

RNA structure alignment ontology

- Part 1 - the Problem - Jim Brown
 - Traditional alignments - what are they?
 - The problems with traditional alignments
 - Part of the solution - “Correspondence”
- Part 2 - the Solution - Rob Knight

Sequence alignment group of the ROC

The task of the “Multiple sequence alignment” group is to develop definitions and concepts for both alignment and secondary structure.

Jim Brown	North Carolina State University
Neocles Leontis	Bowling Green State University
Eric Westhof	Université Louis Pasteur
Fabrice Jossinet	Université Louis Pasteur
Amanda Birmingham	Thermo Fisher Scientific
Rob Knight	University of Colorado-Boulder
Franz Lang	Université de Montréal
Rym Kachouri-Lafond	Université Louis Pasteur
Gerhard Steger	Heinrich-Heine-Universität Düsseldorf
Jesse Stombaugh	Bowling Green State University
Paul Griffiths	University of Sydney

... and anyone else who's willing to help, such as Peter Clote, Dave Mathews, Tom Bittner, Karen Eilbeck, Francois Major, &c.

What do we want to accomplish?

- Make large RNA alignments manageable
- Make the assignment of correspondence (homology or structural similarity) explicit and specific
- Allow explicit labeling and grouping of “things” in columns and rows of the alignment
- Provide tools and standards to make these work

Sequence Alignments - a review

- Sequence alignments are 2-dimensional matrices

Se-Align v2.0a11 Carbon File Edit Sequences Alignment Analysis
seq_align

1510 1515 1520 1525 1530 1535 1540 1545 1550 1555 1560 1565 1570 1575 1580 1585 1590 1595

Cyls.7417
Clrg.spHTF
Chms.sglbs
Plec.borya
Phrm.ectoc
Phrm.minut
Prtx.holla
Bor.burgdo
Spi.litora
Trp.bryant
Spi.stenos
Lps.weilii
Lps.borgpe
Lps.biflex
R.centenum
Rpl.globif
Ric.ricket
Rb.capsul2
Ps.diminut
Cau.cres2
Ntb.winogr
Bdr.japoni
Hyp.vulgar
Ag.tumefac
Bru.aborts
Vit.sterco
Alc.eutrop

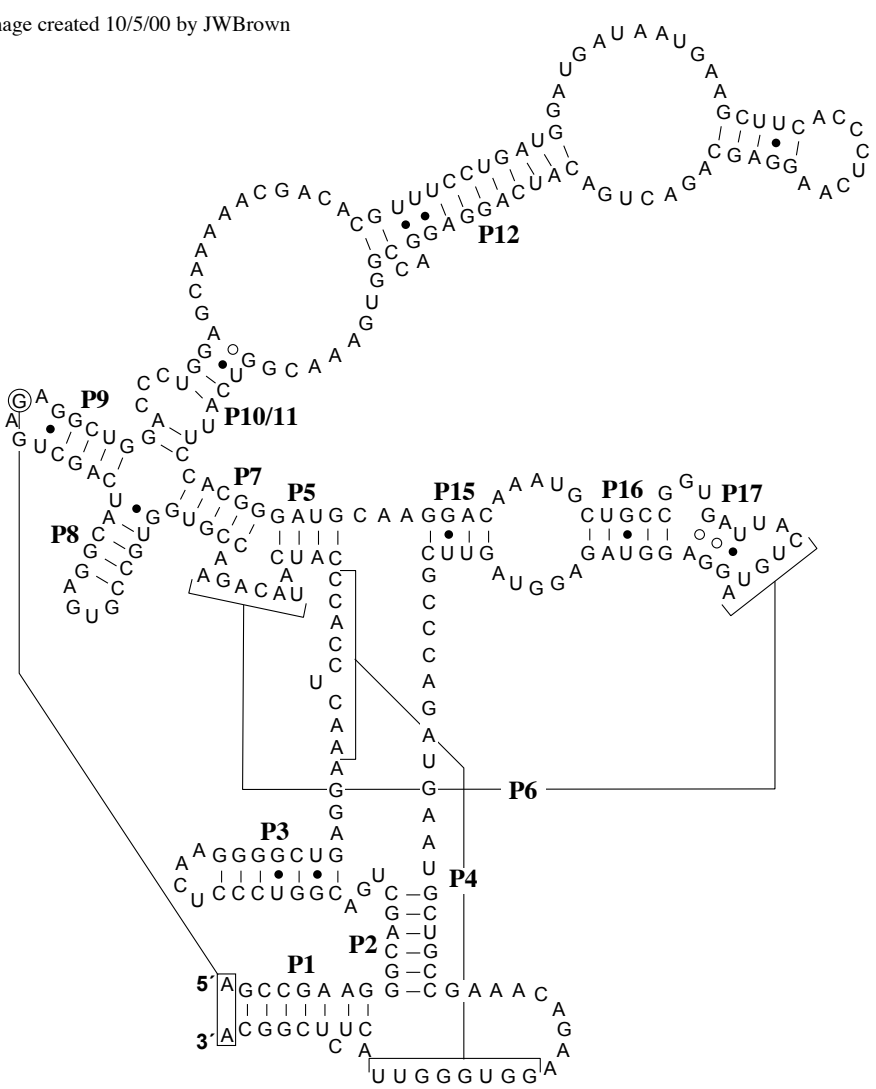
GACAGAGGGCAGCAAGCUAGCGAUAGC - AAGCUAAUCCCGG - AAACCGUAGCUCAGUUCAGAUJGCAGGCUGCAACUCGCCUGCGUGAAGG
GACAAAGGGCAGCGAGACUGCGAAGUC - AAGCAAUCCAG - AAACGAAGGCUCAGUUCAGAUJGCAGGCUGCAACUCGCCUGCGUGAAGG
GACAAAGGGCAGCCAACUGGCGACAGU - GCGCUAAUCCCAU - AAACC AUGGCUCAGUUCAGAUJGCAGGCUGCAACUCGCCUGCAUGAAGG
GACAAAGAGCAGCCAGCCAGCGAUGGU - GAGCCNAUCUCAU - AAACCGGCGCUCAGUUCAGAUJGCAGGCUGCAACUCGCCUGCAUGAAGG
GACAAAGGGCAGCCAGCUAGCGAUAGU - GAGCUNAUCCCAU - AAACCGNUGCUCAGUUCAGAUJGCAGGCUGCAACUCGCCUNCAUGAAGG
GACAAAGGGCAGCAAGCGUGCGAGCGC - AAGCUAAUCCCAU - AAACCGAGGCACAGUUCAGAUJGCAGGCUGCAACUCGCCUGCAUGAAGG
GACAAAGAGCAGCCAACUCGCGAGAGU - NAGCUNAUUCUCAU - MAACCCUNGCUCAGUUCGGAUJGCAGGCUGCAACUCGCCUNCAUGAAGU
UACAAAGCGAAGCGAAACAGUGAUGUG - AAGCAAACGCAU - AAAGCAGGUCUCAGUCCGGAUJGAAGUCUGAAACUCGACUUCUAUGAAGU
UACAGAGUGAUGCGAAGCCGCGAGGUG - AAGCAAACGCAU - AAAGCCGGUCUCAGUUCGGAUJGGAGUCUGAAACUCGACUCCUAUGAAGG
UACAGAGUGAAGCGAAGCAGUGAUGUG - GAGCAAACGCAU - AAAGCCUGCCUCAGUCCGGAUJGGAGUCUGAAACCCGACUCCUAUGAAGU
UACAGAGCGCAGCGAACC CGAGGGAU AAGCGAACC GCAA - AAAGCCGGCCGUAGUUCGGAUJGAAGUCUGAAACCCGACUUCUAUGAAGG
UACAAAGGGUAGCCAACUCGCGAGGGG - GAGCUAAUCUCA - AAAGCCGGUCCAGUUCGGAUJGGAGUCUGCAACUCGACUCCUAUGAAGU
UACAAAGGGUAGCCAACUCGCGAGGGG - GAGCUAAUCUCA - AAAUCCGGUCCAGUUCGGAUJGGAGUCUGCAACUCGACUCCUAUGAAGU
UACAGAGGGUCUCCAAACUGCC AAGUG - GAGCUAAUCUCU - AAAACCGGUCCAGUUCAGAUJGGAGUCUNCAACUCGACUCCUAUGAAGU
GACAGUGGGNAGCGACCACGCGAGUGG - AAGCGAAUCUCC - AAAAGCCAUUCAGUUCGGAUJGCACUCUGCAACUCGGGUGCAUGAAGU
GACAGUGGGACGCCAGGCCGCGAGGCN - GUGCUGAUCCCGA - AAAGGCCGUCUCAGUUCGGAUJGCACUCUGCAACUCGGGUGCAUGAAGG
UACAGAGGGAAAGCAAGACGGCGACGUG - GAGCAAUCCCU - AAAAGACAUCUCAGUUCGGAUJGUUCUCUGCAACUCGAGAGCAUGAAGU
GACAAUJGGG - - - - - CAUCCCA - AAAAGCCAUUCAGUUCGGAUJGGGUCUGCAACUCGACCCCAUGAAGU
UACAGAGGGU - - - - - UAAUCCUU - AAAAGUJGUCUCAGUUCGGAUJGUCCUCUGCAACUCGAGGGCAUGAAGU
UACAGAGGGNUNCAAUCNGGCGAGGGG - GAGCCAAUCCUN - AAAGGUJGUCUCNGUUCGGAUJGUUCUCUGCNNCUCGAGAGCAUGAAGU
GACAAUJGGAAAGCAAAGGGGUGACCC - UAGCAAUCUCA - AAAAACCGUCUCAGUUCGGAUJGGGUCUGCAACCCGAGCCCAUGAAGU
GACAAUJGGGAUGCUAAGGGGCGACCCU - UCGCAAUCUCA - AAAAGCCGUCUCAGUUCGGAUJGGGUCUGCAACUCGAGCCCAUGAAGU
GACAAUJGGGACGCAACUCAGCAAUJGG - AAGCUAAUCUCA - AAAAGCCGUCUCAGUUCGGAUJGGGUCUGCAACUCGAGCCCAUGAAGU
GACAGUGGGCAGCGAGACAGCGAUGUC - GAGCUAAUCUCC - AAAAGCCAUUCAGUUCGGAUJGCACUCUGCAACUCGAGUGCAUGAAGU
GACAGUGGGCAGCGAGCACGCGAGUGU - GAGCUAAUCUCC - AAAAGCCAUUCAGUUCGGAUJGCACUCUGCAACUCGAGUGCAUGAAGU
UACAGAGGGUAGCCAAGCCGCGAGGUG - GAGCCAAUCUCAU - AAAGCCGAUCGUAGUCCGGAUJGCACUCUGCAACUCGAGUGCGUGAAGU
UACAGAGGGUUGCCAACC CGAGGGG - GAGCUAAUCCAG - AAAACGCAUCGUAGUCCGGAUJGCAGUCUGCAACUCGACUACGUGAAGG

RNAs have well-defined 2° structure

Ribonuclease P RNA
Methanobacterium thermoautotrophicum ΔH

Sequence : U42986, Pannucci 1999 PNAS **96**:7803
 Structure : Harris, *et al.*, RNA (in press)

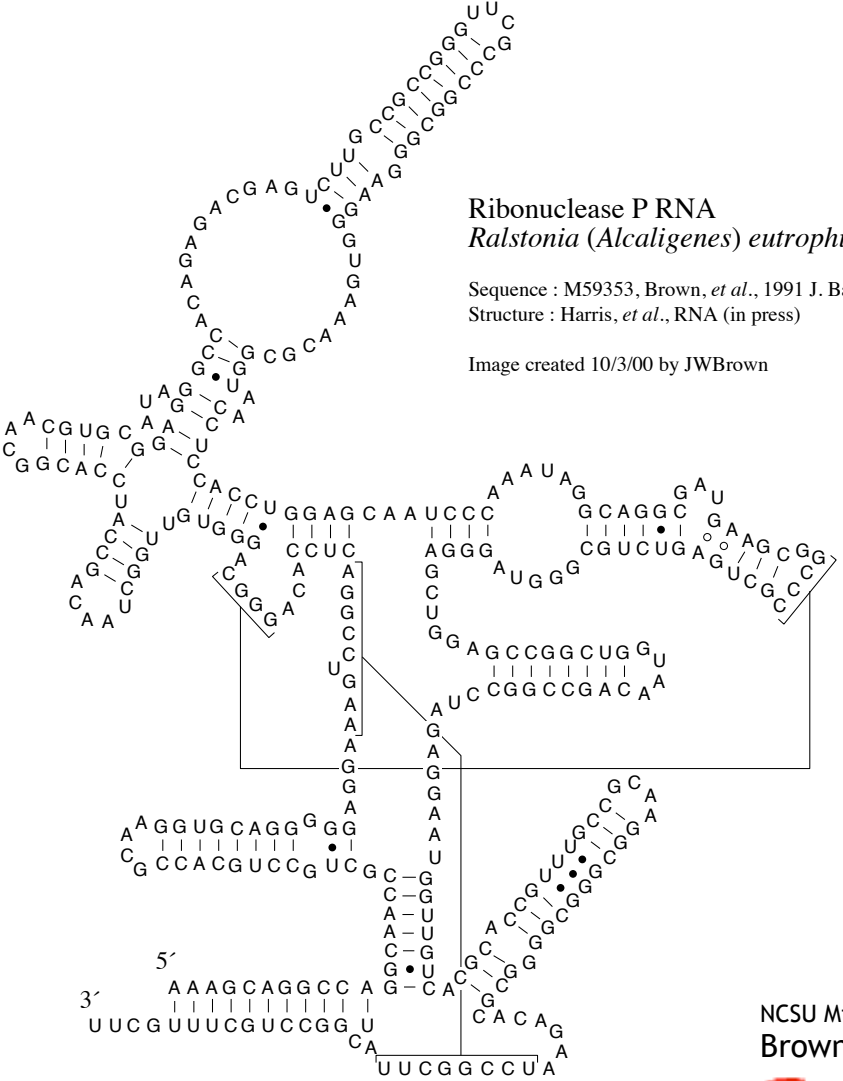
Image created 10/5/00 by JWBrown



Ribonuclease P RNA
Ralstonia (Alcaligenes) eutrophus DSM 531

Sequence : M59353, Brown, *et al.*, 1991 J. Bacteriol. **173**:3855
 Structure : Harris, *et al.*, RNA (in press)

Image created 10/3/00 by JWBrown

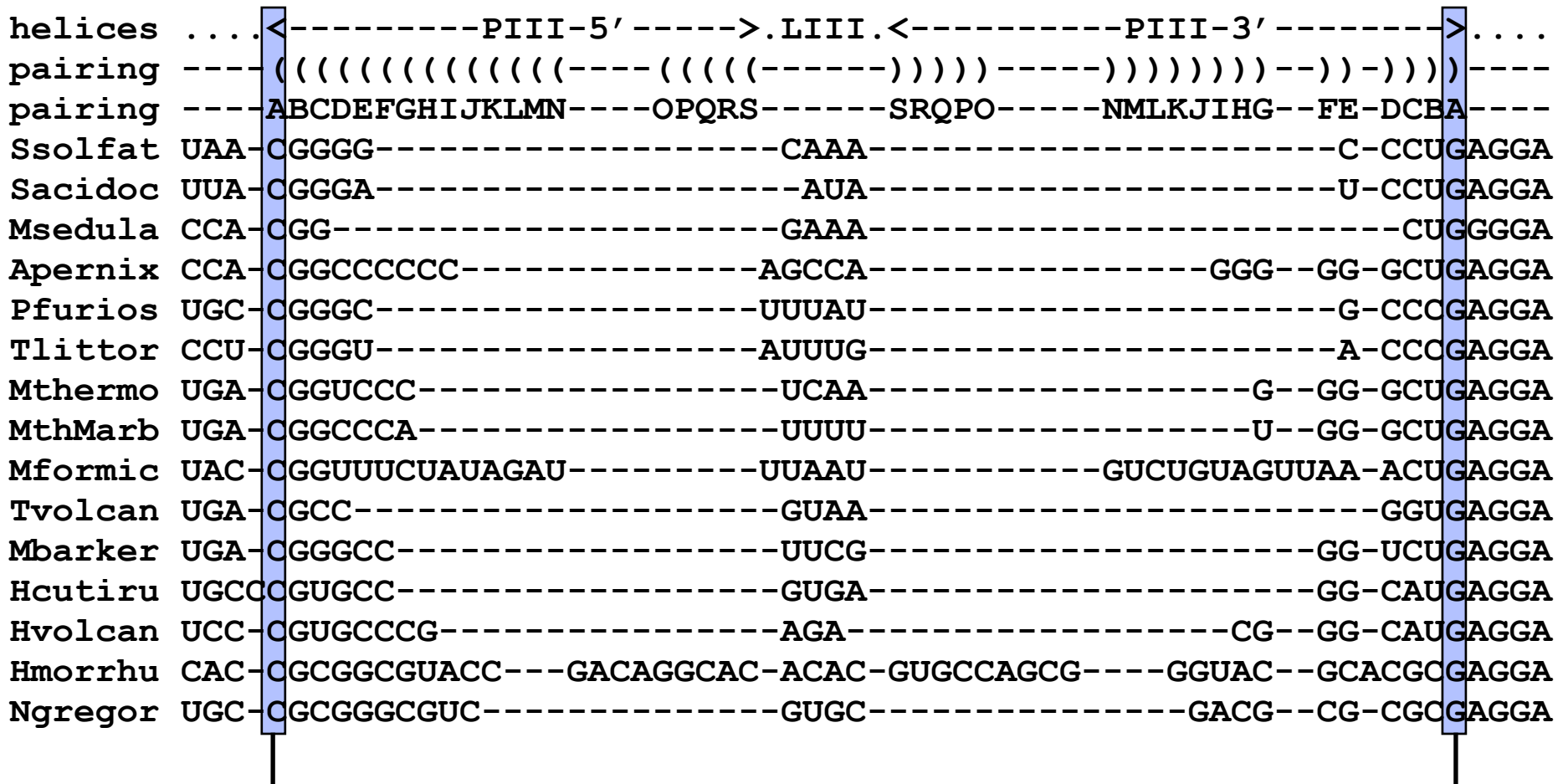


The building block of secondary structures are base pairs



RNA sequence/structure alignments

Sequences are aligned on the basis of *both* sequence similarity *and* secondary structure



If residue X of any sequence pairs to residue Y, then so must the corresponding residues in *all* sequences.

Problems with alignments

You're forced to align the helices, both *en mass* and nt-by-nt, in all sequences. Do these nts really correspond?



Problem : What about regions that are alignable between some sequences but not others?

Problems with alignments

You're forced to align the helices, both *en mass* and nt-by-nt, in all sequences. Do these nts really correspond?

helices	...<-----PIII-5'----->.LIII.<-----PIII-3'----->....
pairing	----((((((((((((((((-----((((-----))))))-----)))))))))----))--))----
pairing	----ABCDEFGHIJKLMN----OPQRS-----SRQPO----NMLKJIHG--FE-DCBA----
Ssolfat	UAA-CGGGG-----CAAA-----C-CCUGAGGA
Sacidoc	UUA-CGGGA-----AUA-----U-CCUGAGGA
Msedula	CCA-CGG-----GAAA-----CUGGGGA
Apernix	CCA-CGGCCCCC-----AGCCA-----GGG--GG-GCUGAGGA
Pfurios	UGC-CGGGC-----UUUAU-----G-CCCGAGGA
Tlittor	CCU-CGGGU-----AUUUG-----A-CCCGAGGA
Mthermo	UGA-CGGUCCC-----UCA-----G--GG-GCUGAGGA
MthMarb	UGA-CGGCCA-----UUUU-----U--GG-GCUGAGGA
Mformic	UAC-CGGUUUCUAUAGAU-----UUAU-----GUCUGUAGUUA-ACUGAGGA
Tvolcan	UGA-CGCC-----GUAA-----GGUGAGGA
Mbarke	CG-----GG-UCUGAGGA
Hcutir	GA-----GG-CAUGAGGA
Hvolcan	UCC-CGUGCCCG-----AGA-----CG--GG-CAUGAGGA
Hmorrh	CAC-CGCGGCGUACC---GACAGGCAC-ACAC-GUGCCAGCG---GGUAC--GCACGCGAGGA
Ngregor	UGC-CGCGGGCGUC-----GUGC-----GACG--CG-CGCGAGGA

The diagram illustrates sequence alignment across various species. Two vertical blue bars highlight specific regions in the sequences. A callout box with a black border and white background contains the text "Who's the corresponding nt?". Two red arrows originate from the box: one points to the 'C' in the Mthermo sequence (UGA-CGGUCCC) and the other points to the 'A' in the Mformic sequence (UAC-CGGUUUCUAUAGAU). The alignment shows that while some sequences have a 'C' at the same position, others have an 'A', creating a discrepancy in the corresponding nucleotide.

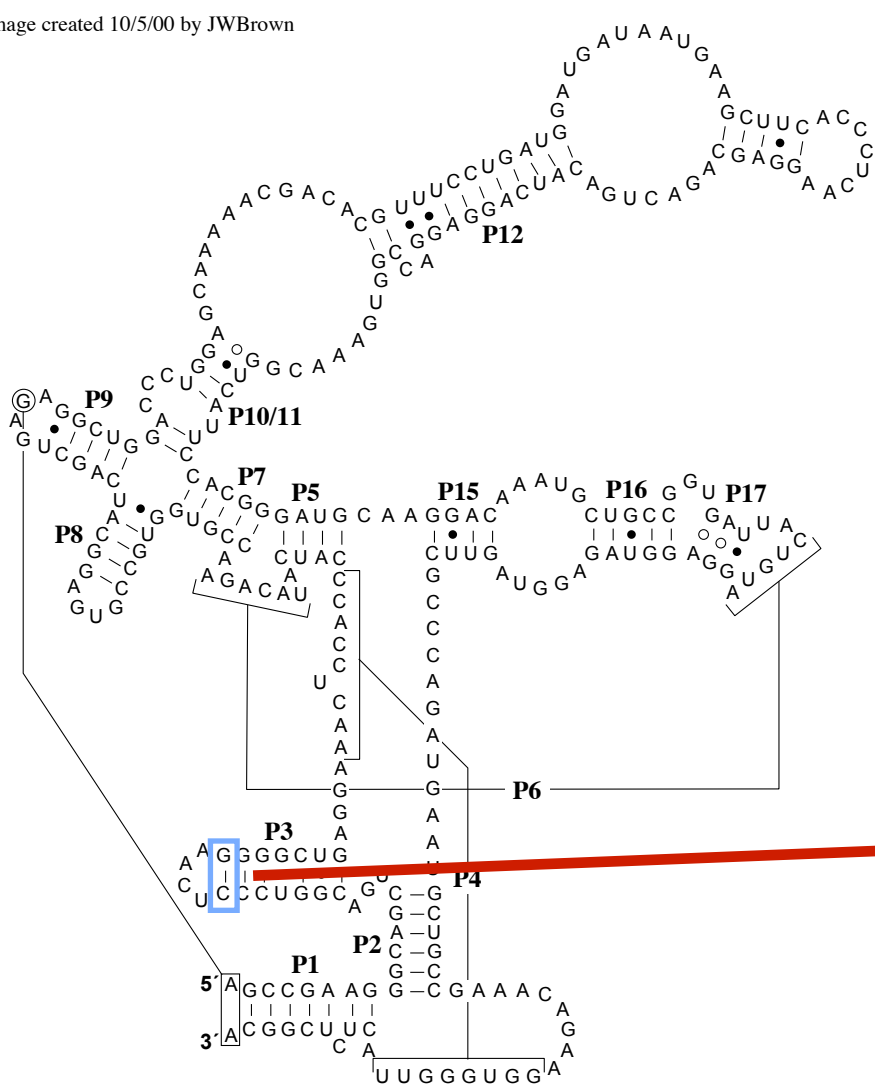
Problem : What about regions that are alignable between some sequences but not others?

Problems with alignments

Ribonuclease P RNA
Methanobacterium thermoautotrophicum ΔH

Sequence : U42986, Pannucci 1999 PNAS **96**:7803
 Structure : Harris, *et al.*, RNA (in press)

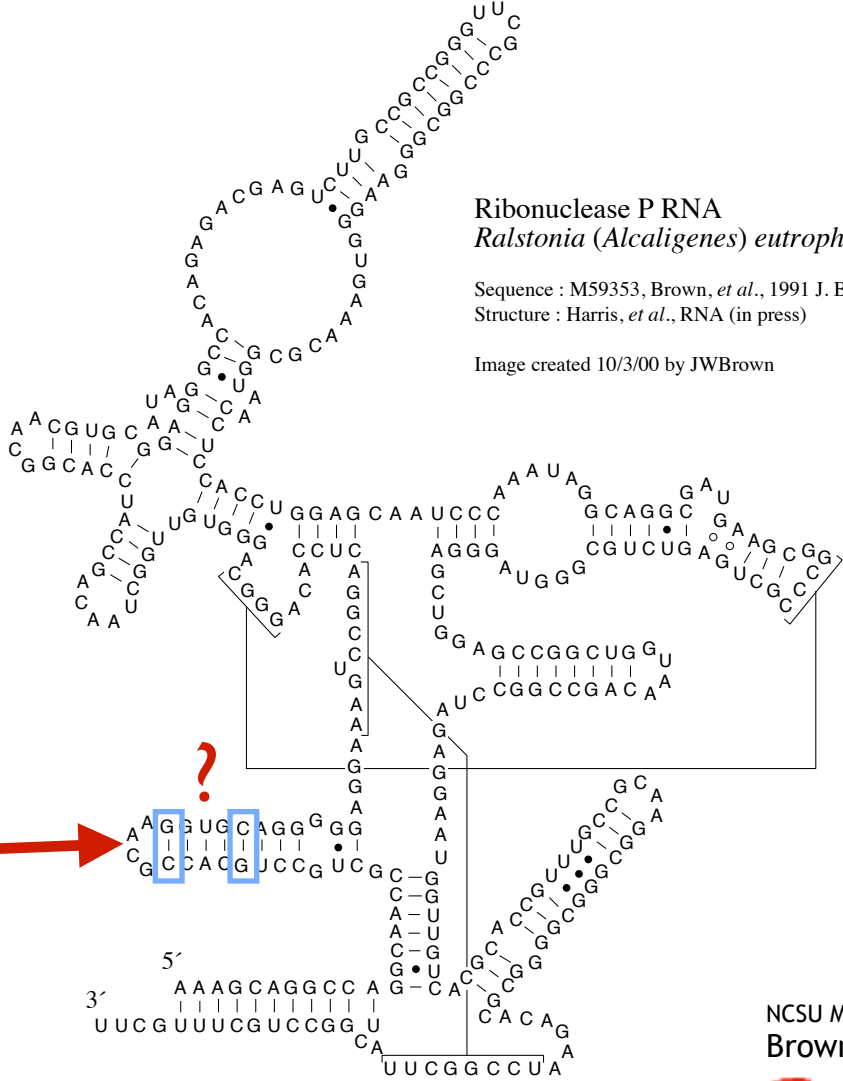
Image created 10/5/00 by JWBrown



Ribonuclease P RNA
Ralstonia (Alcaligenes) eutrophus DSM 531

Sequence : M59353, Brown, *et al.*, 1991 J. Bacteriol. **173**:3855
 Structure : Harris, *et al.*, RNA (in press)

Image created 10/3/00 by JWBrown



NCSU Microbiology
 Brown lab



Problems with alignments

Correspondence must be assigned nt-by-nt; columns are 1 nt wide by definition.

helices<-----PIIII-5'----->	.LIIII.<-----PIIII-3'----->....
pairing	----((((((((((((((((-----((((-----)))))))))---)))))	----((((((((((((((((-----)))))))))---)))))
pairing	----ABCDEFGHIJKLMN----OPQRS	----SRQPO----NMLKJIHG--FE-DCBA----
Ssolfat	UAA-CGGGG-----	CAAA-----C-CCUGAGGA
Sacidoc	UUA-CGGGA-----	AUA-----U-CCUGAGGA
Msedula	CCA-CGG-----	GAAA-----CUGGGGA
Apernix	CCA-CGGCCCCC-----	AGCCA-----GGG--GG-GCUGAGGA
Pfurios	UGC-CGGGC-----	UUUAU-----G-CCCGAGGA
Tlittor	CCU-CGGGU-----	AUUUG-----A-CCCGAGGA
Mthermo	UGA-CGGUCCC-----	UCAA-----G--GG-GCUGAGGA
MthMarb	UGA-CGGCCA-----	UUUU-----U--GG-GCUGAGGA
Mformic	UAC-CGGUUUCUAUAGAU-----	UUAAU-----GUCUGUAGUUAA-ACUGAGGA
Tvolcan	UGA-CGCC-----	GUAA-----GGUGAGGA
Mbarker	UGA-CGGGCC-----	UUCG-----GG-UCUGAGGA
Hcutiru	UGCCCGUGCC-----	GUGA-----GG-CAUGAGGA
Hvolcan	UCC-CGUGCCCG-----	AGA-----CG--GG-CAUGAGGA
Hmorrhhu	CAC-CGCGGCGUACC---GACAGGCAC	ACAC-GUGCCAGCG---GGUAC--GCACGCGAGGA
Ngregor	UGC-CGCGGGCGUC-----	GUGC-----GACG--CG-CGCGAGGA

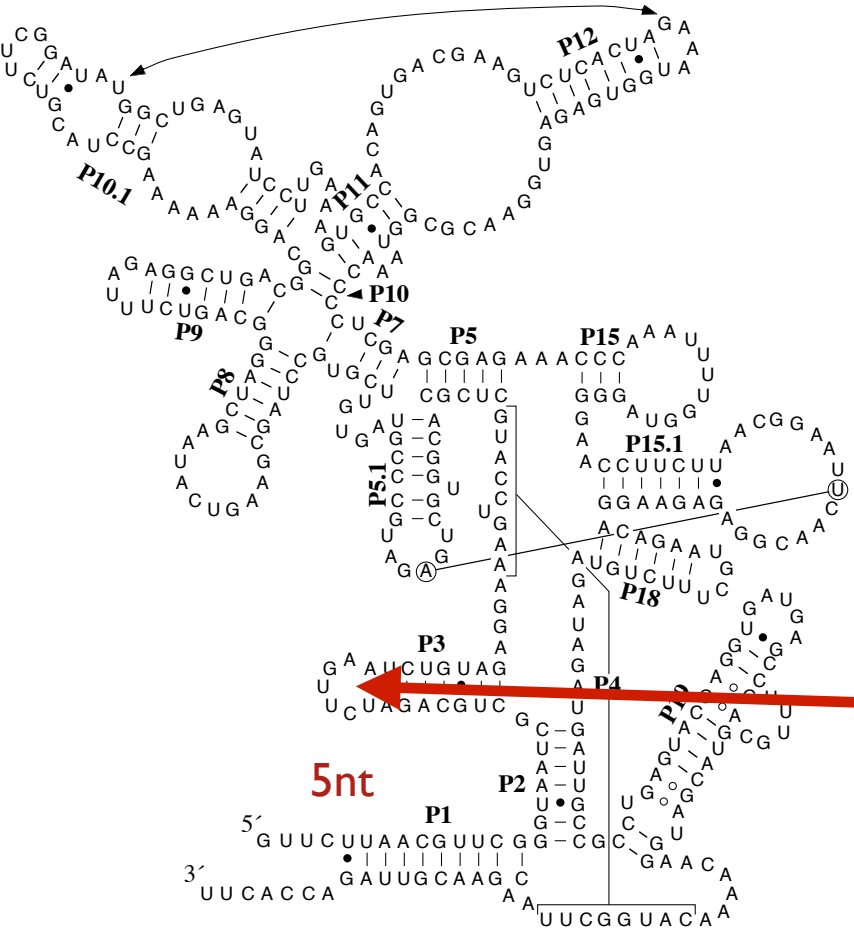
Problems :

- Alignments contain regions correspond *en masse* but not alignable nt-by-nt
- Sequences do not correspond along their entire length

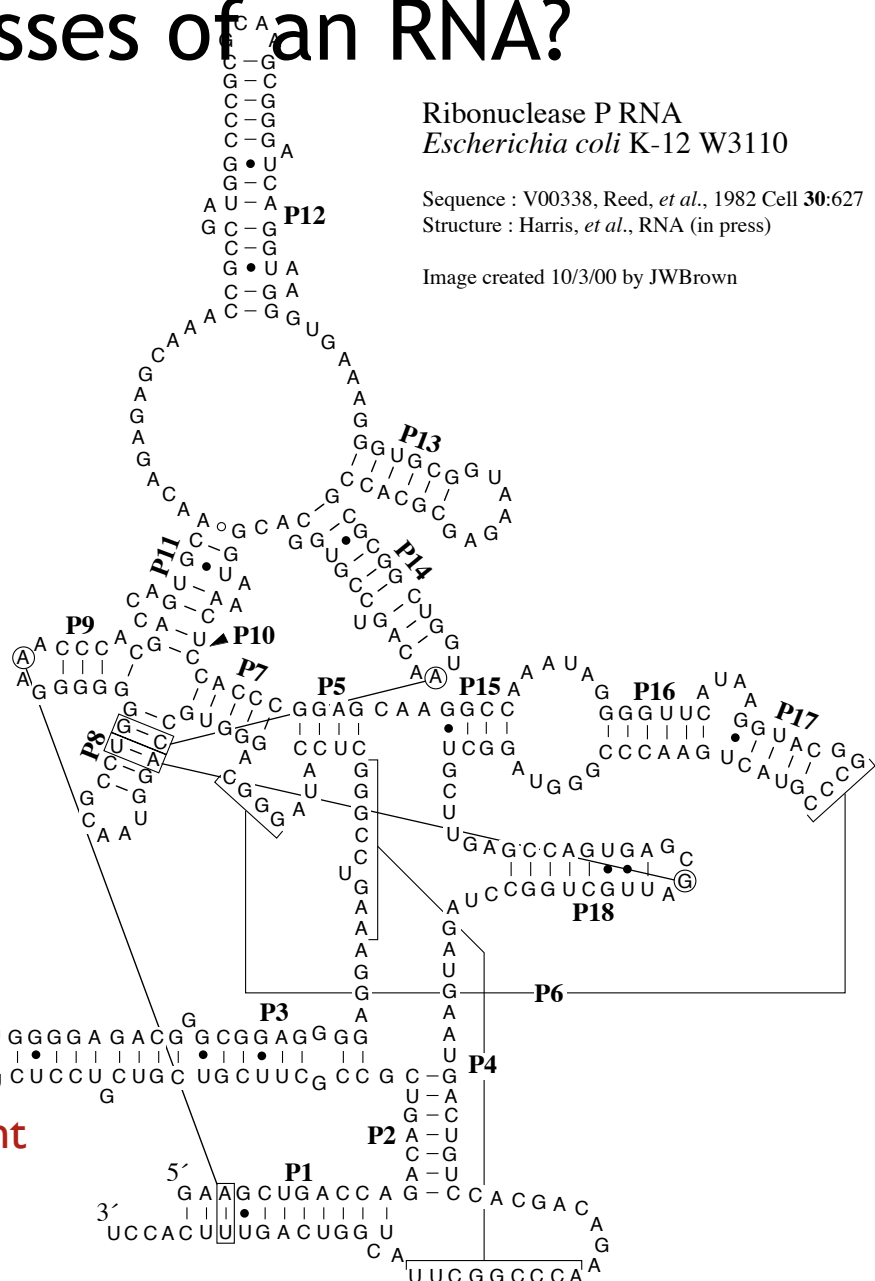
Problems with alignments

How do we align regions of non-corresponding structure in different classes of an RNA?

Ribonuclease P RNA
Bacillus subtilis 168
 Sequence : M13175, Reich, *et al.*, 1986 J. Biol. Chem. 261:7888
 Structure : Harris, *et al.*, RNA (in press)
 Image created 10/3/00 by JWBrown



Ribonuclease P RNA
Escherichia coli K-12 W3110
 Sequence : V00338, Reed, *et al.*, 1982 Cell 30:627
 Structure : Harris, *et al.*, RNA (in press)
 Image created 10/3/00 by JWBrown



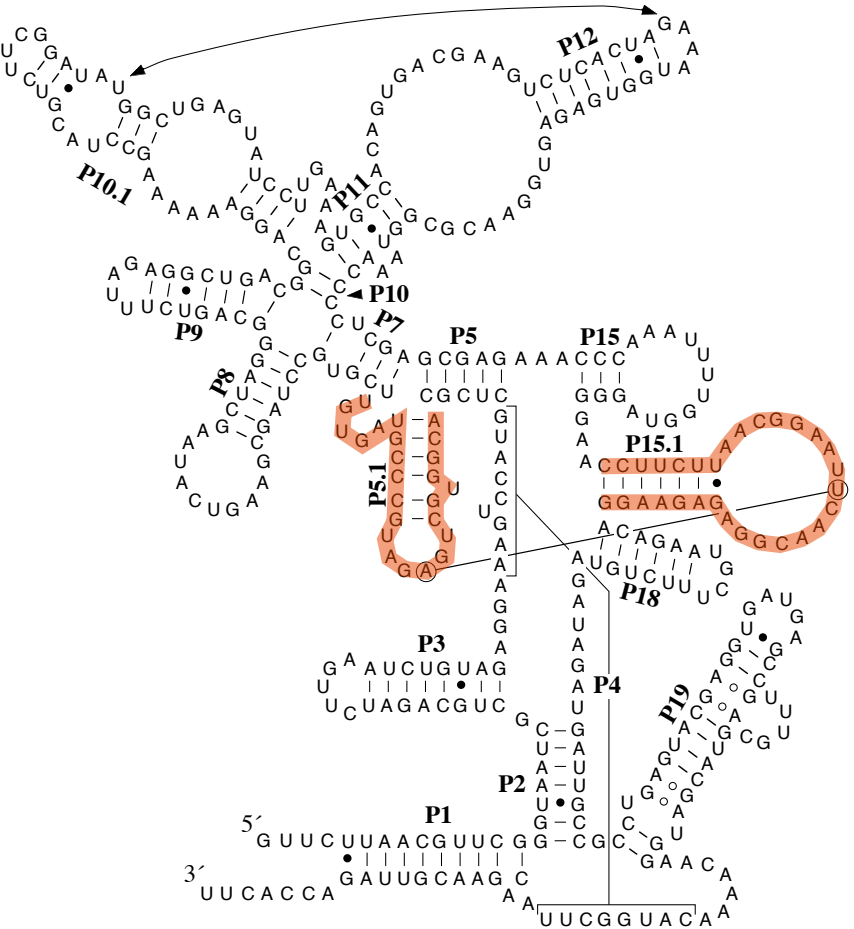
Problems with alignments

How do we align regions of non-corresponding structure in different classes of an RNA?

Ribonuclease P RNA
Bacillus subtilis 168

Sequence : M13175, Reich, *et al.*, 1986 J. Biol. Chem. 261:7888
Structure : Harris, *et al.*, RNA (in press)

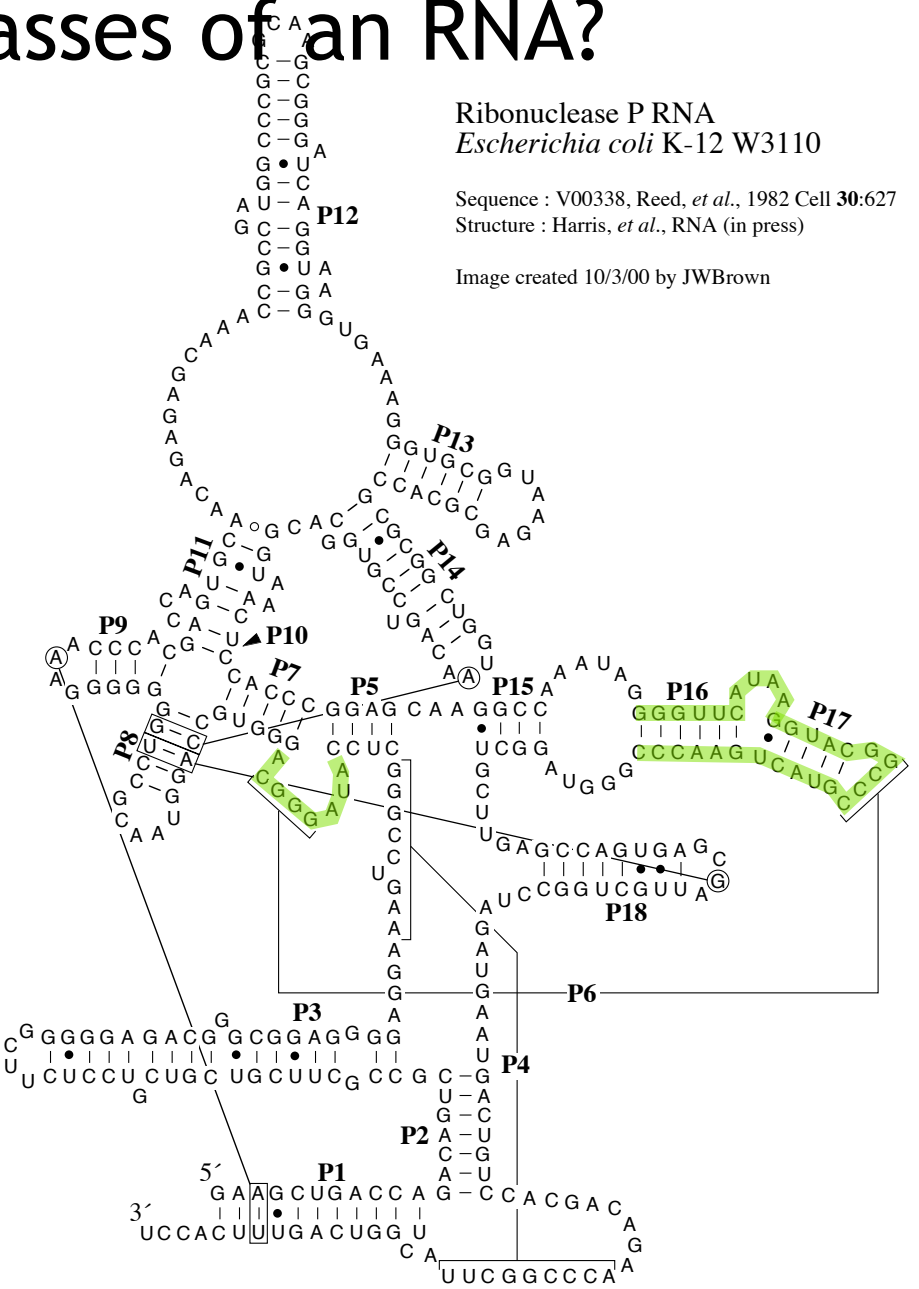
Image created 10/3/00 by JWBrown



Ribonuclease P RNA
Escherichia coli K-12 W3110

Sequence : V00338, Reed, *et al.*, 1982 Cell 30:627
Structure : Harris, *et al.*, RNA (in press)

Image created 10/3/00 by JWBrown



Problems with the 2D matrix paradigm

1. The 2D matrix paradigm forces alignment nt-by-nt even in regions where only alignment of “regions” is meaningful.
2. The 2D matrix forces alignment of non-corresponding regions between structure classes.
3. Alignments expand as large numbers of similar sequences and gaps accumulate, to the point that they are unmanageable.
4. Annotation of data within an alignment is problematic, and *ad hoc* solutions are highly constrained and usually result in data lost in translation.

Solution: Assign correspondence explicitly

Alignment is just the assignment of sets of “*corresponds_to*” relations

```
helices  . . . .<-----PIII-5' ----->.LIII.<-----PIII-3' ----->. . . .
pairing  ----((((((((((((((((-----((((-----))))))-----))))))----))--))--))----
pairing  ----ABCDEFGHIJKLMN----OPQRS-----SRQPO-----NMLKJIHG--FE-DCBA----
Ssolfat  UAA-CGGGG-----CAAA-----C-CCUGAGGA
Sacidoc  UUA-CGGGA-----AUA-----U-CCUGAGGA
Msedula  CCA-CGG-----GAAA-----CUGGGGA
Apernix  CCA-CGGCCCCC-----AGCCA-----GGG--GG-GCUGAGGA
Pfurios  UGC-CGGGC-----UUUAU-----G-CCCGAGGA
Tlittor  CCU-CGGGU-----AUUUG-----A-CCCGAGGA
Mthermo  UGA-CGGUCCC-----UCAAA-----G--GG-GCUGAGGA
MthMarb  UGA-CGGCCA-----UUUU-----U--GG-GCUGAGGA
Mformic  UAC-CGGUUUCUAUAGAU-----UUAUU-----GUCUGUAGUUAA-ACUGAGGA
Tvolcan  UGA-CGCC-----GUAA-----GGUGAGGA
Mbarker  UGA-CGGGCC-----UUCG-----GG-UCUGAGGA
Hcutiru  UGCCCUGGCC-----GUGA-----GG-CAUGAGGA
Hvolcan  UCC-CGUGCCCG-----AGA-----CG--GG-CAUGAGGA
Hmorrhru CAC-CGCGGCGUACC---GACAGGCAC-ACAC-GUGCCAGCG---GGUAC--GCACGCGAGGA
Ngregor  UGC-CGCGGGCGUC-----GUGC-----GACG--CG-CGCGAGGA
```

Traditional alignments implicitly assign correspondence (“homology”) to all columns, and all such assignments are nt-by-nt

Solution: Assign correspondence explicitly

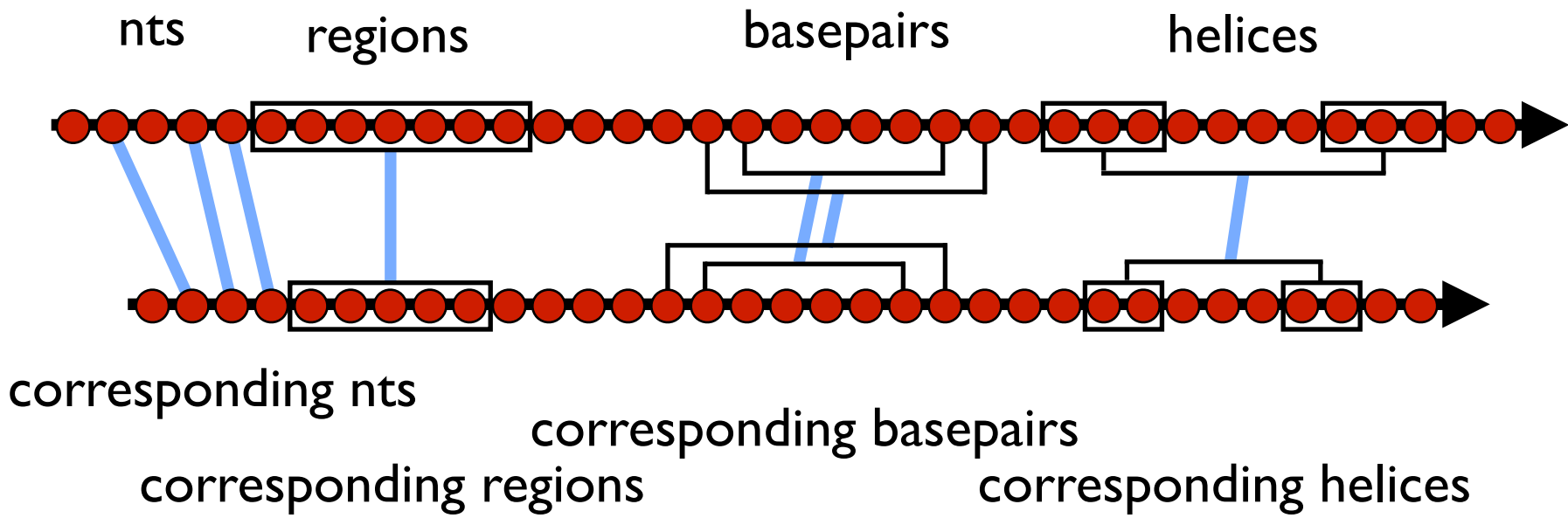
Alignment is just the assignment of sets of “*corresponds_to*” relations

```

helices  ....<-----PIII-5' ----->.LIII.<-----PIII-3' ----->....
pairing  ----((((((((((((((((-----((((-----))))))-----))))))----))--))----
pairing  ----ABCDEFGHIJKLMN----OPQRS-----SRQPO-----NMLKJIHG--FE-DCBA----
Solfat   UAA  CGGGG  CAAA  C  CCUGAGGA
Sacidoc  UUA  CGGGA  AUA   U  CCUGAGGA
Msedula  CCA  CGG    GAAA  CUGGGGA
Apernix  CCA  CGGCCCCC  AGCCA  GGG  GG  GCUGAGGA
Pfurios  UGC  CGGGC  UUUAU  G  CCCGAGGA
Tlittor  CCU  CGGGU  AUUUG  A  CCCGAGGA
Mthermo  UGA  CGGUCCC  UCAA  G  GG  GCUGAGGA
MthMarb  UGA  CGGCCCA  UUUU  U  GG  GCUGAGGA
Mformic  UAC  CGGUUUCUAUAGAU  UUAAU  GUCUGUAGUUAA  ACUGAGGA
Tvolcan  UGA  CGCC    GUAA  GGUGAGGA
Mbarker  UGA  CGGGCC  UUCG  GG  UCUGAGGA
Hcutiru  UGCG  CGUGCC  GUGA  GG  CAUGAGGA
Hvolcan  UCC  CGUGCCCG  AGA  CG  GG  CAUGAGGA
Hmorrh  CAC  CGCGGCGUACC  GACAGGCAC  ACAC  GUGCCAGCG  GGUAC  GCACGCGAGGA
Ngregor  UGC  CGCGGGCGUC  GUGC  GACG  CG  CGCGAGGA
    
```

These homology relations should be assigned explicitly and specifically rather than implicitly and indiscriminantly

Corresponding elements in RNA



Corresponding elements in RNA

Ribonuclease P RNA
Bacillus subtilis 168

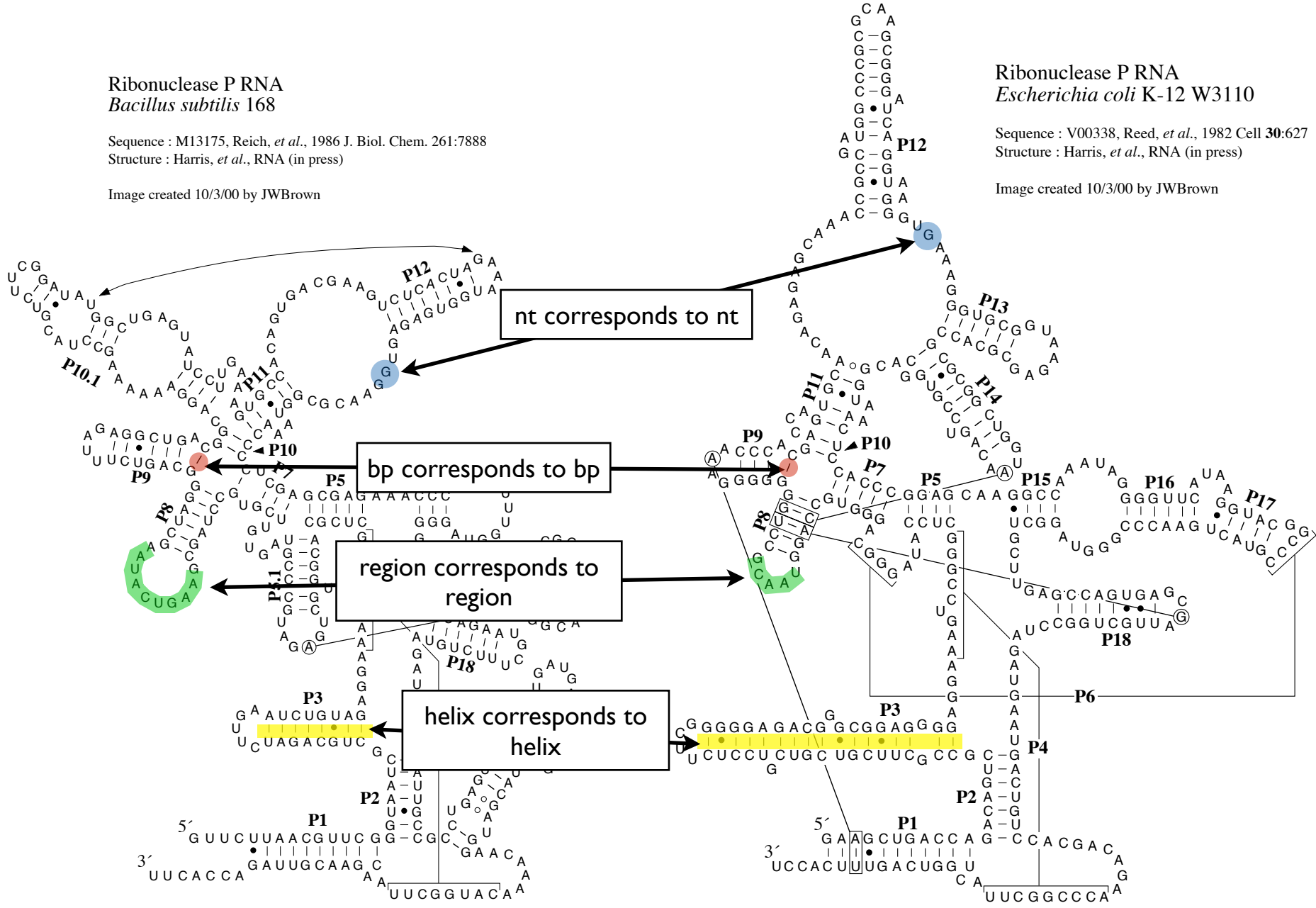
Sequence : M13175, Reich, *et al.*, 1986 J. Biol. Chem. 261:7888
Structure : Harris, *et al.*, RNA (in press)

Image created 10/3/00 by JWBrown

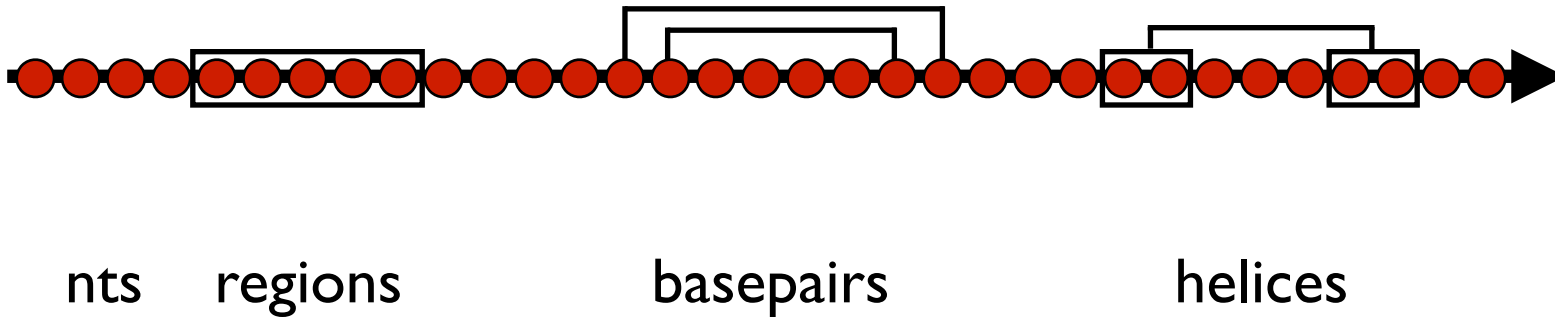
Ribonuclease P RNA
Escherichia coli K-12 W3110

Sequence : V00338, Reed, *et al.*, 1982 Cell 30:627
Structure : Harris, *et al.*, RNA (in press)

Image created 10/3/00 by JWBrown



Corresponding elements in RNA



This is the rudimentary RNA 2° structure ontology

What is needed in an alignment editor?

An ontology-based approach to alignment

The traditional approach of manipulating gaps to move nts into columns (in a 2D matrix) that *imply* correspondence is fundamentally backwards.

The user (machine or human) should assign correspondence, and then the *editor* should display this appropriately, perhaps (but not necessarily) by putting corresponding elements into columns of a 2D or multidimensional matrix.

A man with a mustache, wearing a white lab coat and a white head covering, is shown from the chest up. He has a concerned expression and is holding his right hand to his forehead. A white speech bubble with a black outline is positioned to the left of his head, containing the text "My brain hurts too!". The background is a dark, out-of-focus wall.

My brain hurts **too!**