Curating and evaluating RNA structure assignments

Jan Gorodkin, Division of Genetics and Bioinformatics, IBHV,
The Royal Veterinary and Agricultural University, Denmark

Outline:

- Part I: Semi-Automated RNA Sequence Editor (SARSE).
  - Curating RNA structural alignments.
  - Rnadbtools and SARSE.
  - Integrating Pfold and Pcluster.
  - The temperature of Rfam.
- Part II: R_K: The K category correlation coefficient.
  - Comparing two K category assignments.
  - Pearson's correlation coefficient and least square fitting.
  - Extending Pearson's correlation coefficient to two K dimensional tables the R_K coefficient.
  - Discretization, an extension of Matthews correlation coefficient.
  - Applications of R_K.

Acknowledgements

KVL Collaborators:

- Allan Lind-Thomsen
- Jakob H. Havgaard
- Troels Arvin
- Steen Lumholdt
- Susi E. Kristensen
- Peter Sestoft

External Collaborators:

- Ebbe S. Andersen (Aarhus University)
- Bjarne Knudsen (CLC Bio)
- Jørgen Kjems (Aarhus University)
- Niels Larsen (Aarhus University)
- Christian Zwieb (University of Texas)
- Jody Burks (Auburn University)

Funding:

- The Danish Research Councils
- The Danish Center for Scientific Computing

Semi-Automated RNA Sequence Editor (SARSE)

Motivation:

- Good curated alignments to compare with predictions.
- Some years ago (still valid): Most ncRNA structural alignments have bad quality and obvious contain inconsistencies.
- Some databases even lack structural assignments corresponding to the multiple RNA alignment.
- Some years ago SRP RNA one of the best alignments, but had many inconsistencies. Old clean: from 20 to 3 pr sequence.
- Doing this kind of work is extremely useful to the community, but also extremely low prestige (and no funding :-(..).)
- Exist no good editor which include basic editing functions combined with structural consistency checks.

RNAdbtools: http://rnadbtool.kvl.dk

- Toolbox to conduct basic consistency checks.
- Highlights any tyupe of non-standard RNA pairs (and check, whether bases assigned to the same pair).
- Extends RNA stems where possible.
- Automated search and align0 realignment of global regions of blast hits. [Now outdated].
- Introduction of the column format: http://colformat.kvl.dk.
- Colformat motivation. Easy to work with while very flexible.
  [hence much man power in time are saved!]

RNAdbtools

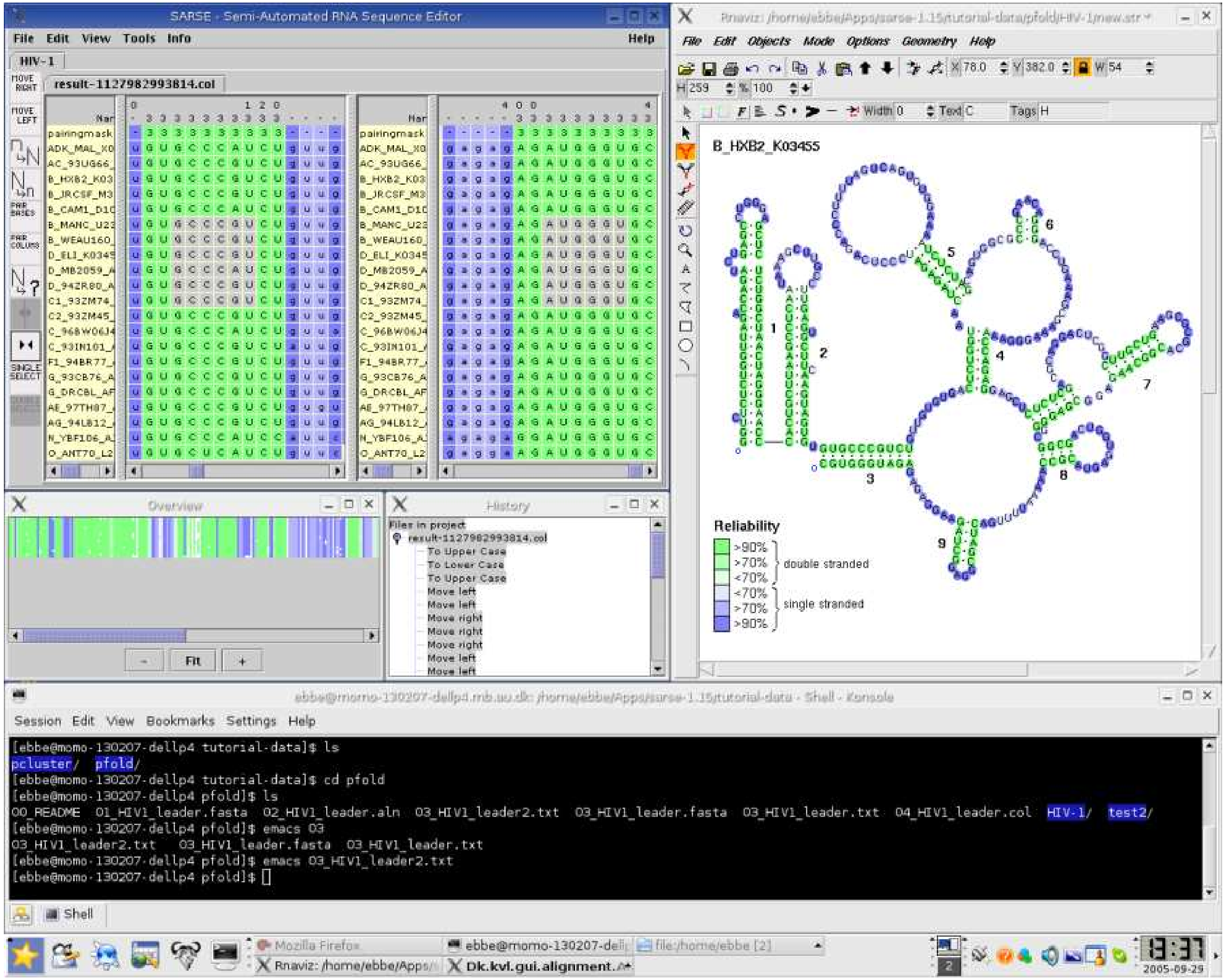
```
              1                                                                                                        100
pairing_mask  aaaaaaa--- ---------- -bbbbb-bbb b-----dddd ddcccceeee ---ffffffg ggggg---hh hhhhhh--gg gggg------ -----hhhhh
Aqu.aeo.      GGGGGCGga- aaggauu-cg aCGGGG-ACa ggcg---GUC Cc---cGAGG a--GCAGGCC GGG------- UGGCU----- CCCGuaac-- ------AGCC
The.mar.      GGGGGCGaa- -cggguu-cg aCGGGG-AUG Gagu---CCC C-----UGGG aa-gCGAGCC GAGGu---cC CCACCU---C CUCGuaaaaa -----AGGUG
The.the.      GGGGGUGaa- acggucu-cg aCGGGG-GUC gccga-gGGC Gu----GGCU ---GCGCGCC GAGG----uG CGGGUg--gC CUCGuaaaa- ------ACCC
Dei.rad.      GGGGGUGac- ccgguuu-cg aCAGGG-Gaa cugaa--GGU G-----aUGU u--GCGUGUC GAGG----uG CCGUUg--gC CUCGuaaaca ------AACG
Por.gin.      GGGGCUGa-- ccggcuu-ug aCAG-C-GUG augaa-gCGG U-----AUGU aa-GCAUGUA GUGCGu--gG GUGgCU--UG CACUauaauc u---cAGaCA
Chl.tep.      GGGGAUGa-- caggcuaucg aCAGGA-UAg gugug-aGAU GU---cGUUG ----CACUCC GAGUUucaGC AUGGAC-gGA CUCGuuaaac a---aGUCUA
Chl.tra.      GGGGGUGua- aagguuu-cg aCUUAG-aAA ugaag--CGU U-----AAUU ---GCAUGCG GAGGgc---G UUGGCUgg-C CUCCuaaaa- -----AGCCG
Chl.pne.      GGGGGUGua- uagguuu-cg aCUUGA-aAA ugaag--UGU U-----AAUU ---GCAUGCG GAGGgc---G UUGGCUgg-C CUCCuaaaa- -----AGCCA
Ana.spe.      GGGUCcGu-- -cgguuu-cg aCAGGU-UGG cgaac--GCU aC---UCUGU gauuCAGGUC gAGAG----U GAGUCUc-CU CU-Gcaaauc a----AGGCU
support       SSSSSSS... .......... .SSSSS.SSS .......SSS S.....SSSS ...SSSSSSS SSSS.....S SSSSSS...S SSSS...... .....SSSSS

              101                                                                                                      200
pairing_mask  hhh------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ------iiii
Aqu.aeo.      G--------- ---------- --cuaaaaca gcucccgaag cugaacuc-- ---------- ---------- ---------- ---------- ---------g
The.mar.      GGacaa---- ---------- -agaauaagu gccaacgaac cuguu----- ---------- ---------- ---------- ---------- ---------g
The.the.      GCaac----- ---------- -ggcauaacu gccaacacca acuac----- ---------- ---------- ---------- ---------- ---------g
Dei.rad.      GCaaagc--- ---------- -cauuuaacu ggcaaccaga acuac----- ---------- ---------- ---------- ---------- ---------g
Por.gin.      UCaaa----- ---------- -aguuuaauu ggcgaaaaua a--------- ---------- ---------- ---------- ---------- -----cuaCG
Chl.tep.      UGUa------ ---------- -ccaauagau gcagacgauu auucguau-- ---------- ---------- ---------- ---------- ---------g
Chl.tra.      ACaaa----- ---------- -acaauaaau gccgaaccua aggcugaaug cgaaauuauc a--------- ------gcuu cgcugaucuc gaagaucuaa
Chl.pne.      ACaaa----- ---------- -acaauaaau gccgaaccua aggcugaaug cgaaauuauu a--------- ------gcuu guuugacuca guagagGAaa
Ana.spe.      CAaaaca--- ---------- -aaaguaaau gcgaauaaca ucguuaaauu u--------- ---------- ---------- ----gcucgu aaggacgcUc
support       SS........ .......... .......... .......... .......... .......... .......... .......... .......... ..........

              201                                                                                                      300
pairing_mask  iiiiijjjjj j--------- ---------- ---------- ---------- jjjjjjiiii iiiii----- -------kkk kkkkk---ll llllmmmmmm
Aqu.aeo.      cucuC-GCUG Ccuaauuaaa ---------- ---------- --------cg GCAGC-G--- ---------- --------CG UCC-----CC GGUA--GGUU
The.mar.      cuguu-GCCG Cuuaauaga- ---------- ---------- -------uaa GCGGC----- ---------- --------CG UCCUC---UC CGA----AGU
The.the.      cucUC-GCGG cuuaau---- ---------- ---------- ---------g aCCGC-GA-- ---------- --------CC UCGC----CC GGU---AGCC
Dei.rad.      cUCUC-GCUg cuu------- ---------- ---------- ---------- -aAGU-GAGA ---------- --------UG ACga-----C CGUG-caGCC
Por.gin.      CUcUC-GCUg cguaaucgaa gaauaguaga uuagacgcuu caucgccgcc aaAGU-GGcA GCGacgaga- --------CA UCGCc----C GAGC-aGCUU
Chl.tep.      caAuG-GCuG ccugauua-- --------gc acaag-uuaa uucagaagcc aUcGU-CcUg cggugaaugc gcuuacucUG AAgCCg--cc ggauGGCaua
Chl.tra.      gAGUA-GCUg cuuaauua-- ---------- ---------- --------gc aaAGU-UGUU accuaaauac gggugaccCG GUGU-----U CGcG--AGCU
Chl.pne.      GaCUA-GCUG cuuaauua-- ---------- ---------- gcaaaaguug uUAGC-UAGa UaaUCucuag g--uaaccCG GUAU----cU GCG---AGCU
Ana.spe.      uagua-GcuG ccuaa----- ---------- ---------- ---auagccu cUuuC-aggu ucGagc---- --------GU CUU------C GGUUuga---
support       ....S.SSSS .......... .......... .......... .......... .SSSS.S... .......... ........SS SSSS.....S SSSS..SSSS

              301                                                                                                      400
pairing_mask  ---------- -nnnnnnnmm mmmm--oooo ----oooo-l lllll----- -kkkkkkkk- ---------- nnnnnnn--- ---------p pp------pp
Aqu.aeo.      ---------- uGCGGGU-GG CC-------- ---------U ACCGGa---- ---GGGCGuc ag-----aga cACCCGCuc- ---------- ----------
The.mar.      u--------- -GGCUGG-GC U--------- ---------- UCGGA----- aGAGGGCGug ag-----aga uCCAGCCua- ---------- ----------
The.the.      cu-------- -GCCGGG-GG CU-------- ---------c ACCGGa---- -aGCGGGGac ac------aa aCCCGGCua- ---------- ----------
Dei.rad.      c--------- -GGCCUUUGG Cgu------- ---------C GCGG------ --aaGUCAcu aaa----aaa GAAGGCUag- ---------- ----------
Por.gin.      u--------- -UUCCCG-AA GU-------- --------aG CUCGa---ug guGCGGUGcu gac----aaa uCGGGAAcc- ---------- ----------
Chl.tep.      accc------ gcGCUUG--- aGCCu----- ---------- ---------a cGGgUUCGcg caa------- gUAAGC---- ---------- ----------
Chl.tra.      cc-------- -ACCAGA-GG UU-------- ---------U uCGAa----- --ACACCGuc a------ugu aUCUGGUua- ---------- ----------
Chl.pne.      cc-------- -ACCAGA-GG CU-------- ---------- UGCAa----- -aAUACCGuc a------uuu aUCUGGUug- ---------- ----------
Ana.spe.      ---------- cUCCGUU-aa ---g------ ---------G ACUG------ ---AAGACca ac-----ccc cAACGGAugc u--------- ----------
support       .......... .SSSSSS.SS SS........ .........S SSSS...... ..SSSSSS.. .......... .SSSSSS... .......... ..........

              401                                                                                                      500
pairing_mask  p--------- ------qqqq qqqqqqrrrr rr-sssssss ssrrrrrrqq qqqqqqqq-- -------sss ssssss--tt ttttyuuuuu --vvvv---w
Aqu.aeo.      ---------- ------gGGC UACU--CGGU c---GCACGG G---GCUG-- AGUAGCUgac acc--uaaCC CGUGC----U aCCC-UCGGG -gaGCUu---
The.mar.      ---------- -------CCG AUUCA-GcUU c---GCCUUC C---GGcC-U GAAUCGGgaa aac--ucaGG AAGGCug--- UGGG-AGaGG acacCCu---
The.the.      ---------- -------GCC CGGG--GccA c---GCCCUC ----UaaC-- CCCGGGCgaa gcu--ugaaG GGGGCuc--- gCUC-CUGGC --cGCCc---
Dei.rad.      ---------- -------CcC a------GGC gauuCUCCAU A---GCCgac ---gGcGaaa cu-----uUA UGGAGcuacG GCCu--GCGa -gaACCu---
Por.gin.      ---------- -------GCU A------CAG GaugCUUCCu -g-CCUG--- ---UGGUcag auc---gaac GGAAGaua-a gGAu-CGuGC auuGGGuc--
Chl.tep.      ---------- -------UCC GUACAU-UCa U---gCCCGA ---GgGG-GU GUGCGGGuaa cc-----aaU CGGG-aua-a gGGg-aCGAa ---cGCug--
Chl.tra.      ---------- ------GAAC UUAgguccUU Ua-AUUCUCG ---AGGaa-- aUGAGUUUga aau--uuaaU GAGAGU---C GUUA-GUCUC u-aUAGg--G
Chl.pne.      ---------- ------GAAC uuACuuUcUC Ua-AUUCUCA ---AGGaA-- GUucGUUCga gau---uuuU GAGAGU---C AUUG-GCUgC u-aUAGag-g
Ana.spe.      ---------- -------CUA GCAAu---GU U---CUCUGG UU-GGC---- UUGCUAGcu- aagauuuAAU CAGAGc---- ---------- ----------
support       .......... .......SSS SSSS..S.SS ....SSSSSS ....SS.S.. SSSSSSS... .........S SSSSS..... .SS..SSSSS ...SSS....
```

We would like to have SARSE

- Make RNAdbtools interactive.
- Features: Split view; drop drag; highlight complement bases; unlimited undo/redo sessions; overview (click and jump to region); history window.
- Integrate your own commandline software into SARSE
  [by dumping data in the colformat].
- Extends RNA stems where possible.
- Automated search and align0 realignment of global regions of blast hits.
  [Now outdated, but similar stuff could be included in SARSE].
- SARSE: Jave based interface. Basic editor funtions directly incoorporated.
- Other software can be executed: RNAdbtools, pfold and pcluster [NEW !].
- Extensive documentation: http://sarse.kvl.dk.

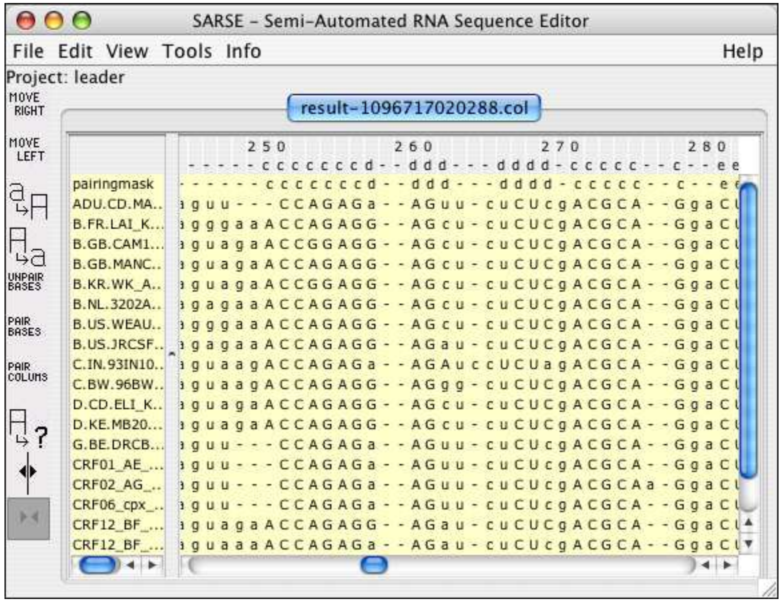
Semi-Automated RNA Sequence Editor

Clean up RNA multiple structural alignments. http://sarse.kvl.dk (See intro)



```
[ebbe@momo-130207-dellp4 tutorial-data]$ ls
pcluster/  pfold/
[ebbe@momo-130207-dellp4 tutorial-data]$ cd pfold
[ebbe@momo-130207-dellp4 pfold]$ ls
00_README  01_HIV1_leader.fasta  02_HIV1_leader.aln  03_HIV1_leader2.txt  03_HIV1_leader.fasta  03_HIV1_leader.txt  04_HIV1_leader.col  HIV-1/  test2/
[ebbe@momo-130207-dellp4 pfold]$ emacs 03
03_HIV1_leader2.txt   03_HIV1_leader.fasta  03_HIV1_leader.txt
[ebbe@momo-130207-dellp4 pfold]$ emacs 03_HIV1_leader2.txt
[ebbe@momo-130207-dellp4 pfold]$ 
```

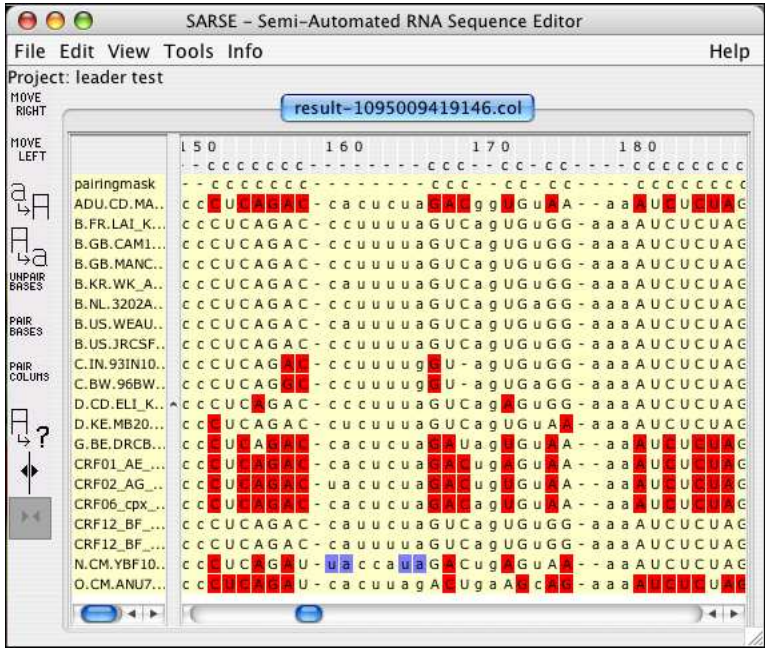
SARSE: Semi-Automated RNA Sequence Editor


SARSE – Semi-Automated RNA Sequence Editor

Project: leader

result-1096717020288.col

```
                250       260       270       280
              -----ccccccccd--ddd---dddd-ccccc--c--ee
pairingmask   ------ccccccccd--ddd---dddd-ccccc--c--e
ADU.CD.MA..   aguu---CCAGAGa--AGuu-cuCUcgACGCA--GgaCU
B.FR.LAI_K... agggaaACCAGAGG--AGcu-cuCUCgACGCA--GgaCU
B.GB.CAM1...  aguagaACCGGAGG--AGcu-cuCUCgACGCA--GgaCU
B.GB.MANC..   aguagaACCAGAGG--AGcu-cuCUCgACGCA--GgaCU
B.KR.WK_A..   aguagaACCGGAGG--AGcu-cuCUCgACGCA--GgaCU
B.NL.3202A..  agagaaACCAGAGG--AGcu-cuCUCgACGCA--GgaCU
B.US.WEAU..   agggaaACCAGAGG--AGcu-cuCUCgACGCA--GgaCU
B.US.JRCSF..  agagaaACCAGAGG--AGau-cuCUCgACGCA--GgaCU
C.IN.93IN10.. aguaagACCAGAGa--AGAuccUCUagACGCA--GgaCU
C.BW.96BW..   aguaagACCAGAGG--AGgg-cuCUCgACGCA--GgaCU
D.CD.ELI_K..  aguagaACCAGAGG--AGcu-cuCUCgACGCA--GgaCU
D.KE.MB20...  aguagaACCAGAGG--AGcu-cuCUCgACGCA--GgaCU
G.BE.DRCB...  aguu---CCAGAGa--AGuu-cuCUcgACGCA--GgaCU
CRF01_AE_...  aguu---CCAGAGa--AGuu-cuCUcgACGCA--GgaCU
CRF02_AG_...  aguu---CCAGAGa--AGuu-cuCUcgACGCAa-GgaCU
CRF06_cpx_... aguu---CCAGAGa--AGuu-cuCUcgACGCA--GgaCU
CRF12_BF_...  aguagaACCAGAGG--AGau-cuCUCgACGCA--GgaCU
CRF12_BF_...  aguaaaACCAGAGa--AGau-cuCUcgACGCA--GgaCU
```


SARSE – Semi-Automated RNA Sequence Editor

Project: leader test

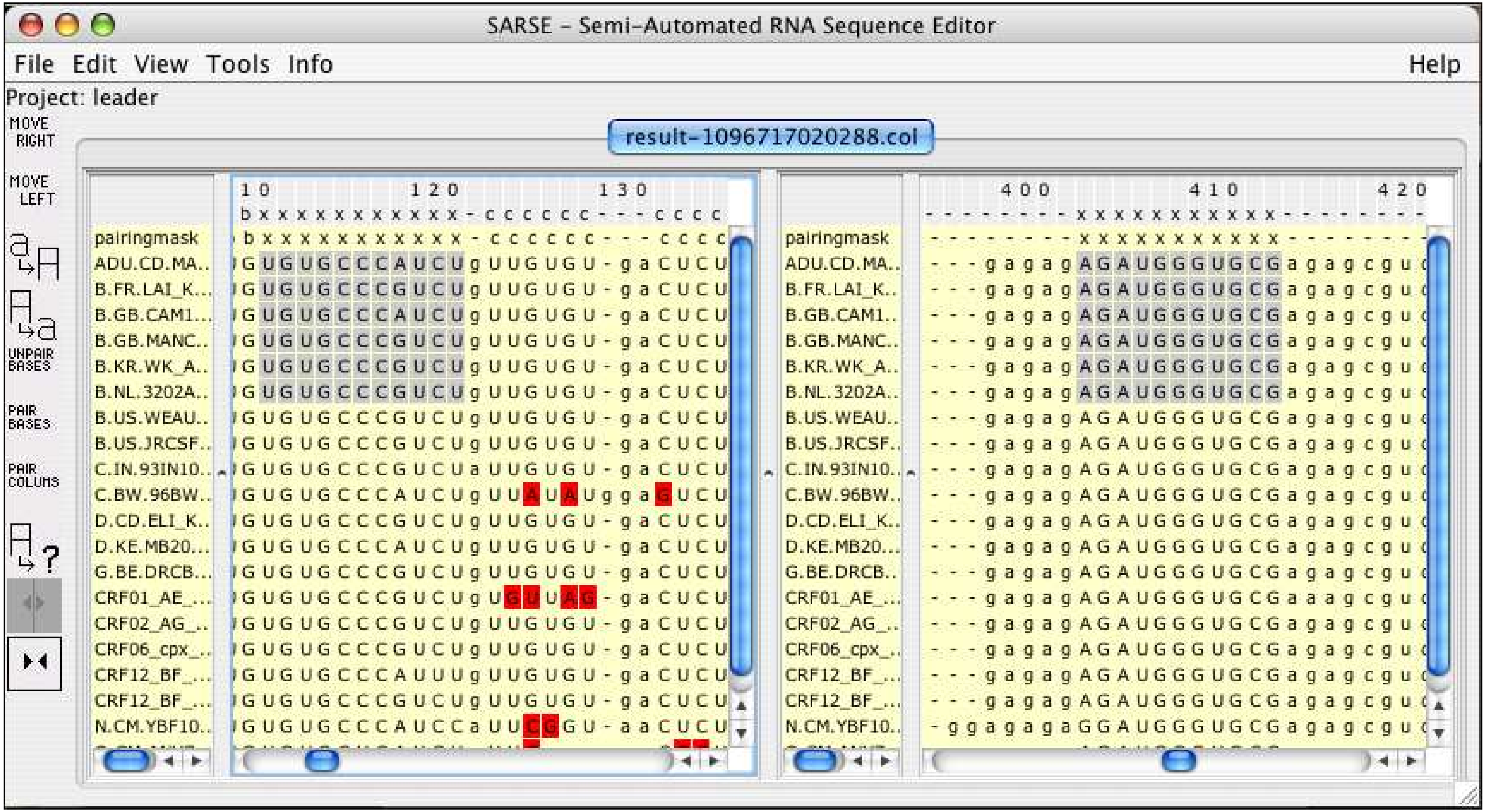
result-1095009419146.col

```
              150       160       170       180
              --ccccccc--------ccc--cc-cc----cccccccc
pairingmask   --ccccccc--------ccc--cc-cc----cccccccc
ADU.CD.MA..   ccCUCAGAC-cacucuaGACggUGuAA--aaAUCUCUAG
B.FR.LAI_K... ccCUCAGAC-ccuuuuaGUCagUGuGG-aaaAUCUCUAG
B.GB.CAM1...  ccCUCAGAC-ccuuuuaGUCagUGuGG-aaaAUCUCUAG
B.GB.MANC..   ccCUCAGAC-ccuuuuaGUCagUGuGG-aaaAUCUCUAG
B.KR.WK_A..   ccCUCAGAC-cauuuuaGUCagUGuGG-aaaAUCUCUAG
B.NL.3202A..  ccCUCAGAC-ccuuuuaGUCagUGaGG-aaaAUCUCUAG
B.US.WEAU..   ccCUCAGAC-cauuuuaGUCagUGuGG-aaaAUCUCUAG
B.US.JRCSF..  ccCUCAGAC-ccuuuuaGUCagUGuGG-aaaAUCUCUAG
C.IN.93IN10.. ccCUCAGAC-ccuuuugGU-agUGuGG-aaaAUCUCUAG
C.BW.96BW..   ccCUCAGGC-ccuuuugGU-agUGaGG-aaaAUCUCUAG
D.CD.ELI_K..  ccCUCAGAC-cccuuuaGUCagAGuGG-aaaAUCUCUAG
D.KE.MB20...  ccCUCAGAC-cucuuuaGUCagUGuAA-aaaAUCUCUAG
G.BE.DRCB...  ccCUCAGAC-cacucuaGAUagUGuAA--aaAUCUCUAG
CRF01_AE_...  ccCUCAGAC-cacucuaGACugAGuAA--aaAUCUCUAG
CRF02_AG_..   ccCUCAGAC-uacucuaGACugUGuAA--aaAUCUCUAG
CRF06_cpx_... ccCUCAGAC-cacucuaGACagUGuAA--aaAUCUCUAG
CRF12_BF_...  ccCUCAGAC-cauucuaGUCagUGuGG-aaaAUCUCUAG
CRF12_BF_...  ccCUCAGAC-cauuuuaGUCagUGuGG-aaaAUCUCUAG
N.CM.YBF10..  ccCUCAGAU-uaccauaGACugAGuAA--aaAUCUCUAG
O.CM.ANU7...  ccCUCAGAU-cacuuagACUgaAGcAG-aaaAUCUCUAG
```

- Invoke RNADBTOOLS to point out inconsistencies.
- Small overview box (not shown) gives a global view.
- Data from programs dumped in a projects directory.

SARSE: Semi-Automated RNA Sequence Editor

Split view

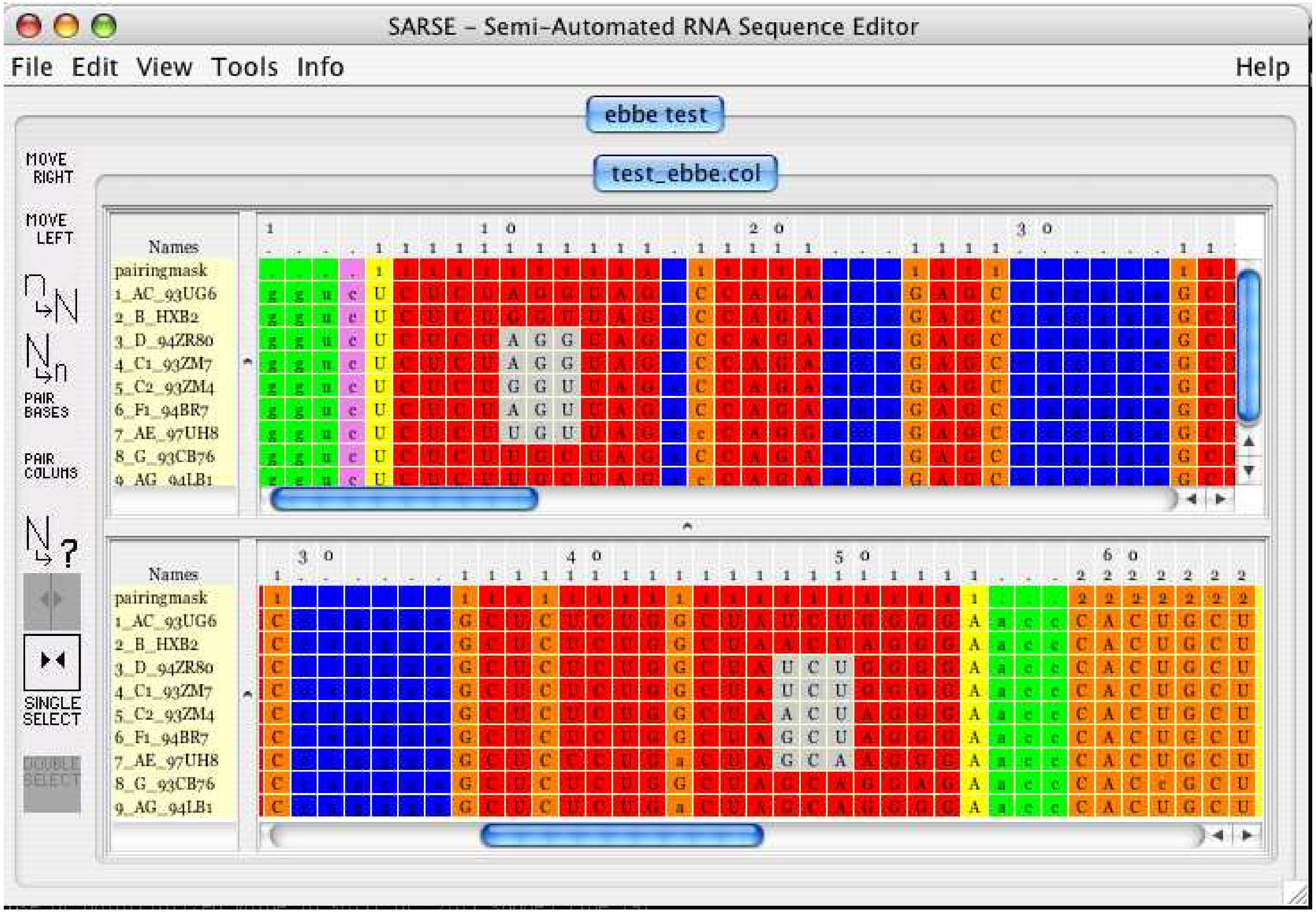


Project: leader

```
             10          120         130                          400          410          420
             b x x x x x x x x x x x - c c c c c c - - - c c c c  - - - - - - - - x x x x x x x x x x x - - - - - - - -
pairingmask  b x x x x x x x x x x x - c c c c c c - - - c c c c  - - - - - - - - x x x x x x x x x x x - - - - - - - -
ADU.CD.MA..  G U G U G C C C A U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
B.FR.LAI_K.. G U G U G C C C G U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
B.GB.CAM1... G U G U G C C C A U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
B.GB.MANC..  G U G U G C C C G U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
B.KR.WK_A..  G U G U G C C C G U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
B.NL.3202A.. G U G U G C C C G U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
B.US.WEAU..  G U G U G C C C G U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
B.US.JRCSF.. G U G U G C C C G U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
C.IN.93IN10.. G U G U G C C C G U C U a U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
C.BW.96BW..  G U G U G C C C A U C U g U U A U A U g g a G U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
D.CD.ELI_K.. G U G U G C C C G U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
D.KE.MB20... G U G U G C C C A U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
G.BE.DRCB... G U G U G C C C G U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
CRF01_AE_... G U G U G C C C G U C U g U G U U A G - g a C U C U  - - - g a g a g A G A U G G G U G C G a a a g c g u
CRF02_AG_..  G U G U G C C C G U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
CRF06_cpx_.. G U G U G C C C G U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
CRF12_BF_... G U G U G C C C A U U U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
CRF12_BF_... G U G U G C C C G U C U g U U G U G U - g a C U C U  - - - g a g a g A G A U G G G U G C G a g a g c g u
N.CM.YBF10.. G U G U G C C C A U C C a U U C G G U - a a C U C U  - g g a g a g a G G A U G G G U G C G a g a g c g u
```

SARSE: Semi-Automated RNA Sequence Editor

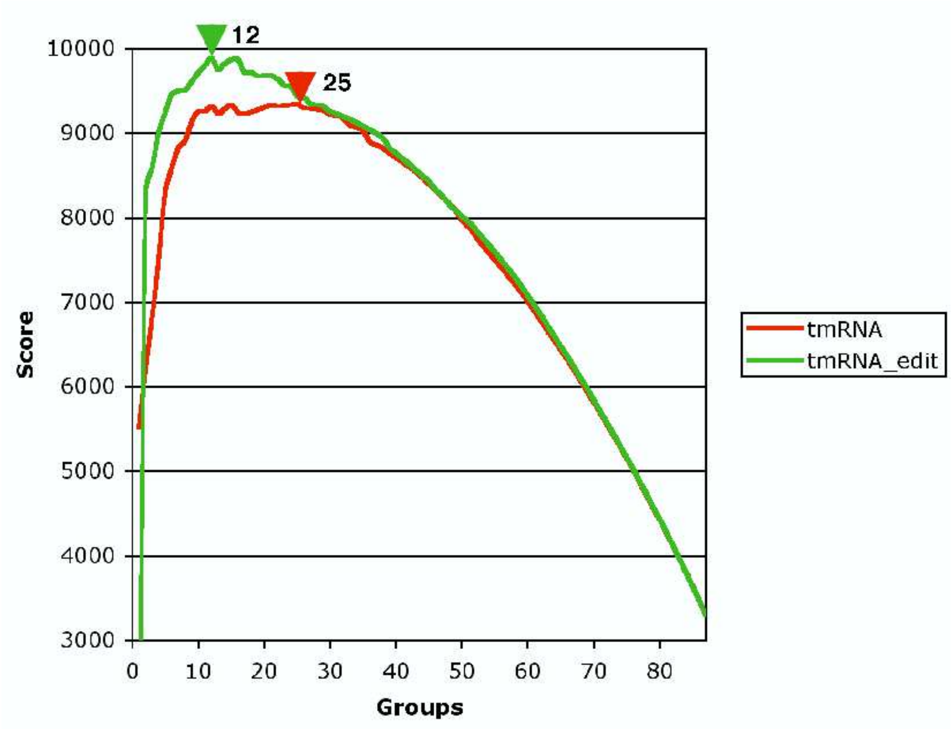
Coloring acocrding to Pfold reliability scores:

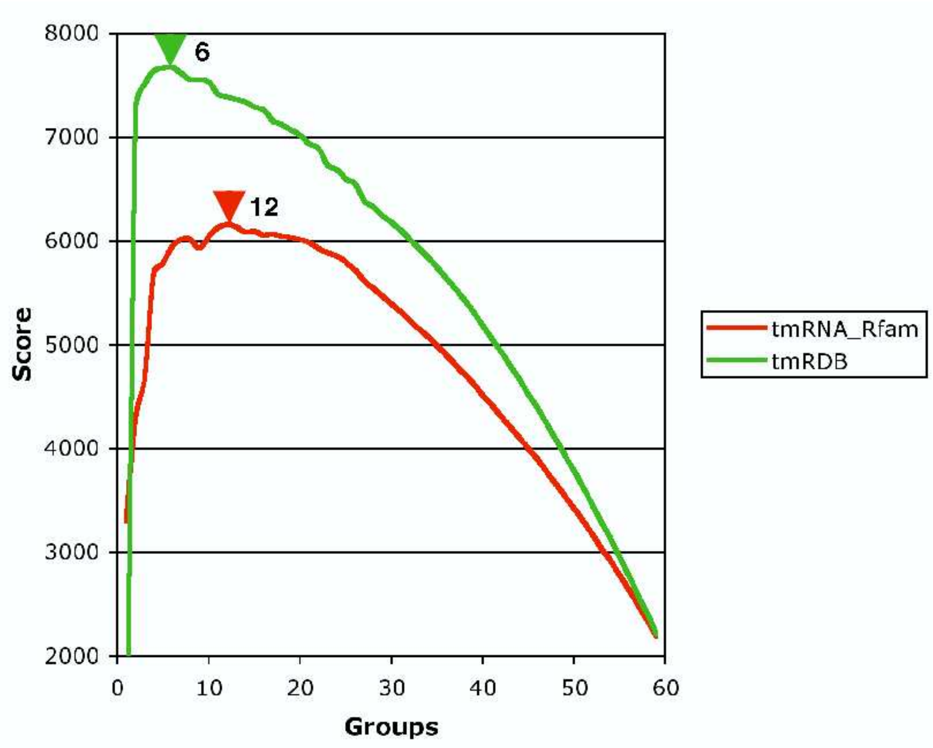


Semi-Automated RNA Sequence Editor: Pcluster

- pfold structure disrupted: poor alignment or variations in structure.
- Detect structural subgroups.
- Score: Reliability weighted sum of base pairs.
- Greedy heuristic method by joining subgroups with highest score.
- Extract: number of subgroups for max score.
- Heuristic to find "best" groupings by interpolation between max score point that score half of max point.

Semi-Automated RNA Sequence Editor: Pcluster





... and to something not so different:

You yeah I mean you: you train and test your prediction methods on these data

SARSE perspectives

- Adding even more programs. RNAalifold etc.
- Adding: auto fecthing of sequences from databases.
- Web server set up.
- Web suites for specialist to curate their set of families.

Part II The K category correlation coefficient

Motivation:

- Predictions can yield more than dichotomies.
  [Eg. protein secodndary structure]

- RNA secondary structure predictions: bp, ¬bp, unassignable.

- Exists measure for comparing multiple categories
  [eg: Escoufier (1973) and review by Baldi *et al.*. (2000).]

- None of the measures have completely desired properties.

- Goal: an extension of Matthews correlation coefficient:

$$C = \frac{P_t N_t - P_f N_f}{\sqrt{(N_t + N_f)(N_t + P_f)(P_t + N_f)(P_t + P_f)}}$$

- Idea: Simple extension of Pearson’s correlation coefficient.

Note on Matthews correlation coefficient for ncRNA evaluation

(... and to something *slightly* different!)

Basepair prediction N_t factor N larger than P_t, P_f, N_f. [$N(N-1)/2$ pairs of bases.]

$$\begin{aligned} C &= \frac{P_t N_t - P_f N_f}{N_t\sqrt{(1+N_f/N_t)(1+P_f/N_t)(P_t+N_f)(P_t+P_f)}} \\ &\approx \frac{P_t N_t - P_f N_f}{N_t\sqrt{(P_t+N_f)(P_t+P_f)}} \\ &= \frac{P_t}{\sqrt{(P_t+N_f)(P_t+P_f)}}\left[1-\frac{P_f N_f}{P_t N_t}\right] \end{aligned}$$

where $N_f/N_t \to 0$ and $P_f/N_t \to 0$ for $N \to \infty$. For any reasonable prediction method ($P_t > 0$), with at least $P_t \sim P_f$ or $P_t \sim N_f$, we can write

$$C \approx \frac{P_t}{\sqrt{(P_t+N_f)(P_t+P_f)}} = \sqrt{\frac{P_t}{P_t+N_f}\frac{P_t}{P_t+P_f}}\,,$$

Pearson’s correlation coefficient and least square fitting

Pearson’s correlation coefficient:

$$r = \frac{COV(X,Y)}{\sqrt{COV(X,X)COV(Y,Y)}}, \qquad COV(X,Y) = \sum_{n=1}^{N} (X_n - \overline{X})(Y_n - \overline{Y})$$

For variables Y and X of length N least. Least square fitting in the coefficient b:

$$Y = a + bX$$

yield an expression for b. Conversely an similary expression can be obtanined for fitting in the coefficient b':

$$X = a' + b'Y$$

For a linear fit:

$$E = \sum_{n=1}^{N} (Y_n - (a + bX_n))^2$$

paritial derivaties in a and b should be zero. It follows that

$$r^2 = bb', \qquad b = \frac{COV(X,Y)}{COV(X,X)}$$

Extending Pearson’s correlation coefficient to two K-dimensional tables, the R_K coefficient

Consider two $N \times K$ tables: $\underline{\underline{X}}$ and $\underline{\underline{Y}}$. Define

$$COV(\underline{\underline{X}},\underline{\underline{Y}}) \;=\; \sum_{k=1}^{K} w_k COV(\underline{\underline{X}}_k,\underline{\underline{Y}}_k) = \frac{1}{K}\sum_{n=1}^{N}\sum_{k=1}^{K}(X_{nk}-\overline{X}_k)(Y_{nk}-\overline{Y}_k)$$

where $\overline{X}_k = \frac{1}{N}\sum_{n=1}^{N} X_{nk}$ and $\overline{Y}_k$ are the respective means of column k. Use ("prior") $w_k = 1/K$.

$$R_K = \frac{COV(\underline{\underline{X}},\underline{\underline{Y}})}{\sqrt{COV(\underline{\underline{X}},\underline{\underline{X}})COV(\underline{\underline{Y}},\underline{\underline{Y}})}}$$

Basic properties: $R_K \in [-1,1]$; $R_1 = r$; $R_2 = r$, when $X_{n1}+X_{n2}=a$ and $Y_{n1}+Y_{n2}=b$. Hence R_2 reduces to Matthews correlation coefficient when X and Y components only take the values $\{0,1\}$.

Relation to least square fitting

K related linear fits $\vec{Y} = \vec{a} + b\vec{X}$ over the N data points. $K = 1$: Pearson case.

Weighted difference in a cost function:

$$E = \sum_{n=1}^{N} \sum_{k=1}^{K} w_k \left(Y_{nk} - (a_k + bX_{nk})\right)^2$$

To obtain minimum. Require: $\partial E / \partial a_k = 0$ (for all $k = 1, \ldots, K$) and $\partial E / \partial b = 0$. After a little algebra:

$$\sum_{k=1}^{K} w_k \left(\sum_{n=1}^{N} X_{nk} Y_{nk} - N \overline{X}_k \overline{Y}_k \right) = b \left\{ \sum_{k=1}^{K} w_k \left(\sum_{n=1}^{N} X_{nk}^2 - N \overline{X}_k^2 \right) \right\}$$

yielding

$$b = \frac{COV\left(\underline{\underline{X}}, \underline{\underline{Y}}\right)}{COV\left(\underline{\underline{X}}, \underline{\underline{X}}\right)} \qquad \text{and} \qquad R_K^2 = bb'$$

The Discrete version of R_K

- The $K \times K$ *confusion* matrix $\underline{\underline{C}}$.
- Let C_{kl} be the number X_{nk}'s predicted to belong to class k, but belong to class l, $l \neq k$.
- For $K = 2$: C_{11}: true positives; C_{22}: true negatives; C_{12}: false positives; C_{21}: false negatives.
- Well known observations:
  - $N = \sum_{kl} C_{kl}$.
  - $\overline{X}_k = \frac{1}{N} \sum_l C_{kl}$.
  - $\overline{Y}_k = \frac{1}{N} \sum_l C_{lk}$.
  - $C_{kk} = \sum_n X_{nk} Y_{nk}$.

The Discrete version of R_K

Plug in the known observations to R_K and obtain

$$R_K = \frac{\sum\limits_{klm} C_{kk}C_{lm} - C_{kl}C_{mk}}{\sqrt{\sum\limits_k \left(\sum\limits_l C_{kl}\right)\left(\sum\limits_{\substack{l' \\ k' \neq k}} C_{k'l'}\right)}\sqrt{\sum\limits_k \left(\sum\limits_l C_{lk}\right)\left(\sum\limits_{\substack{l' \\ k' \neq k}} C_{l'k'}\right)}}$$

or equivalently

$$R_K = \frac{N\,Tr(\underline{\underline{C}}) - \sum_{kl} \tilde{\underline{\underline{C}}}_k \hat{\underline{\underline{C}}}_l}{\sqrt{N^2 - \sum_{kl} \tilde{\underline{\underline{C}}}_k (\hat{\underline{\underline{C}}^\top})_l}\sqrt{N^2 - \sum_{kl} (\tilde{\underline{\underline{C}}^\top})_k \hat{\underline{\underline{C}}}_l}}$$

- $\tilde{\underline{\underline{C}}}_k$ the kth row of $\underline{\underline{C}}$.
- $\hat{\underline{\underline{C}}}_l$ the lth column of $\underline{\underline{C}}$.
- $\underline{\underline{C}}^\top$ is $\underline{\underline{C}}$ transposed.

Applications of R_K

Comparison to other measures of evaluating protein secondary structure predictions [From EVA (Rost and Co-workers)]

- Numerous approaches for protein secondary structure prediction.
- Predicting the three classes, α-helix, β-sheet and coil.
- Comparing to Q_3 ranking, the fraction of correctly predictions over all three clasess.
- Comparing to SOV (Segment OVerlap), measure that take continuous stretches of helices and sheet into consideration in the evaluation.

Applications of R_K

- Eva (as of August 2003) have several classes of different set sizes.

- Each set covers different number of predictions methods.

R	1	0	0	1	0	0
L	1	0	0	1	0	0
R	1	0	0	1	0	0
V	1	0	0	1	0	0
H	1	0	0	1	0	0
Q	1	0	0	1	0	0
I	1	0	0	1	0	0
A	1	0	0	1	0	0
E	1	0	0	1	0	0
E	1	0	0	1	0	0
H	0	0	1	1	0	0
G	0	0	1	0	0	1
L	0	1	0	0	0	1
R	0	1	0	0	0	1
H	0	1	0	0	0	1
D	0	1	0	0	0	1
S	0	1	0	0	0	1
S	0	1	0	0	0	1
G	0	0	1	0	0	1
E	0	0	1	0	0	1
G	0	0	1	0	0	1
K	0	0	1	0	0	1
	α	β	C	α	β	C

Set	method	R_3	rank	sov	Q_3
1	profsec	0.621	1	74.8	75.09
	psipred	0.619	1	73.6	75.11
	apssp2	0.613	1	71.4	74.74
	samt99_sec	0.613	1	71.1	74.68
	sspro2	0.598	1	69.1	73.81
	phdpsi	0.581	1	69.7	72.61
	jpred	0.570	1	70.3	71.81
	prospect	0.567	1	69.8	71.77
	prof_king	0.555	1	69.8	70.83
	phd	0.526	2	64.5	69.01
2	profsec	0.600	1	71.5	74.00
	samt99_sec	0.586	1	67.1	73.28
	psipred	0.579	1	69.8	72.86
	sspro2	0.573	1	67.4	72.50
	phdpsi	0.560	1	66.9	71.62
	prof_king	0.544	1	66.5	70.33
	jpred	0.536	1	66.7	69.92
	phd	0.505	2	62.6	68.03
3	profsec	0.600	1	71.5	74.00
	psipred	0.579	1	69.8	72.86
	samt99_sec	0.586	1	67.1	73.28
	phdpsi	0.560	1	66.9	71.62
	prof_king	0.544	1	66.5	70.33
	jpred	0.536	1	66.7	69.92
	phd	0.505	2	62.6	68.03
4	profsec	0.608	1	71.7	74.61
	psipred	0.591	1	71.1	73.69
	samt99_sec	0.591	1	68.9	73.69
	phdpsi	0.568	1	67.7	72.22
	jpred	0.545	1	67.5	70.63
	phd	0.512	2	64.5	68.54
5	psipred	0.608	1	71.0	74.82
	profsec	0.606	1	70.2	74.52
	samt99_sec	0.600	1	69.5	74.32
	phdpsi	0.566	2	66.6	72.13
	phd	0.533	3	64.8	69.95
6	psipred	0.617	1	72.1	75.44
	profsec	0.610	1	71.2	74.82
	phdpsi	0.565	2	67.5	72.10
	phd	0.540	3	65.9	70.43

Applications of R_K

RNA example of R_3. First some background:

- Study applying pfold to structure prediction of the HIV leader (713 nt.) using 20 aligned sequences (Knudsen *et al.*, 2004).
- Study: Compare pfold predictions to predictions with pertubed rate values.
- Aim of study: can the evolutionary rates estimated from rRNA and tRNA be applied on the much faster evolving HIV-1 sequence.
  [NOTE: this assumption is implicit for all prediction methods estimating parameters from *e.g.*, rRNA and tRNA, such as for QRNA.]
- The answer was yes; the HIV-1 prediction is fine, just within the limits stabil

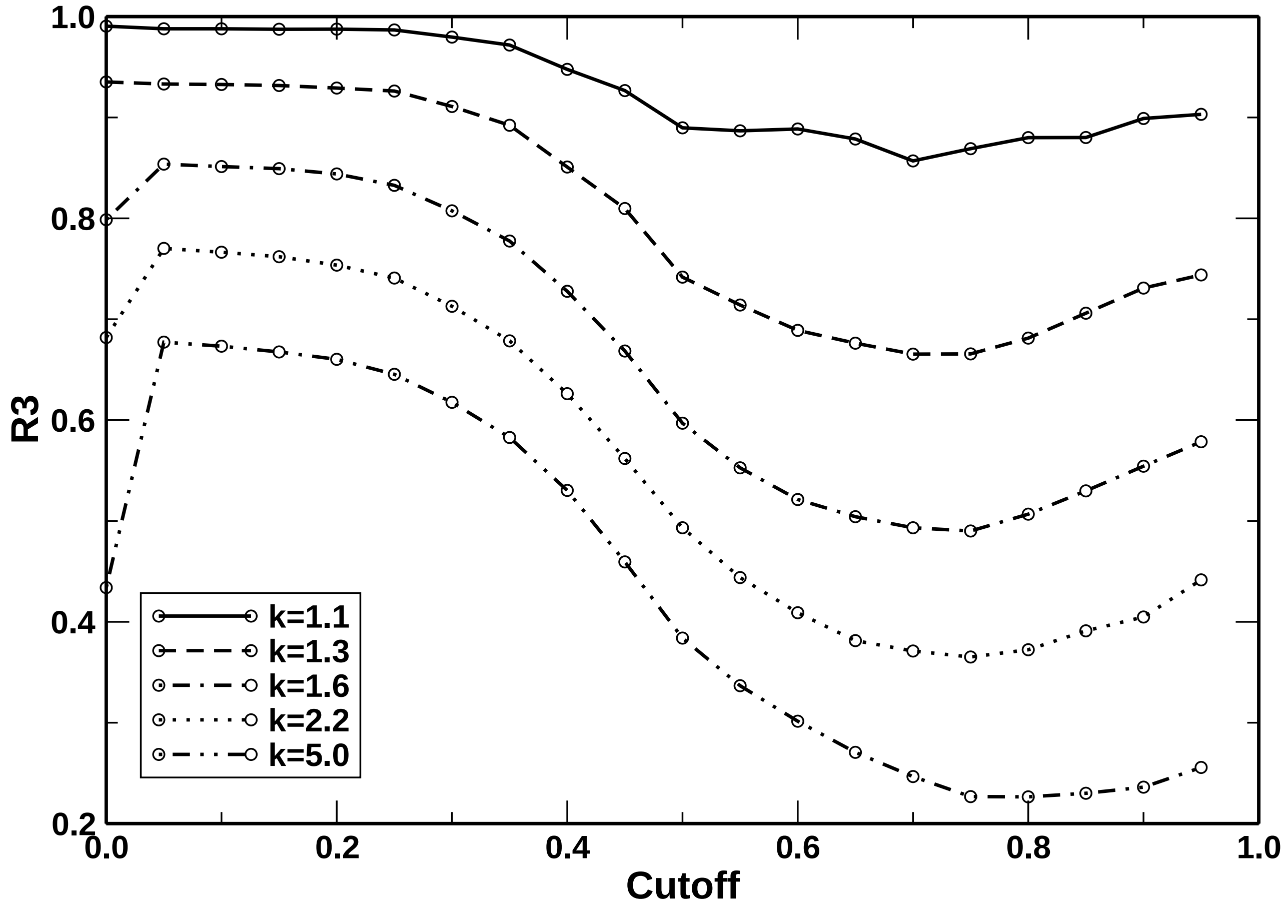
  predictions obtained when perturbing the evolutionary rates.
- Rates were essentially pertubed by having 50% chance of making the rate k times larger or k times smaller [and adding some normalization constraints].

Applications of R_K

- Introducing a third category, the unknown or unassignable categories.
- pfold predictions uncertain for low reliability score.
- Statement: If pfold score low for positions that are manually hard to assign basepairs, the overall prediction should be higher than if an assignment was enforced on these positions.
- Use different reliability score cut-offs for sending a basepair in the third category "unknown". The reference structure is still the original pfold prediction.

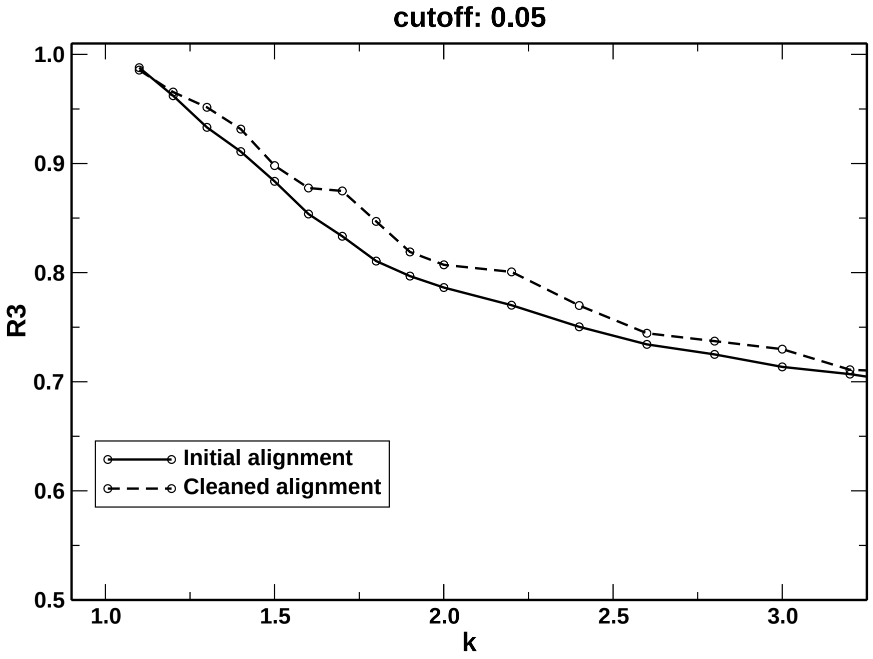
Applications of R_K

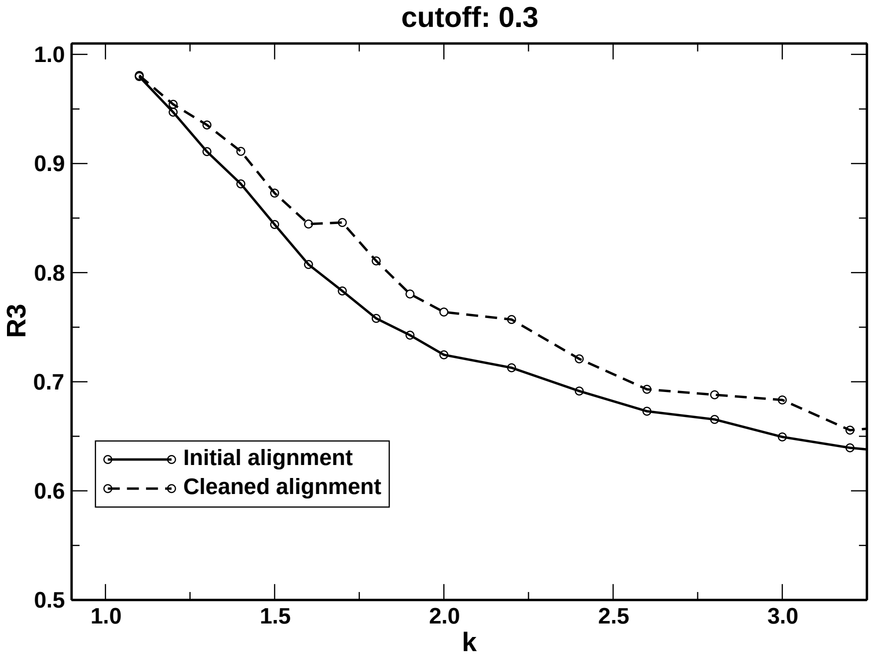
Rate variation performance for various cut-offs.



Applications of R_K

RNA strcuture predictions for different reliability (pfold) cut-offs for varying rate pertubations k. A comparison between initial and cleaned alignment.





Perspectives for R_K

Measures for comparing a predicted structure assignment to a curated structure assignment.

- Applying R_K to cases of $K > 3$.
- A measure as SOV would be needed to take prediction of entire helices into consideration.
- Further extension: Comparing $L\ N \times K$ tables and compute one correlation coefficient.