Computational analysis of non-coding RNA *(continued)*

Andrew Uzilov

auzilov@ucsc.edu

BME110

Tue, Nov 23, 2010

Corrected/updated talk slides are here:

http://tinyurl.com/UzilovRna

redirects to:

http://users.soe.ucsc.edu/~auzilov/BME110/Fall2010/

Talk progress

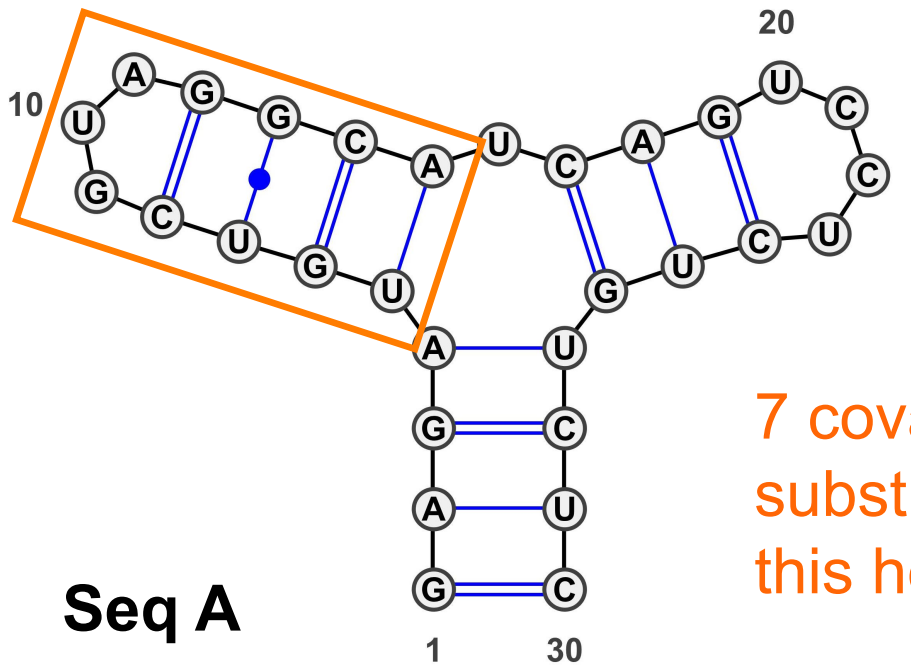
- **Introduction to RNA alignment**
- Tools for editing and viewing RNA alignments
- RNA alignment algorithms
  - theory
  - practice (example: FOLDALIGN)
- Practical advice

Why make RNA multiple sequence alignments?

- Required as input to:
  - train Rfam covariance models
  - generic gene-finders like RNAz, QRNA, Pfold, EvoFold, etc.

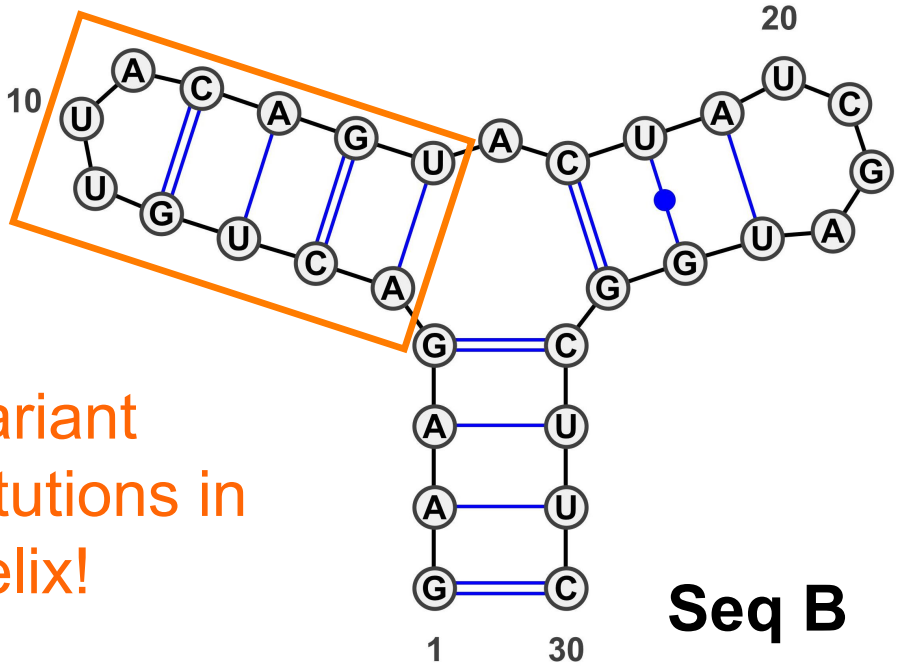
- Good MSAs are absolutely essential for the above tools to make reasonably accurate predictions!

What is covariance?



Seq A

7 covariant substitutions in this helix!



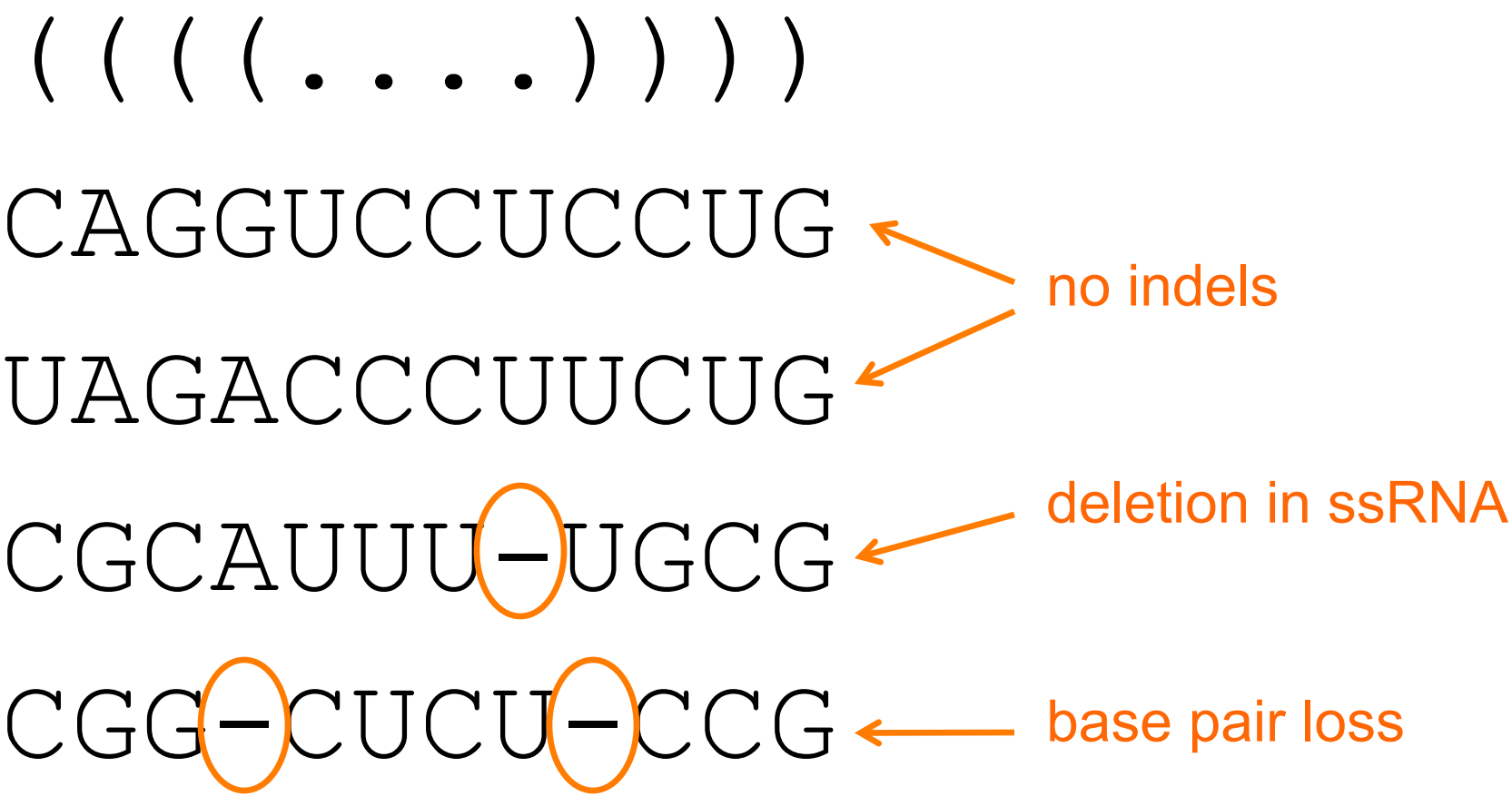
Seq B

```
seqA           GAGAUGUCGUAGGCAUCAGUCCUCUGUCUC
seqB           GAAGACUGUUACAGUACUAUCGAUGGCUUC
#=GC SS_cons   ((((((((...)))))).(((....)))))))
```

RNA structural alignment

- Key principle: alignment columns represent conserved base pairing
  - If column *i* base pairs with column *j*, then a base in *i* must form a base pair with the base in *j* in each sequence (or both bases are gaps)
  - How to align ssRNA is much less important

RNA structural MSA example



Talk progress

- Introduction to RNA alignment
- **Tools for editing and viewing RNA alignments**
- RNA alignment algorithms
  - theory
  - practice (example: FOLDALIGN)
- Practical advice

Tools for editing RNA alignments

- The bad news: slim pickings!
- Common hacks:
  - Emacs RALEE mode
  - spreadsheets
- Other options
  - Jalview RNA mode (forthcoming)
  - SARSE (clunky GUI, many dependencies)
  - S2S (sophisticated – best option?)

The new Jalview RNA support

- http://jalview-rnasupport.blogspot.com/
  - Added by Lauren Lui (Lowe lab)
  - should be in the next Jalview release
  - **LIVE DEMO!**

Visualizing RNA alignments

`colorstock.pl` (in the DART package)

- http://biowiki.org/RNAAlignmentViewers
- Written in Perl (NOT a webtool)
- Input in Stockholm format

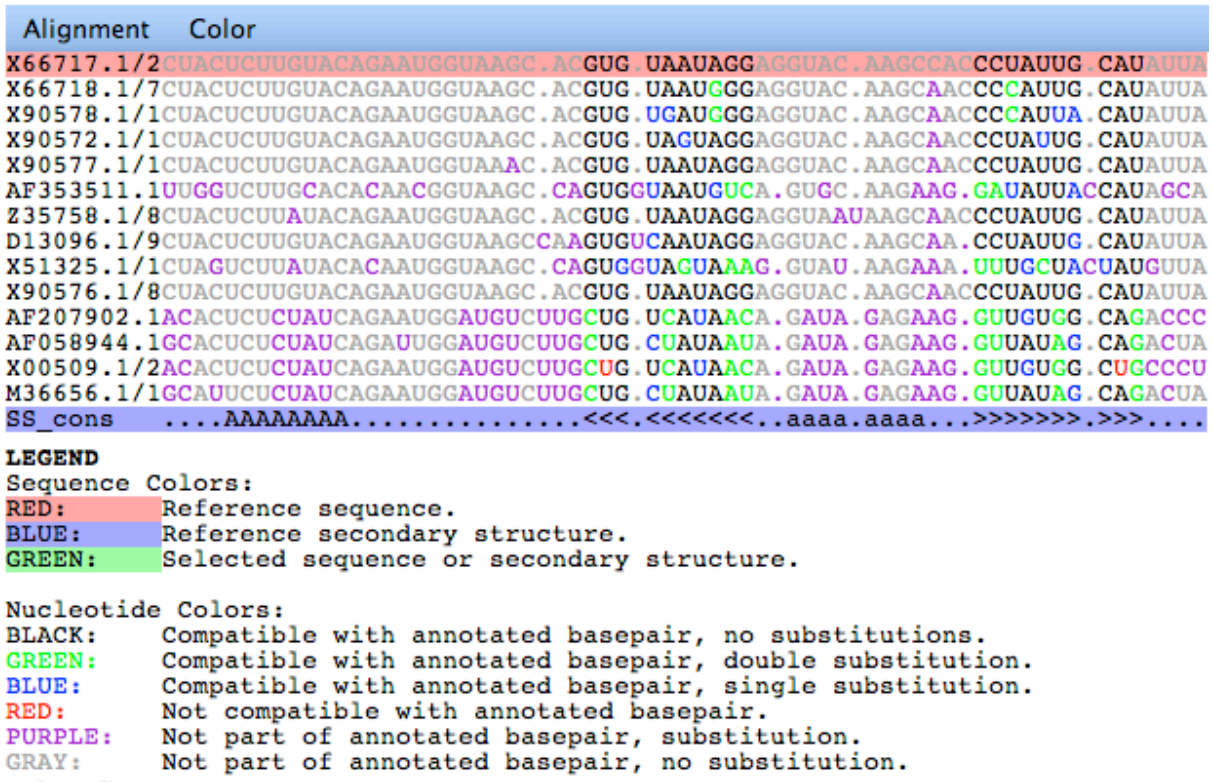
```
# STOCKHOLM 1.0
#=GC CS                 ..........................2.2.3432334............4332343.2.2....
#=GC BP                 ..........................ede.eeeeeee............eeeeeee.ede....
X66717.1/2272-2334      CUACUCUUGUACAGAAUGGUAAGC.ACGUG.UAAUAGGAGGUAC.AAGCCACCCUAUUG.CAUAUUA
X66718.1/795-857        CUACUCUUGUACAGAAUGGUAAGC.ACGUG.UAAUGGGAGGUAC.AAGCAACCCCAUUG.CAUAUUA
X90578.1/1000-1062      CUACUCUUGUACAGAAUGGUAAGC.ACGUG.UGAUGGGAGGUAC.AAGCAACCCCAUUA.CAUAUUA
X90572.1/1000-1062      CUACUCUUGUACAGAAUGGUAAGC.ACGUG.UAGUAGGAGGUAC.AAGCAACCCUAUUG.CAUAUUA
X90577.1/1000-1062      CUACUCUUGUACAGAAUGGUAAAC.ACGUG.UAAUAGGAGGUAC.AAGCAACCCUAUUG.CAUAUUA
AF353511.1/27820-27882  UUGGUCUUGCACACAACGGUAAGC.CAGUGGUAAUGUCA.GUGC.AAGAAG.GAUAUUACCAUAGCA
Z35758.1/8161-8224      CUACUCUUAUACAGAAUGGUAAGC.ACGUG.UAAUAGGAGGUAAUAAGCAACCCUAUUG.CAUAUUA
D13096.1/9359-9422      CUACUCUUGUACAGAAUGGUAAGCCAAGUGUCAAUAGGAGGUAC.AAGCAA.CCUAUUG.CAUAUUA
X51325.1/1443-1505      CUAGUCUUAUACACAAUGGUAAGC.CAGUGGUAGUAAAG.GUAU.AAGAAA.UUUGCUACUAUGUUA
X90576.1/809-871        CUACUCUUGUACAGAAUGGUAAGC.ACGUG.UAAUAGGAGGUAC.AAGCAACCCUAUUG.CAUAUUA
AF207902.1/31022-31083  ACACUCUCUAUCAGAAUGGAUGUCUUGCUG.UCAUAACA.GAUA.GAGAAG.GUUGUGG.CAGACCC
AF058944.1/9309-9370    GCACUCUCUAUCAGAUUGGAUGUCUUGCUG.CUAUAAUA.GAUA.GAGAAG.GUUAUAG.CAGACUA
X00509.1/2201-2262      ACACUCUCUAUCAGAAUGGAUGUCUUGCUG.UCAUAACA.GAUA.GAGAAG.GUUGUGG.CUGCCCU
M36656.1/1479-1540      GCAUUCUCUAUCAGAAUGGAUGUCUUGCUG.CUAUAAUA.GAUA.GAGAAG.GUUAUAG.CAGACUA
#=GC SS_cons            ....AAAAAAAA.............<<<.<<<<<<..aaaa.aaaa...>>>>>>>.>>>....
#=GC SN                 ..........................000.1111111............1111111.000....

#=GF CS                 67 columns, 10 paired, 9 covariant; 2 stems
//
```

Visualizing RNA alignments

RATON webserver

- http://harmony.biowiki.org/raton/
- Input in Stockholm format



Talk progress

- Introduction to RNA alignment
- Tools for editing and viewing RNA alignments
- **RNA alignment algorithms**
  - **theory**
  - practice (example: FOLDALIGN)
- Practical advice

Algorithms for RNA alignment

More than one way to do it!

- Machado-Lima, A. *et al*. "Computational methods in noncoding RNA research." *Journal of Mathematical Biology* **56** (2008): 15-49. http://www.ncbi.nlm.nih.gov/pubmed/17786447

- That review covers pretty much anything you ever wanted to know about computational RNA work!

Algorithms for RNA alignment

One common approach is to extend minimum free energy approaches (see Nov 16 lecture) to do alignment

- Alignment terms (usually just indel penalties) added to scoring function
- Simultaneous alignment and structure prediction (they are co-optimized)
- Not gene-specific!
- Variants of the Sankoff 1985 algorithm are used for this

Algorithms for RNA alignment

Sankoff is inefficient – $O(L^{3N})$ – how to constraint?

- Do pairwise alignment only
- Do progressive multiple alignment
  - Guide tree is important!
- Pruning optimizations exist

Talk progress

- Introduction to RNA alignment
- Tools for editing and viewing RNA alignments
- **RNA alignment algorithms**
  - theory
  - **practice (example: FOLDALIGN)**
- Practical advice

FOLDALIGN

- http://foldalign.ku.dk/server/index.html

- Pairwise RNA structural alignment
- Three types of alignment:
  - scan: a set of local, ranked alignments
  - local: only one local alignment (can be as long as whole sequence)
  - global: every base is aligned

FOLDALIGN

- Max sequence length is
  - 500 nt for “scan” mode
  - 200 nt for “local” and “global” modes

- Maximum length difference (Delta) setting
  - Constrains distance between aligned base pairs
  - Similar to “M” option in Dynalign, other Sankoff algorithms

FOLDALIGN

- **DEMO!**

Other webservers

- ClustalW and MUSCLE
  - Work remarkably well given that they don't model RNA structure (or anything about RNA) explicitly
  - See benchmarks in Gardner *et al*, NAR 2005

- R-Coffee
  - Theoretically elegant, but has never given me a structure that I believe

Alignment using gene-specific models

- Infernal (used by Rfam) provides command-line tools to use covariance models for alignment

- Failing that, can align several sequences to the model, then post-process to align them to each other

Talk progress

- Introduction to RNA alignment
- Tools for editing and viewing RNA alignments
- RNA alignment algorithms
  - theory
  - practice (example: FOLDALIGN)
- **Practical advice**

Practical advice summary

- Do you know the structure of at least one of the homologs?
  - Manually align some homologs to it (make "seed" alignment)
  - Build covariance model (Infernal software)
  - Use it to align other homologs (make "full" alignment)
- Do you know nothing about the structure?
  - Dynalign, Stemloc, FoldAlign, R-Coffee, etc.

Practical advice summary

- There are no perfect alignment tools
- Manual review and tweaking of the alignment is very important
- Useful to know which bases are important (from genetics and biochemistry)
- There may be more than one answer!

Making pretty RNA figures

VARNA is the best option for this

- http://varna.lri.fr/
- Customizable Java applet, but also runs from command line and has full Java API
- Accepts many common RNA formats
- Exports as EPS, PNG, JPEG, etc. (EPS is especially handy for additional tweaks)
- **LIVE DEMO!**