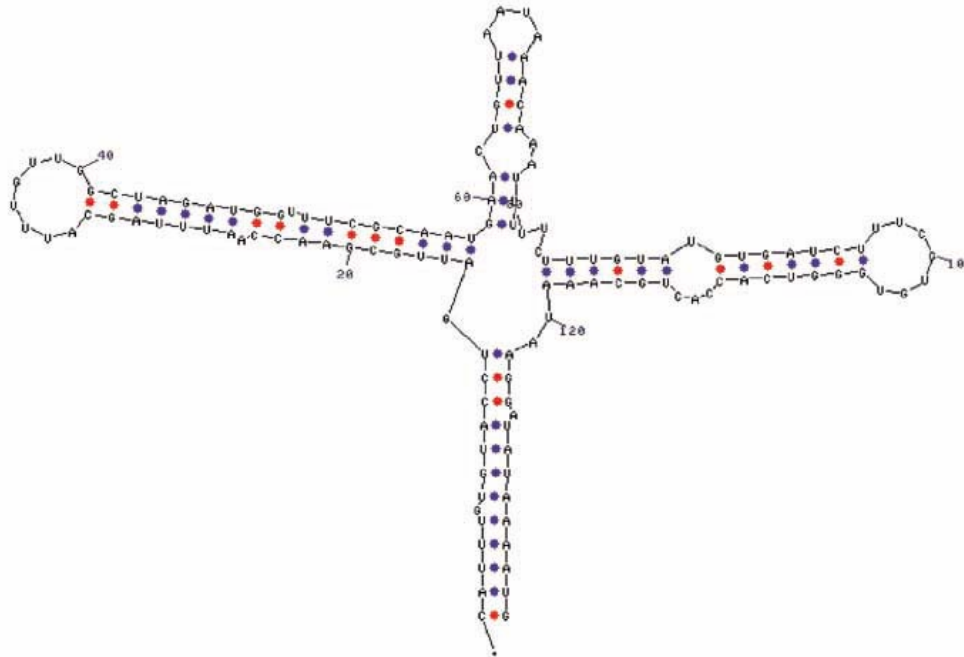


Lab 8.3: RNA Secondary Structure



Jennifer Gardy

Centre for Microbial Diseases and Immunity Research

University of British Columbia

jennifer@cmdr.ubc.ca

Outline

- The chocolate receptor – yum!
- MFOLD: Secondary structure prediction
- Building a consensus sequence
- Forcing base pairs
- Searching for RNA motifs in the genome

The chocolate receptor

- Human chocolate receptor found!



- You suspect that the chocolate receptor is posttranscriptionally regulated in response to complex behavioral stimuli, but how?

LOCUS HSCHOCR 5010 bp mRNA PRI 11-OCT-1999
 DEFINITION Human mRNA for chocolate receptor.
 ACCESSION X010601
 NID g374321
 VERSION X010601.1 GI:374321
 KEYWORDS chocolate receptor.
 SOURCE human.
 ORGANISM Homo sapiens
 Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria;
 Primates; Catarrhini; Hominidae; Homo.
 REFERENCE 1 (bases 1 to 5010)
 AUTHORS Hershey, A.
 TITLE Primary structure of human chocolate receptor deduced from the
 mRNA sequence
 JOURNAL Nature Junk Food 3 (23), 675-678 (1999)
 MEDLINE 850127431
 COMMENT Data kindly reviewed (14-OCT-1999) by W. Wonka.

FEATURES
 source 1..5010
 /organism="Homo sapiens"
 /taxon="taxon:9606"
 CDS 264..2546
 /product="chocolate receptor (aa 1-760)"
 /codon_start=1
 /protein_id="CAA255271.1"
 /db_xref="PID:g374331"
 /db_xref="GI:374331"
 /db_xref="SWISS-PROT:P027861"
 /translation="MMDQARSAFSNLFGGEPLSYTRFSLARQVDGDN SHVEMKLAVDE
 EENADNNTKANVTKPKRCSGSICYGTIAVIVVFFLIGFMIGYLG YCKGVEPKTECERLA
 GTESPVREEPGEDFPAARRLYWDDLKRKLSEKLDSTDFSTIKLLNENS YVPREAGSQ
 KDENLALYVENQFREFKLSKVWRDQHFVKIQVKDSAQNSVIIVDKNGRLVYLVENPGG
 YVAYSKAATVTGKLVHANFGTKKDFEDLYTPVNGSIVIVRAGKITFAEKVANAESLNA
 IGVL IYMDQTKFPIVNAELSFFGHAHLGTGDPYTPGFPSFNHTQFPPSRSSGLPNIPV
 QTISRAAAEKLFGNMEGDCPSDWKTDSTCRMVTSSEKNVKLTVSNVLKEIKILNIFGV
 IKGFVEPDHYVVVGAQRDAWGPGAAKSGVGTALLLKL AQMFSDMVLDKGFQPSRSIIF
 ASWSAGDFG SVGATEWLEGYLSSLHLKAFETYINLDKAVLGTSNFKVSASPLLYTLIEK
 TMQNVKHPVTGQFLYQDSNWASKVEKLTLDNAAFPFLAYSGIPAVSFCFCEDTDYPYL
 GTTMDTYKELIERIPELNKVARAAA EVAGQFVIKLT HDVELNLDYERYNSQLLSFVRD
 LNQYRADIKEMGLSLQWLYSARGDFFRATSRLTTDFGNAEKTDRFVMKKLNDRVMRVE
 YHFLSPYVSPKESPF RHVFWGSGSHTLPALLENLKL RKQNNGAFNETLFRNQLALATW
 TIQGAANALSGDVWDIDNEF"

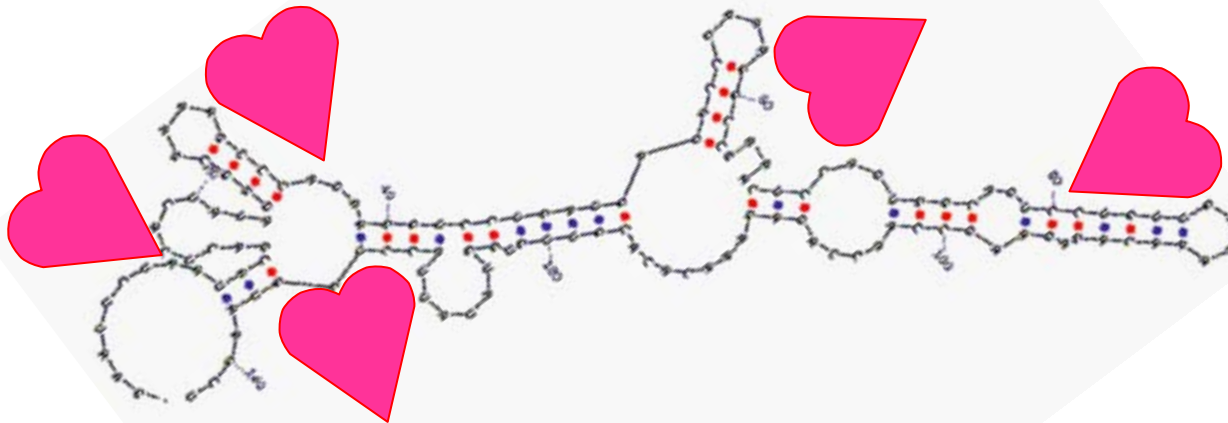
Unusually long 3' UTR

Analysis of the 3' UTR

- Lab studies show bases 3430-4025 necessary and sufficient for posttranscriptional regulation
- Cross-linking data shows a protein factor (the Valentine factor) binds this region in at least two, if not more, places
- Sequence analysis doesn't turn up anything
- **RNA binding proteins often don't recognize SEQUENCE, they recognize STRUCTURE**
 - Many RNA secondary structures involved in gene regulation are short hairpins with a bulge in the stem

Hypothesis

- The Valentine factor proteins binds the chocolate receptor 3' UTR via a conserved secondary structure RNA motif to regulate expression of the receptor



- Computational analysis (2° structure prediction) can identify this motif in the chocolate receptor
- We can use our motif to find more Valentine-regulated genes in the human genome

2° Structure Prediction with MFOLD

- MFOLD developed by Mike Zuker in 1989
- Uses energy minimization to predict folding of an RNA sequence into a secondary structure:
 - Uses a set of base pairing rules (e.g. G:C, A:U, G:U only) to create “optimal” structure (lowest free energy) and “suboptimal” structures (within 12kcal/mol of optimum)
 - 41% overall accuracy when tested on known RNA structures (Doshi et al., BMC Bioinformatics 5:105)

Day 8 wiki > HSCHOCR_utr.txt file (FASTA file of key region involved in gene regulation) – open onscreen

Open a 2nd browser to the MFOLD web server:

<http://www.bioinfo.rpi.edu/applications/mfold/old/rna/>

On your own: Viewing MFOLD Predictions – ~15min.

☐ View Individual Structures:

[Click Here for New Structure Viewing Options](#)

■ Structure 1 : Initial dG = -160.5 kcal/mole, (*[Thermodynamic Details](#)*).

Different file formats: *[PostScript](#)*, *[png](#)*, *[jpg](#)*, *[new](#)*, *[.ct file](#)*, *[Vienna](#)*, *[RNAML](#)*, *[RnaViz ct](#)*, *[Mac ct](#)*, *[GCG](#)*, *[XRNA ss](#)*.

■ Structure 2 : Initial dG = -160.4 kcal/mole, (*[Thermodynamic Details](#)*).

Different file formats: *[PostScript](#)*, *[png](#)*, *[jpg](#)*, *[new](#)*, *[.ct file](#)*, *[Vienna](#)*, *[RNAML](#)*, *[RnaViz ct](#)*, *[Mac ct](#)*, *[GCG](#)*, *[XRNA ss](#)*.

■ Structure 3 : Initial dG = -160.3 kcal/mole, (*[Thermodynamic Details](#)*).

Different file formats: *[PostScript](#)*, *[png](#)*, *[jpg](#)*, *[new](#)*, *[.ct file](#)*, *[Vienna](#)*, *[RNAML](#)*, *[RnaViz ct](#)*, *[Mac ct](#)*, *[GCG](#)*, *[XRNA ss](#)*.

■ Structure 4 : Initial dG = -159.8 kcal/mole, (*[Thermodynamic Details](#)*).

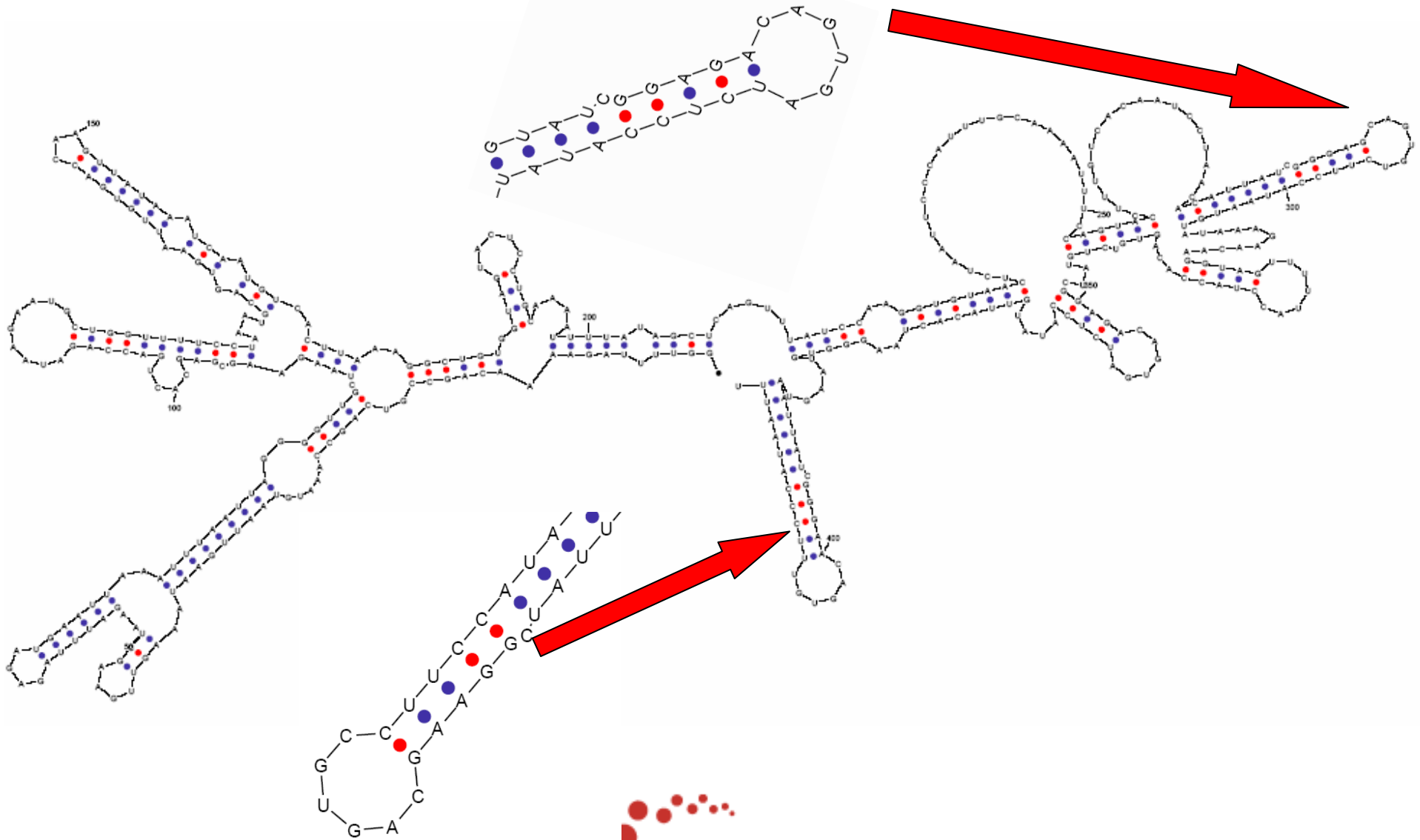
Different file formats: *[PostScript](#)*, *[png](#)*, *[jpg](#)*, *[new](#)*, *[.ct file](#)*, *[Vienna](#)*, *[RNAML](#)*, *[RnaViz ct](#)*, *[Mac ct](#)*, *[GCG](#)*, *[XRNA ss](#)*.

■ Structure 5 : Initial dG = -159.0 kcal/mole, (*[Thermodynamic Details](#)*).

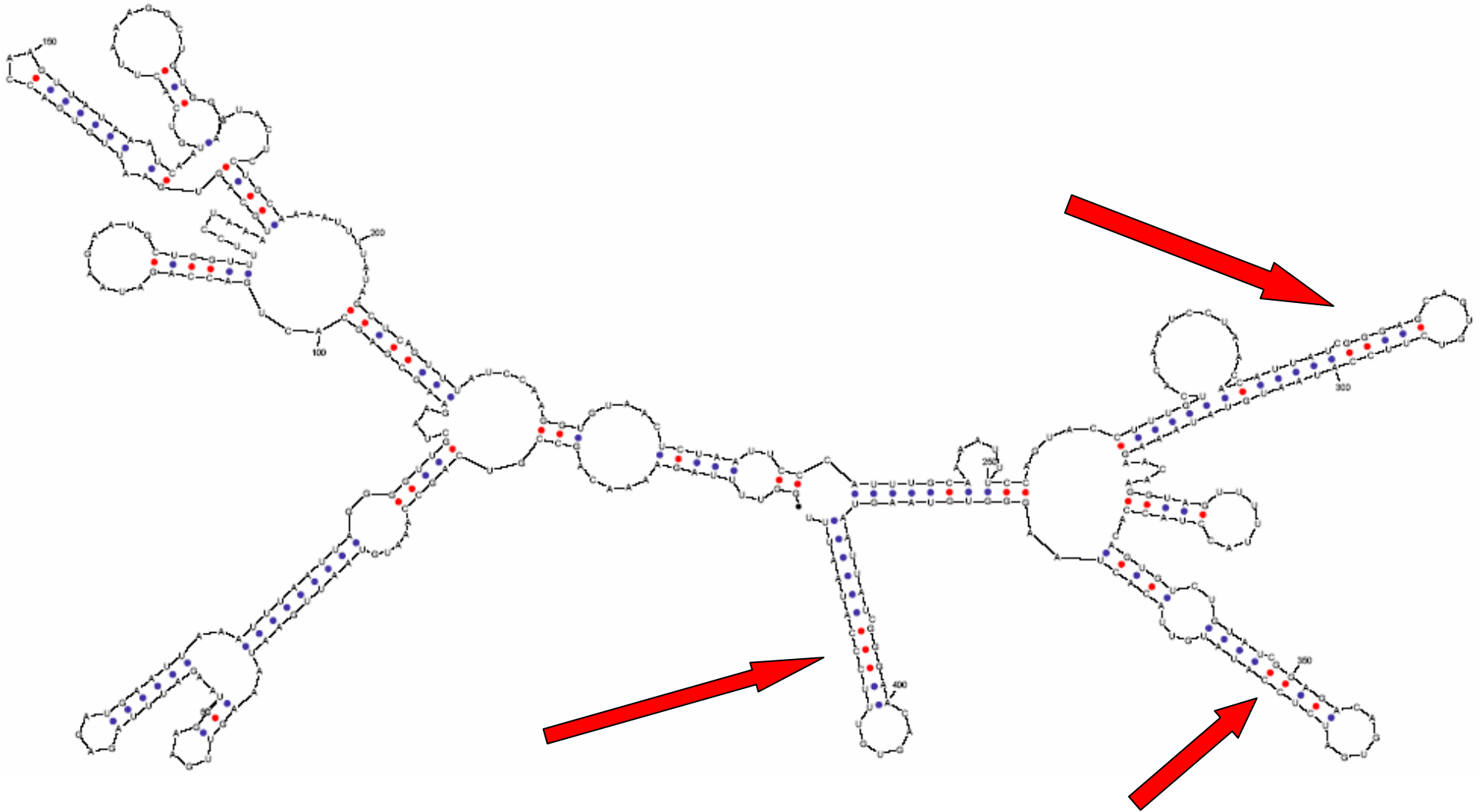
Different file formats: *[PostScript](#)*, *[png](#)*, *[jpg](#)*, *[new](#)*, *[.ct file](#)*, *[Vienna](#)*, *[RNAML](#)*, *[RnaViz ct](#)*, *[Mac ct](#)*, *[GCG](#)*, *[XRNA ss](#)*.

- Have each team member open one of the top four structures onscreen:
 - Save file locally, at shell prompt, type **gv filename**
- Look for conserved 2° structure motifs
 - Hint: Pay special attention to hairpin loops with a bulge!
- How many motifs do you find in each structure?

2 Motifs in Optimal Structure

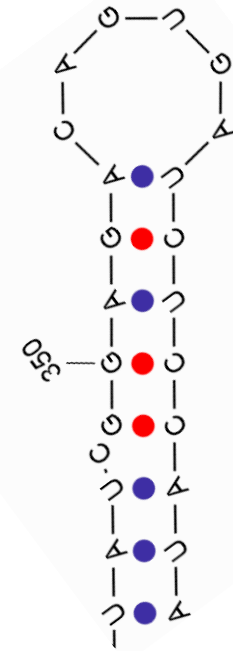


3 Motifs in Structures 2 & 3



From Motif to Sequence

- 6bp loop, 3bp – bulge – 5bp stem
- Grab sequences of the motifs found:
 - 3rd motif in structures 2, 3 is shortest
 - Use as a guide for the lengths of the others



UUAUCGGAAGCAGUGCCUCCAUAA
GUAUCGGAGACAGUGAUCUCCAUAU
UUAUCGGGAGCAGUGUCUCCAUAA

On your own: Consensus Sequence – ~5 min.

- What would the motif's consensus sequence be?

```
TAT | C | GGAAG | CAGTGC | CTTCC | ATA
TAT | C | GGAGA | CAGTGA | TCTCC | ATA
TAT | C | GGGAG | CAGTGT | CTTCC | ATA
>>> * >>>>> <<<<<< <<<<
TAT C GG--- CAGTG- --TCC ATA
```

- Search 3' UTR FASTA file for more occurrences:

```
>HSCHOCR Human mRNA for chocolate receptor; positions 3430..4025
tatttatcagtgacagagttcactataaatgggtgtttttttaatagaata
taattatcggaagcagtgcttccataattatgacagttatactgtcggg
tttttttaataaaaagcagcatctgctaataaaacccaacagatactgga
agttttgcatttatgggtcaacacttaaggggttttagaaaacagccgtcag
ccaaatgtaattgaataaagttgaagctaagatttagagatgaattaaat
ttaattagggggttgctaagaagcgcagcactgaccagataagaatgctggg
tttccataaatgcagtggaattgtgaccaagttataaatcaatgtcacttaa
aggctgtggtagtactcctgcaaaaattttatagctcagtttatccaagggt
gtaactctaattcccatttgcaaaaatccagtacctttgtcacaatcct
aacacattatcgggagcagtgcttccataatgtataaagaacaaggtag
tttttacctaccacagtgctgttatcggagacagtgatcctccatatggtta
cactaaggggtgtaagtaattatcgggaacagtgtttccataat
```

Are There Even More Motifs?

- Search for sub-segments of the consensus



```

>HSCHOCR Human mRNA for chocolate receptor; positions 3430..4025
tattatcagtgacagagttcactataaatggtgtttttttaatagaata
taattatcggagcagtgccctccataattatgacagttatactgtcgggt
tttttttaaataaaagcagcatctgctaataaaaccaacagatactgga
agttttgcatttatggtcaacacttaaggggttttagaaaacagccgtcag
ccaatgtaattgaataaagtgaagctaagatttagagatgaattaaat
ttaattaggggttgctaagaagcgagcactgaccagataagaatgctggt
ttcctaaatgcagtgaattgtgaccaagttataaatcaatgtcacttaa
aggctgtggtagtactcctgcaaaaatttatagctcagttatccaaggt
gtaactctaattcccatttgcaaaaatccagtacctttgtcacaatcct
aacacatatcgggagcagtgctccataatgtataaagaacaaggtag
ttttacctaccacagtgctgtatcggagacagtgatctccatatgtta
cactaaggggtgtaagtaatatcgggaacagtgtttccataatt
  
```

- Our sequence contains **FIVE** motifs!

Revisit MFOLD with New Information

- MFOLD allows you to “constrain” base pairing
 - When you have prior information about some of the secondary structure elements in a sequence (e.g. which bases should pair with each other to form a helix)

Enter **constraint information** in the box at the right. (optional) You may:

1. force bases $i, i+1, \dots, i+k-1$ to be double stranded by entering:

D i 0 k on 1 line in the constraint box.

2. force consecutive base pairs $i, j, i+1, j-1, \dots, i+k-1, j-k+1$ by entering:

D i j k on 1 line in the constraint box.

3. force bases $i, i+1, \dots, i+k-1$ to be single stranded by entering:

S i 0 k on 1 line in the constraint box.

4. prohibit the consecutive base pairs

$i, j, i+1, j-1, \dots, i+k-1, j-k+1$ by entering:

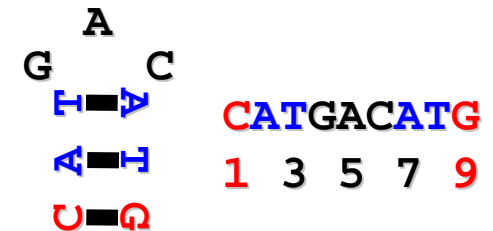
P i j k on 1 line in the constraint box.

5. prohibit bases i to j from pairing with bases k to l by entering:

P i-j k-l on 1 line in the constraint box.

On your own: Forcing Base Pairs – ~15min.

- Force command = `F a b c`
 - *a*: residue number of first base pair
 - *b*: residue number of last base pair
 - *c*: how many consecutive bases to pair
 - E.g. `F 1 9 3`



- Given the information below, set up constraints in MFOLD to force the formation of our five motifs:

```

5      |TAT|C|AGTGA|CAGAGT|TCACT|ATA| 27
55     |TAT|C|GGAAG|CAGTGC|CTTCC|ATA| 77
458    |TAT|C|GGGAG|CAGTGT|CTTCC|ATA| 480
523    |TAT|C|GGAGA|CAGTGA|TCTCC|ATA| 545
570    |TAT|C|GGGAA|CAGTGT|TTCCC|ATA| 592
>>> * >>>>>          <<<<<< <<<

```

- Rerun MFOLD with your constraints, view optimal structure

MFOLD Constraints

FASTA format may be used.

```
>HSCHOCR Human mRNA for chocolate receptor; positions 3430..4025
tatttatcagtgacagagttcactataaaatgggtgttttttaataagaata
taattatcgggaagcagtgccctccataaattatgacagttatactgtcggg
tttttttaataaaaagcagcatctgctaataaaaacccaacagatactgga
agttttgcatthtatgggtcaacacttaagggtttagaaaacagccgtcag
ccaaatgtaattgaataaagttgaagctaagatttagagatgaattaaat
ttaattaggggttgctaagaagcgagcactgaccagataagaatgctggt
tttcctaaatgcagtgattgtgaccaagttataaatcaatgtcacttaa
aggctgtggtagtagtactcctgcaaaaattttatagctcagttatccaaggt
gtaactcctaattcccatttgcaaaaattccagtagctttgtcacaaatcct
aacacattatcgggagcagtggtcttccataaatgtataaagaacaaggtag
```

- Enter **constraint information** in the box at the right. (optional) You may:

- force bases $i, i+1, \dots, i+k-1$ to be double stranded by entering:
F i 0 k on 1 line in the constraint box.
- force consecutive base pairs $i, j, i+1, j-1, \dots, i+k-1, j-k+1$ by entering:
F i j k on 1 line in the constraint box.
- force bases $i, i+1, \dots, i+k-1$ to be single stranded by entering:
P i 0 k on 1 line in the constraint box.
- prohibit the consecutive base pairs
 $i, j, i+1, j-1, \dots, i+k-1, j-k+1$ by entering:
P i j k on 1 line in the constraint box.
- prohibit bases i to j from pairing with bases k to l by entering:
P i-j k-l on 1 line in the constraint box.

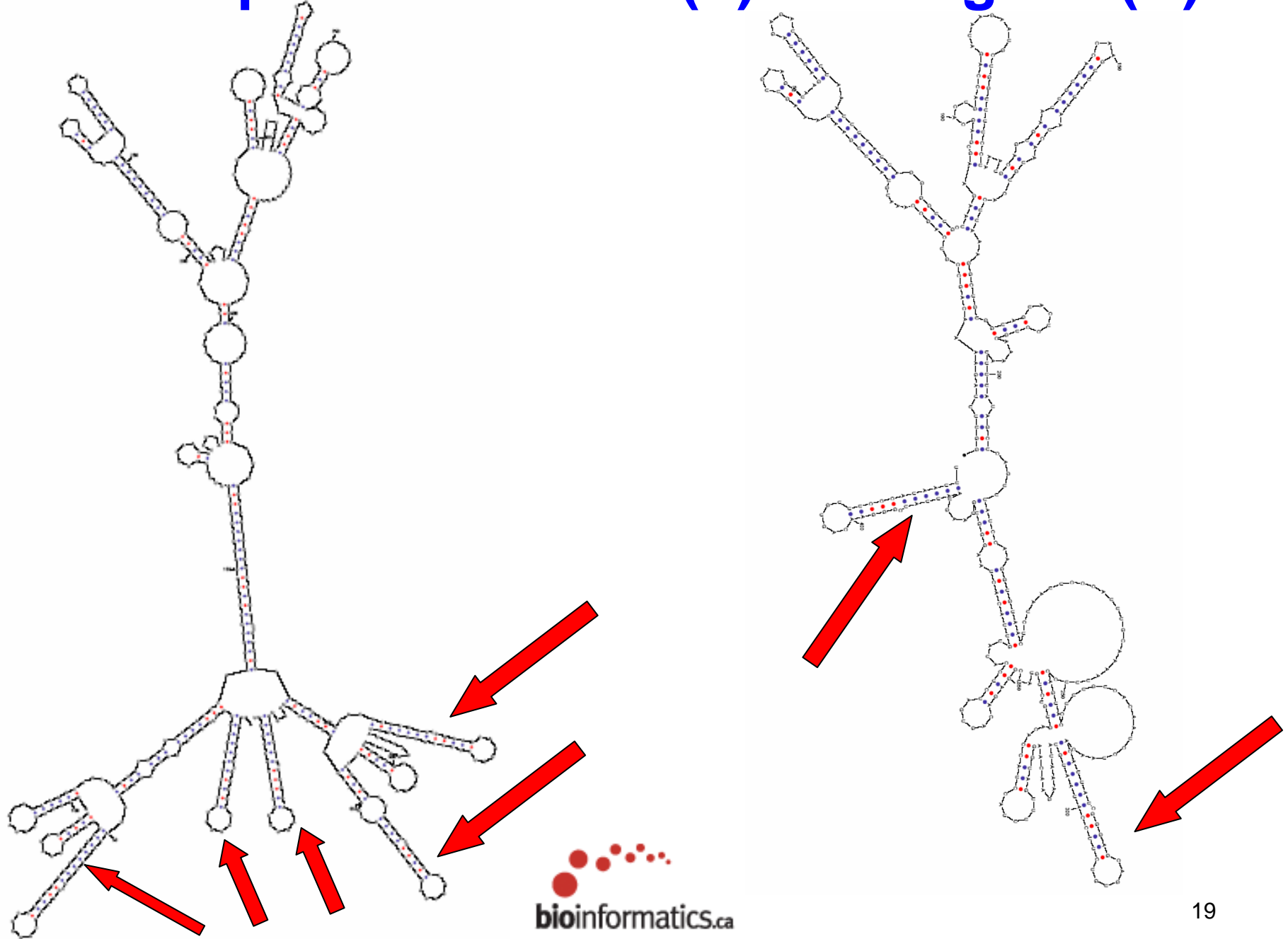
- The RNA sequence is .

- Folding temperature is fixed at 37°.

- Ionic conditions: 1M NaCl, no divalent ions.

```
F 5 27 3
F 9 24 5
F 55 77 3
F 59 74 5
F 458 480 3
F 462 477 5
F 523 545 3
F 527 542 5
F 570 592 3
F 574 589 5
```

Comparison – Final (L) vs. Original (R)



Demonstration: Finding More Motifs in the Genome

- Are other human genes regulated by Valentine factor binding to a motif in the UTR?
- Find dataset of human genes to search against:
 - Can we constrain our search to a subset of the genome?
 - ***Must have 5' or 3' UTR***
 - What resource could we use to create this dataset?
 - ***Ensembl's BioMart***
- Create a description of our motif and use this to search the database
 - RNAMOT

Creating the Dataset

- Ensembl > BioMart > Homo sapiens genes
- Export > Sequence > 5' UTR
 - Gene ID
 - Description
- Same for 3' UTR
- Gunzip each file
- Concatenate:
 - `cat file1 file2 > file3`

The screenshot shows the Ensembl BioMart interface for configuring a dataset export. The 'Sequences' attribute page is selected. The 'Type of Sequence to Export' section shows two diagrams of a gene structure with the 5' UTR highlighted in red. The 'Unspliced (Transcript)' option is selected. The 'Upstream flank' and 'Downstream flank' are both set to 100. The 'Header Information' section shows 'Ensembl Gene ID' and 'Description' selected. The 'Summary' panel on the right shows the dataset is 'Homo sapiens genes' with 34294 entries total and 34294 entries passing filters. The output is set to 'Sequences'.

Select the Attribute Page

Sequences

SEQUENCES:

Type of Sequence to Export (all in 5'-3' direction):

Unspliced (Transcript) (selected)
Fank (Transcript)
Fank-coding region (Transcript)
5' UTR
Exon sequences (Transcript)
cDNA sequences
Peptide

Unspliced (Gene)
Fank (Gene)
Fank-coding region (Gene)
3' UTR
Exon sequences (Gene)
Coding sequence

Upstream flank: 100

Downstream flank: 100

Header Information

Gene Attributes

Chromosome
Ensembl Gene ID (versioned)
External Gene DB
Description (selected)

Ensembl Gene ID (checked)
External Gene ID
Gene Family

Summary

start

Dataset: Homo sapiens genes

34294 Entries Total

filter

None

34294 Entries pass Filters

output

Sequences

34294 Results in Output

RNAMOT

- Simple motif-searching algorithm (1990)
 - Other motif searching methods available (e.g. RNAMotif), but require more complex input scripts
- Command-line usage
- Basic command:
 - `rnamot -s -s datasettosearchagainst -d motifdescriptor -o results`
- -s: sequences to be scanned for the motif
 - UTRs from Ensembl
- -d: motif descriptor file
- -o: where to write the output to

RNAMOT Analysis of Human UTRs

- 4560 genes, 5381 motifs

```
--- ENSG00000136305|Cell death activator CIDE-B (Cell death-inducing DFFA-like effector B). [Source:Uniprot/SWISSPROT;Acc:Q9UHD4]|protein_coding --- (2728 bases)
|SCO: 200.76|POS:2303-2325|MIS: 0|WOB: 1| | |
|UUU|C|UUUGU|CAGGAC|ACAGA|AAA|
|SCO: 200.37|POS:2696-2718|MIS: 0|WOB: 0|
|AGA|A|GUUCC|AGGGAA|GGAAC|UCU|

--- ENSG00000168986|Leukotriene B4 receptor 1 (LTB4-R 1) (P2Y purinoceptor 7) (P2Y7) (Chemoattractant receptor-like 1). [Source:Uniprot/SWISSPROT;Acc:Q15722]|protein_coding --- (1725 bases)
|SCO: 200.47|POS:310-332|MIS: 0|WOB: 1|
|CUG|U|CAUUC|AGGCUG|GAGUG|CAG|

--- ENSG00000100442|FK506-binding protein 3 (EC 5.2.1.8) (Peptidyl-prolyl cis-trans isomerase) (PPIase) (Rotamase) (25 kDa FKBP) (FKBP-25) (Rapamycin-selective 25 kDa immunophilin). [Source:Uniprot/SWISSPROT;Acc:Q00688]|protein_coding --- (411 bases)
|SCO: 200.67|POS:76-98|MIS: 0|WOB: 0|
|UUU|A|AAGGC|AUAGCU|GCCUU|AAA|

--- ENSG00000165527|ADP-ribosylation factor 6. [Source:Uniprot/SWISSPROT;Acc:P62330]|protein_coding --- (536 bases)
|SCO: 200.33|POS:427-449|MIS: 0|WOB: 1|
|GGC|G|CUCUC|GCGGCC|GAGAG|GCU|

--- ENSG00000100505|Tripartite motif protein 9 (RING finger protein 91). [Source:Uniprot/SWISSPROT;Acc:Q9C026]|protein_coding --- (1122 bases)
|SCO: 200.34|POS:582-604|MIS: 0|WOB: 1| | |
|UCC|G|GCAGC|CCACUC|GCUGC|GGG|
|SCO: 200.35|POS:859-881|MIS: 0|WOB: 1|
|CCC|A|AUGGG|CCGCUG|CUCAU|GGG|

--- ENSG00000100505|Tripartite motif protein 9 (RING finger protein 91). [Source:Uniprot/SWISSPROT;Acc:Q9C026]|protein_coding --- (390 bases)
|SCO: 200.35|POS:127-149|MIS: 0|WOB: 1|
|CCC|A|AUGGG|CCGCUG|CUCAU|GGG|

--- ENSG00000100505|Tripartite motif protein 9 (RING finger protein 91). [Source:Uniprot/SWISSPROT;Acc:Q9C026]|protein_coding --- (390 bases)
|SCO: 200.35|POS:127-149|MIS: 0|WOB: 1|
|CCC|A|AUGGG|CCGCUG|CUCAU|GGG|

--- ENSG00000177947|sperm tail protein SHIPP01 [Source:RefSeq_peptide;Acc:NP_444510]|protein_coding --- (309 bases)
|SCO: 200.46|POS:85-107|MIS: 0|WOB: 1|
results.txt.sol
```

Take Home Messages

- MFOLD and other RNA secondary structure prediction tools rarely give the right answer first (or at all)
 - Too many possible structures in the low energy neighbourhood
- Can be used as a “first-pass” tool
 - Eyeball key conserved motifs
 - Collect sequences to build a consensus
- Often need to adjust parameters
 - Use prior knowledge to force base pairing
- Motif-searching tools can be used to identify conserved secondary structure motifs in a sequence database
 - Retrieves more results than sequence-based searches

Other (Optional) Activities

- The Valentine factor binding motif in the chocolate receptor is actually IRE - the iron response element.
- The chocolate receptor is transferrin and the Valentine factor is IRF/IRE-BP.
- Visit UTRSite to learn about IRE. What is UTRSite?
 - <http://www2.ba.itb.cnr.it/UTRSite/>
 - Signal Manager > U0002
- Visit Rfam's IRE entry. What is Rfam?
 - <http://www.sanger.ac.uk/cgi-bin/Rfam/getacc?RF00037>
- Read about the IRE in its biological context
 - PMID: 8710843