The RNA Ontology (RNAO): An Ontology for Integrating RNA Sequence and Structure Data

Neocles Leontis

RNA Ontology Consortium

Bowling Green State University

ROC Participants/Contributors

Craig Zirbel, Eric Westhof, Jesse Stombaugh (Ph.D. 2009)

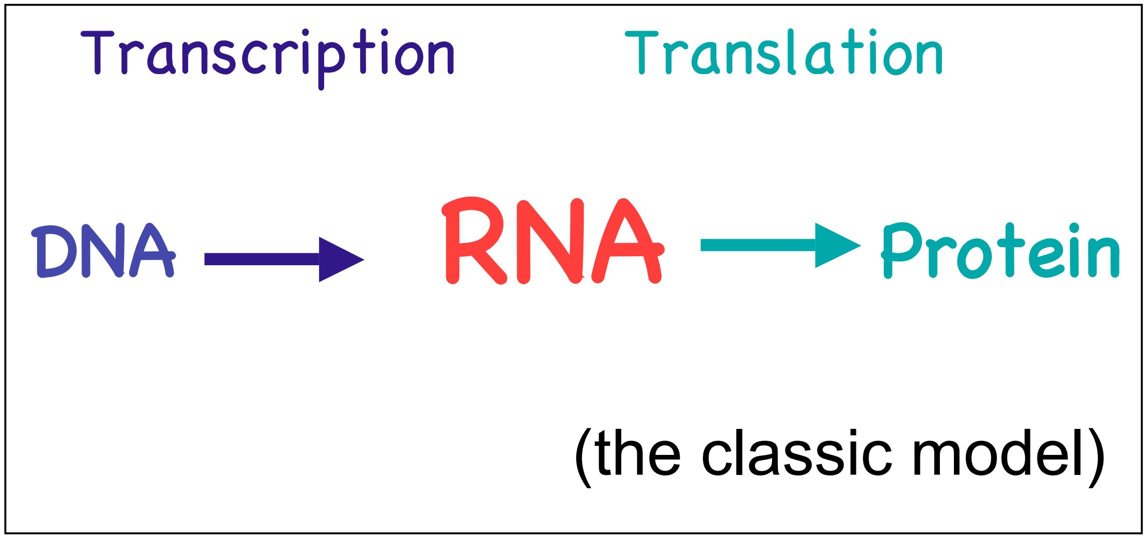
Colin Batchelor, Helen Berman (PDB/NDB), Thomas Bittner, James Brown, Karen Eilbeck (SO), Janna Hastings (ChEBI), Robert Hoendorf, Rob Knight, Franz Lang, Alain Laederach, Christopher Mungall (GO), Jane Richardson, Gerhard Steger, John Westbrook (partial list)

Funding: NSF RCN, NIH GMS

Outline

- Motivation for Developing RNAO
- The RNA Ontology Consortium (ROC)
- Relation to other Ontologies
- RNA 3D structure - Entities and Relations
- Formalization - definitions and axioms starting from a minimal set of primitives
- Annotating RNA Sequence Alignments

Motivation: RNA and the Genome


Transcription

Translation

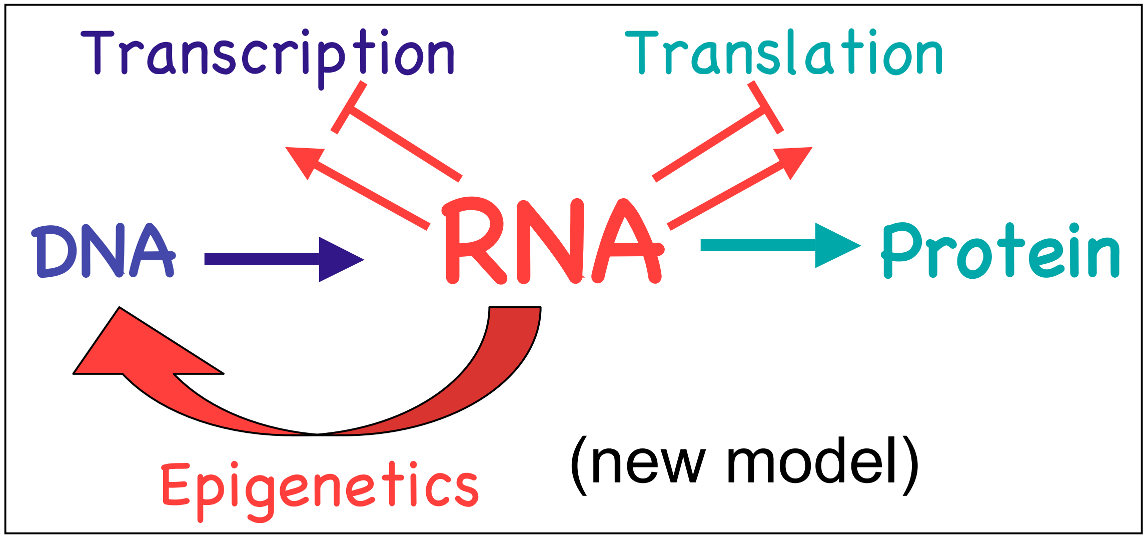
DNA → RNA → Protein

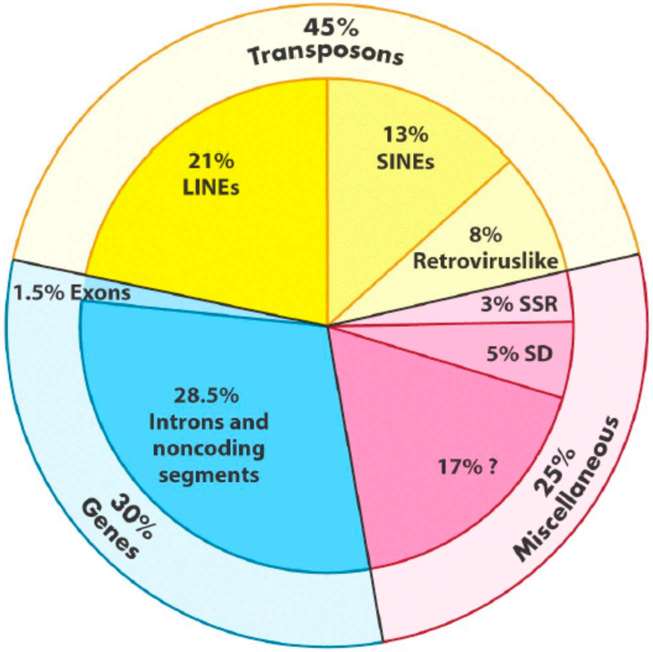
(the classic model)

Only 1.5% of Genome Codes for Protein
What about the rest?

- **Most of eukaryotic genome is transcribed to RNA & most is probably functional**
- Significant fraction of functional RNAs are structured (function depends on 2D or 3D structure, not just sequence)
- Larger amounts of diverse and heterogeneous RNA data are rapidly accumulating:
  - 3D Structures
  - Homologous Sequences
  - Functional data

Motivation: RNA and the Genome

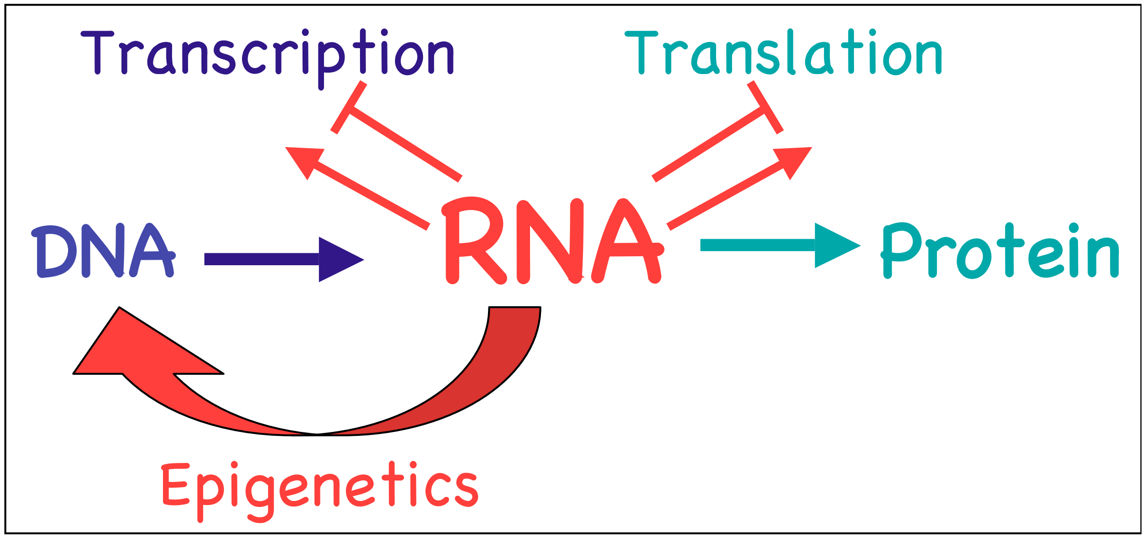


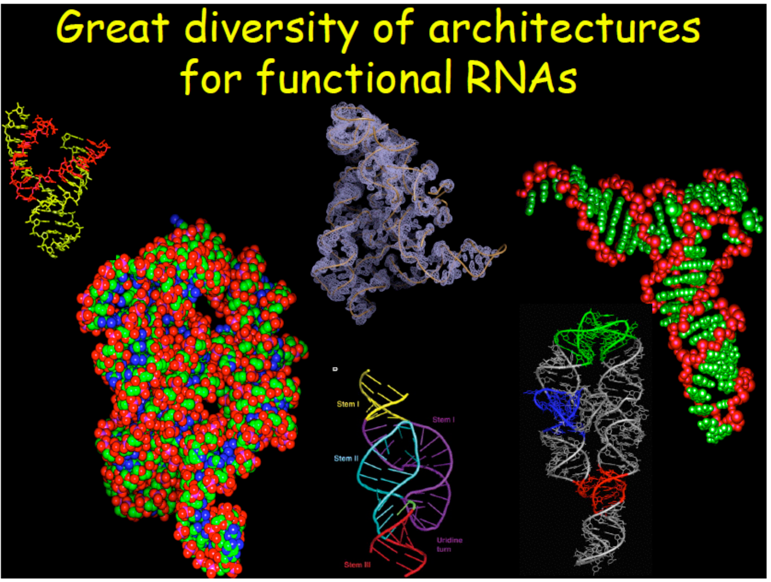


- **Most of eukaryotic genome is transcribed to RNA & most is probably functional**
- Significant fraction of functional RNAs are structured (function depends on 2D or 3D structure, not just sequence)
- Larger amounts of diverse and heterogeneous RNA data are rapidly accumulating:
  - 3D Structures
  - Homologous Sequences
  - Functional data

RNA
Ontology
Consortium

Motivation: RNA and the Genome




Great diversity of architectures for functional RNAs

- Most of eukaryotic genome is transcribed to RNA & most is probably functional
- **Significant fraction of functional RNAs are structured (function depends on 2D or 3D structure, not just sequence)**
- Larger amounts of diverse and heterogeneous RNA data are rapidly accumulating:
  - 3D Structures
  - Homologous Sequences
  - Functional data

RNA Ontology (RNAO) Consortium

ROC was established in 2005 under Auspices of RNA Society

NSF Funding obtained to support Collaborative Activities

ROC is an International Consortium -- open to all interested persons

See: http://roc.bgsu.edu

Current version of RNAO can be downloaded and viewed with Protégé:

See: http://code.google.com/p/rnao/

ROC Works with curators of related Open Biomedical Ontologies (OBO) to ensure:

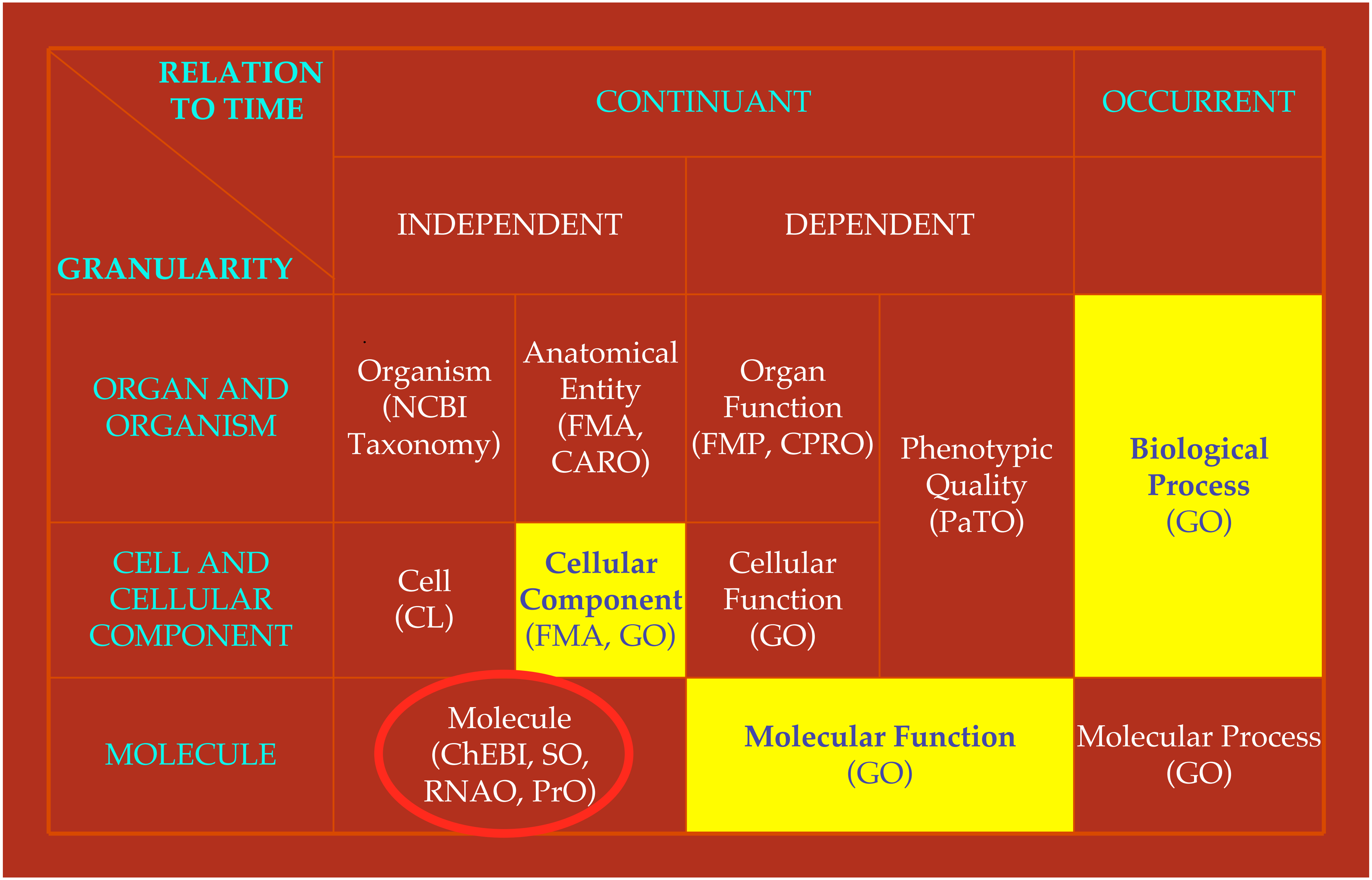
-Orthogonality

-Interoperability

ROC invites your participation!

Objectives of RNAO

- To integrate diverse and heterogeneous data regarding RNA molecules especially 3D and sequence data
- To Describe the 3D Structures of Individual RNA Molecules -their parts and the relations between the parts -- in computer and human understandable formats
- To define relations between "corresponding" parts of homologous RNA molecules
- To use these relations to annotate genomic sequences and RNA sequence alignments
- To draw new inferences regarding RNA structure, evolution and function from the integrated data



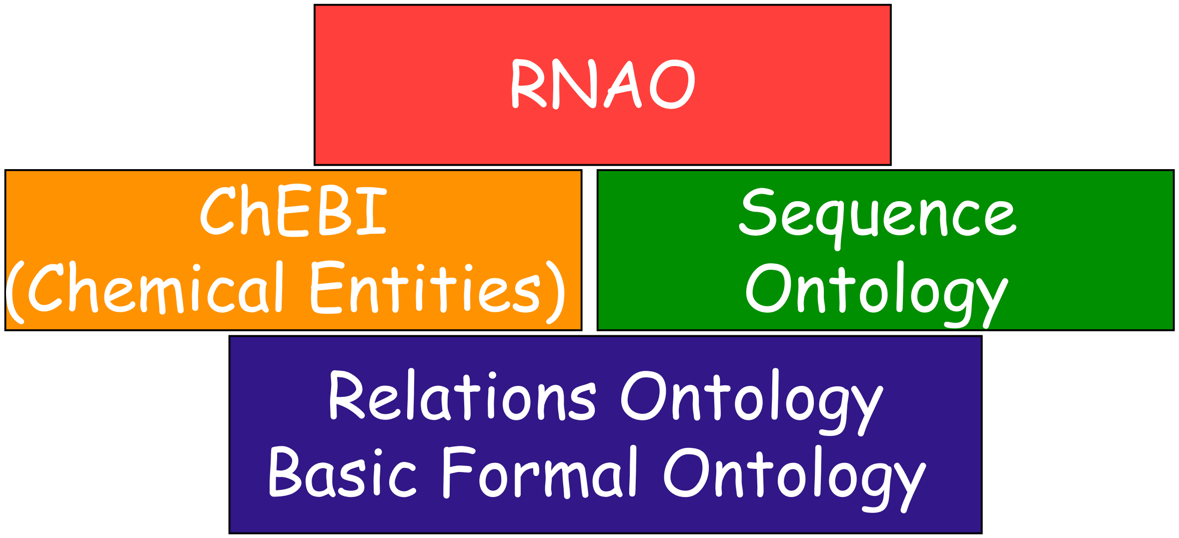
RELATION TO TIME	CONTINUANT				OCCURRENT
GRANULARITY	INDEPENDENT		DEPENDENT		
ORGAN AND ORGANISM	Organism (NCBI Taxonomy)	Anatomical Entity (FMA, CARO)	Organ Function (FMP, CPRO)	Phenotypic Quality (PaTO)	**Biological Process** (GO)
CELL AND CELLULAR COMPONENT	Cell (CL)	**Cellular Component** (FMA, GO)	Cellular Function (GO)		
MOLECULE	Molecule (ChEBI, SO, RNAO, PrO)		**Molecular Function** (GO)		Molecular Process (GO)

OBO Foundry http://obofoundry.org

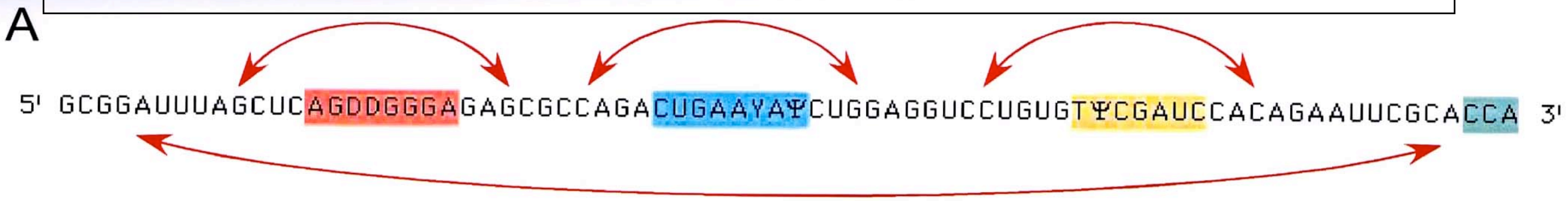
Smith et al. Nature Biotechnology 2008

Related ontologies in this space

- The Sequence Ontology (SO)
- ChEBI (Chemical Entities of Biological Interest)
- Gene Ontology (GO)



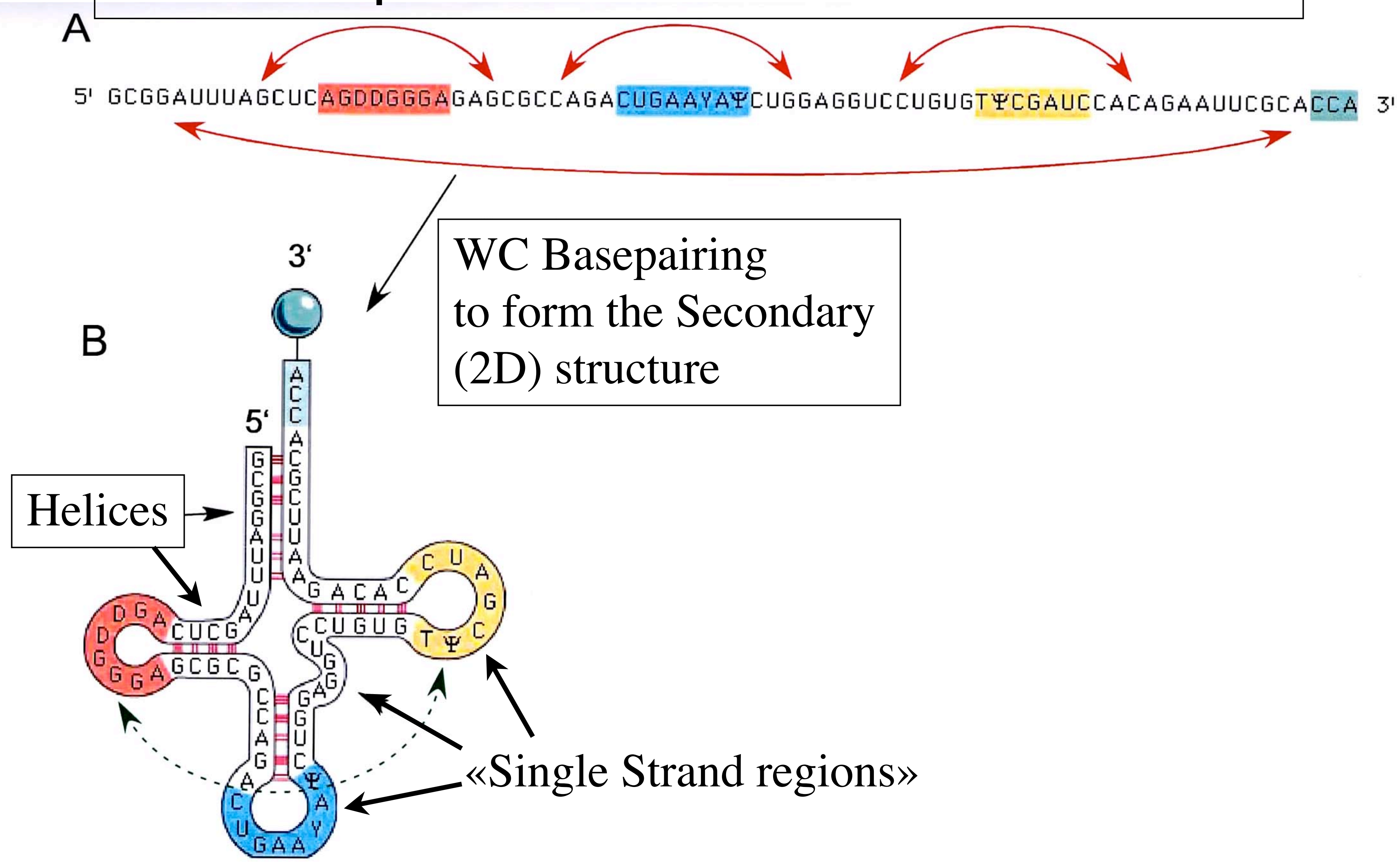
RNA sequence ---> 2D ---> 3D Structure


A
5' GCGGAUUUAGCUCAGDDGGGAGAGCGCCAGACUGAAYAΨCUGGAGGUCCUGUGTΨCGAUCCACAGAAUUCGCACCA 3'

RNA molecules are linear chains that fold back on themselves through diverse inter-nucleotide interactions to form unique 3D structures:

- •Base-pairing Interactions
- •Base-stacking Interactions
- •Base-backbone (phosphate) Interactions

RNA sequence ---> 2D ---> 3D Structure

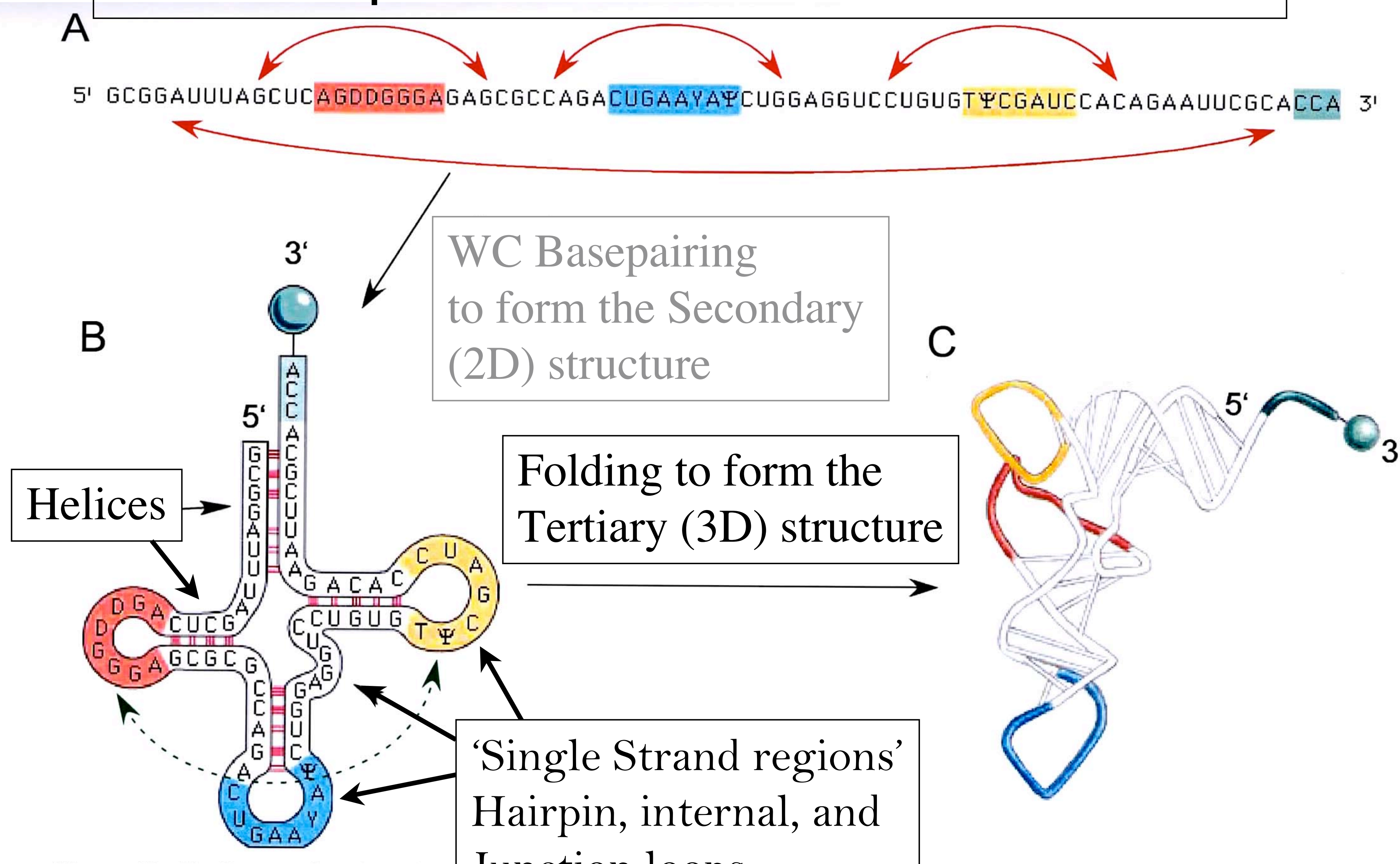

A
5' GCGGAUUUAGCUCAGDDGGGAGAGCGCCAGACUGAAYAΨCUGGAGGUCCUGUGTΨCGAUCCACAGAAUUCGCACCA 3'
B
3'
5'

WC Basepairing to form the Secondary (2D) structure

Helices

«Single Strand regions»

RNA sequence ---> 2D ---> 3D Structure


A
5' GCGGAUUUAGCUCAGDDGGGAGAGCGCCAGACUGAAYAΨCUGGAGGUCCUGUGTΨCGAUCCACAGAAUUCGCACCA 3'
B
3'
5'
C
5'
3'

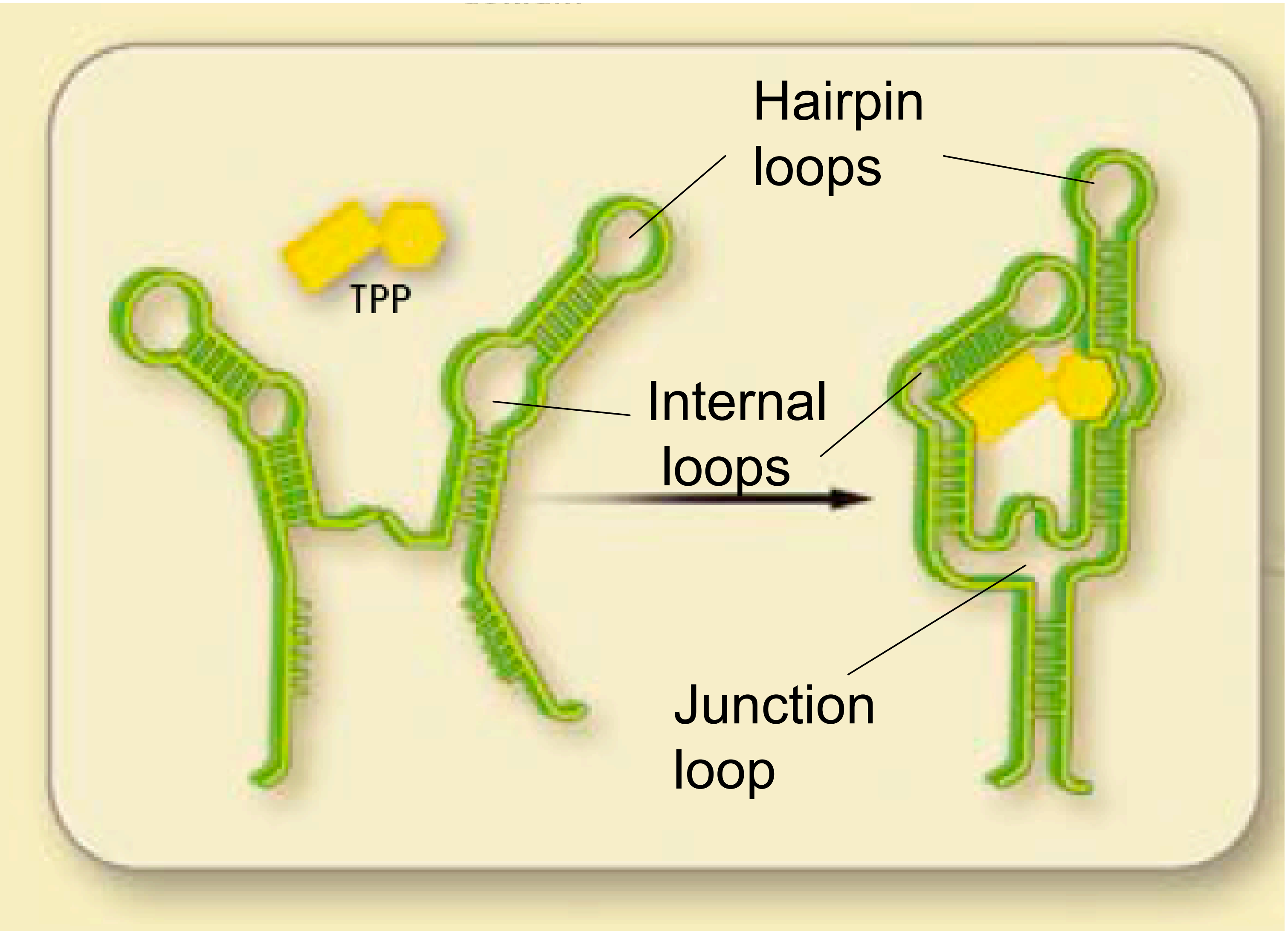
WC Basepairing to form the Secondary (2D) structure

Helices

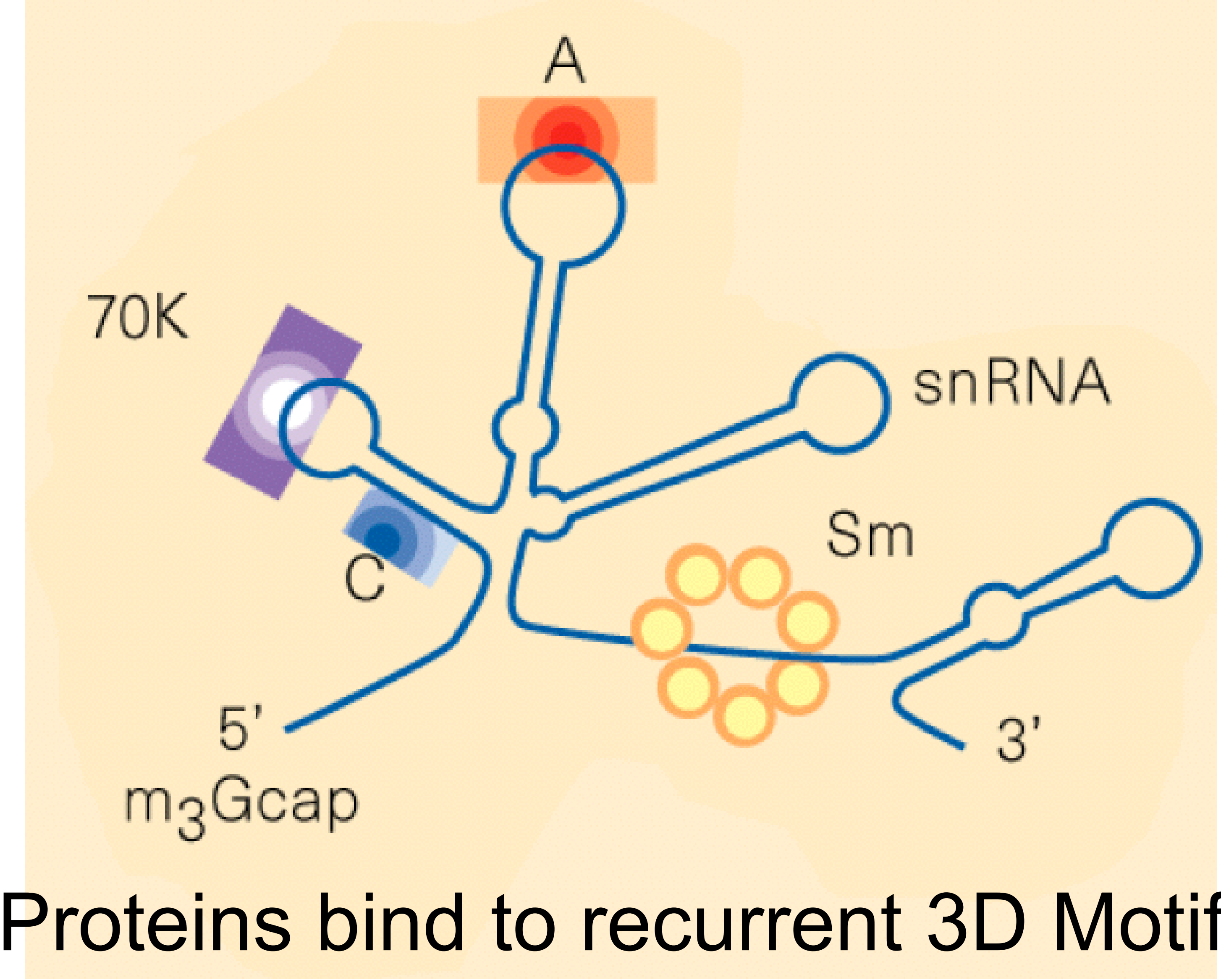
Folding to form the Tertiary (3D) structure

'Single Strand regions' Hairpin, internal, and Junction loops

RNA Folding forms “Stems” and “Loops” = “3D Motifs” = Binding and Docking Sites

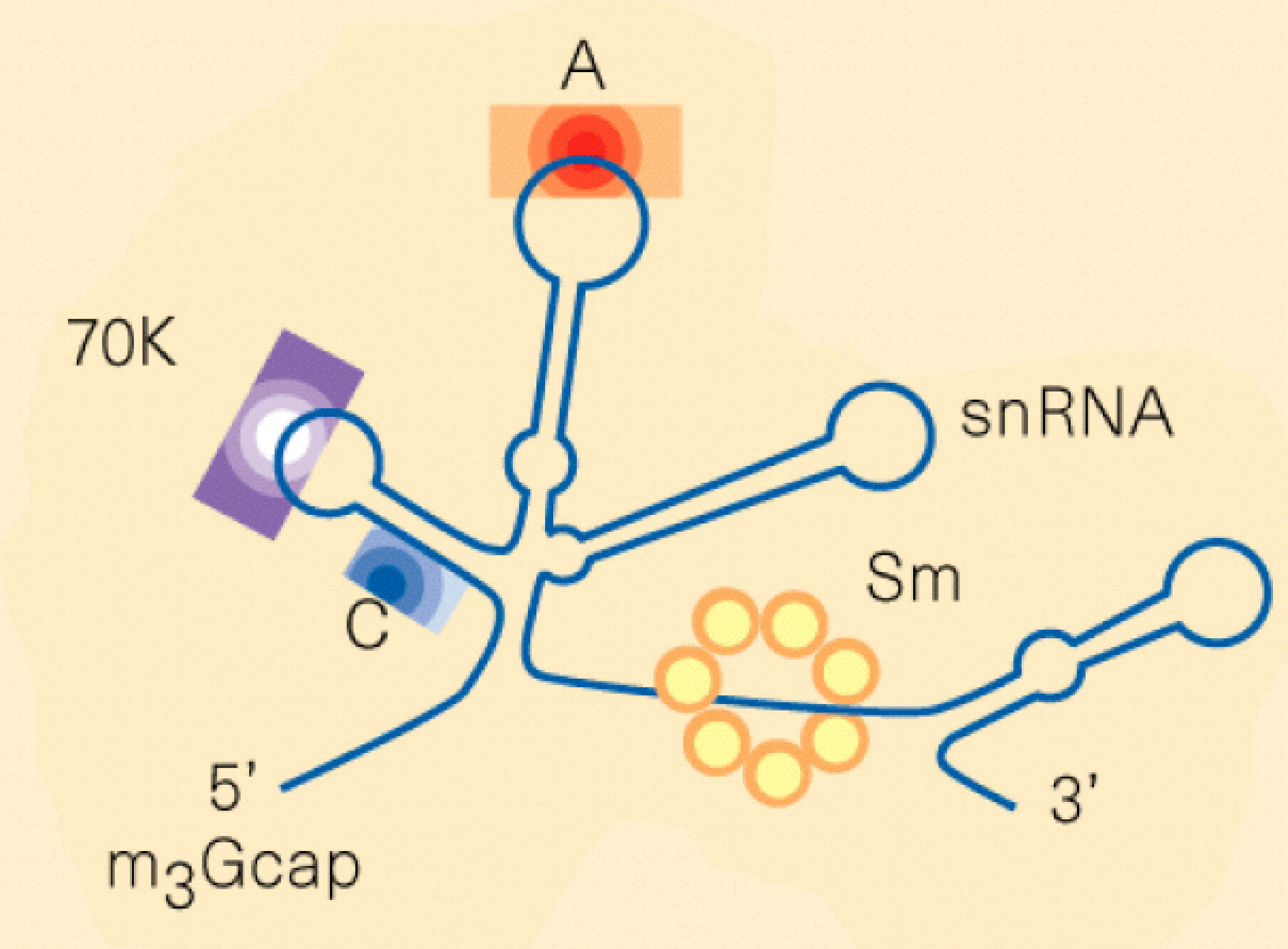


RNAs recruit multiple protein co-factors:



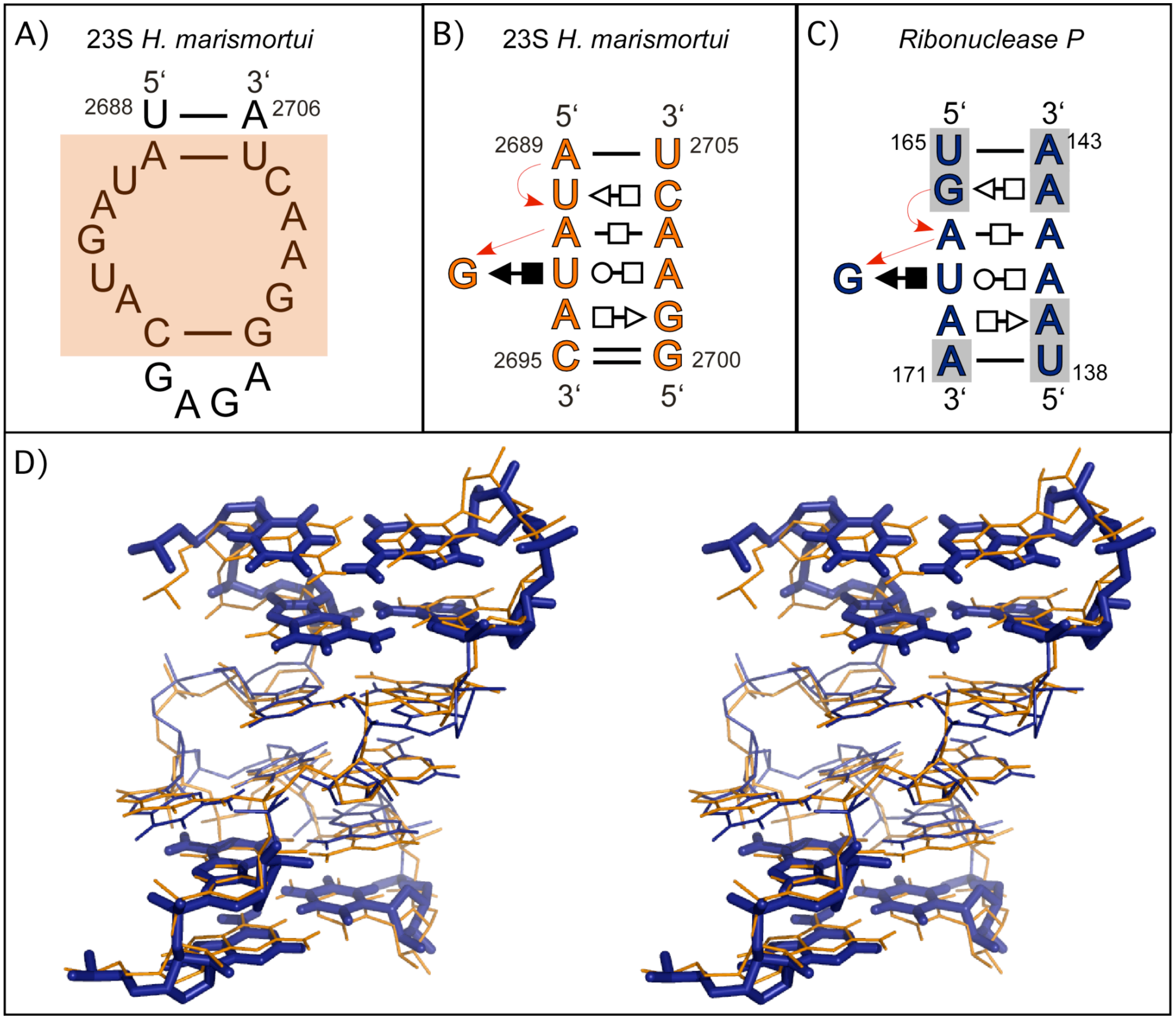
Proteins bind to recurrent 3D Motifs

There is no NAKED RNA in the Cell!

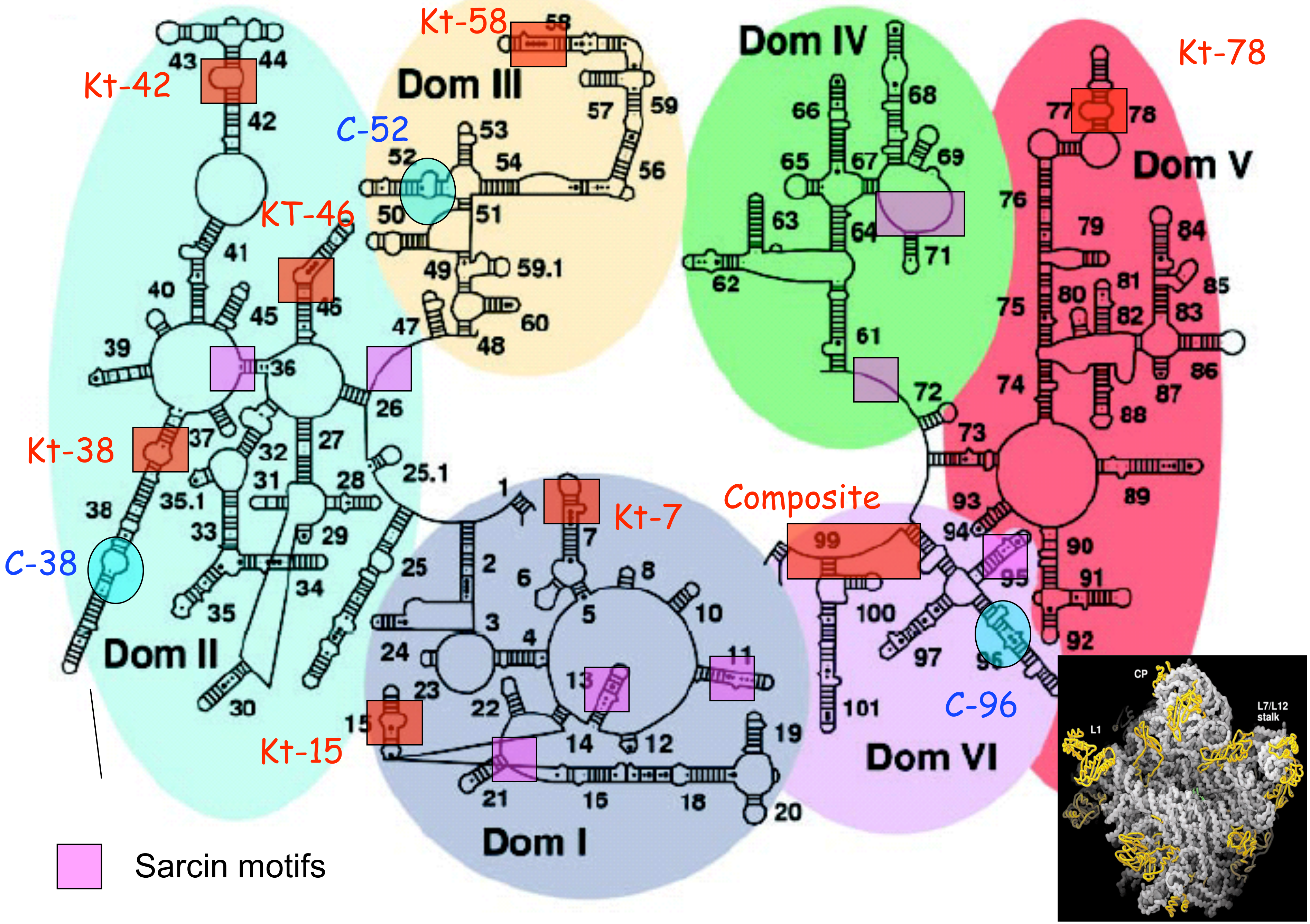


Proteins bind to recurrent 3D Motifs

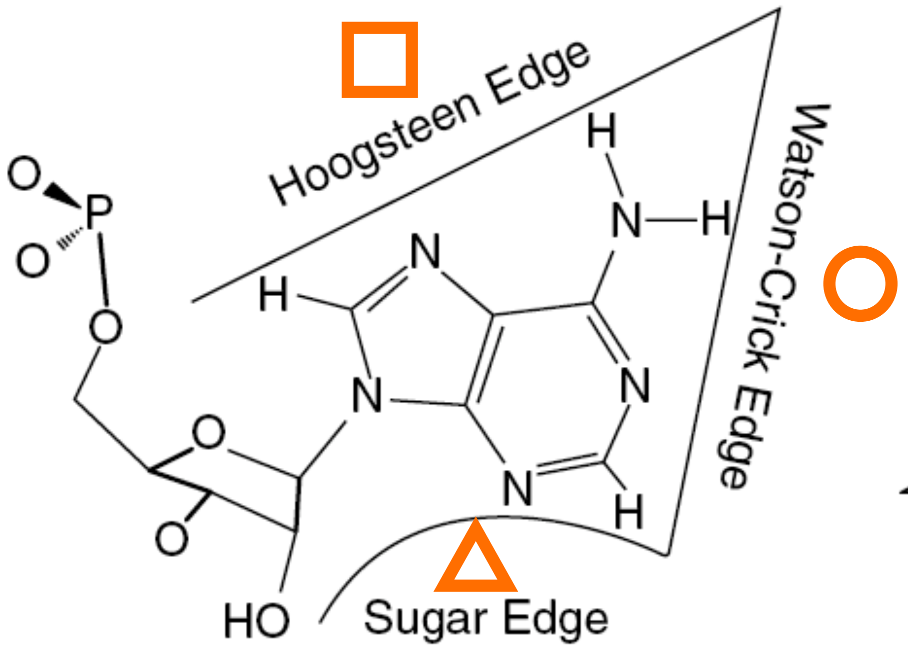
Example of a Recurrent Modular 3D Motif: Sarcin/Ricin "loop"

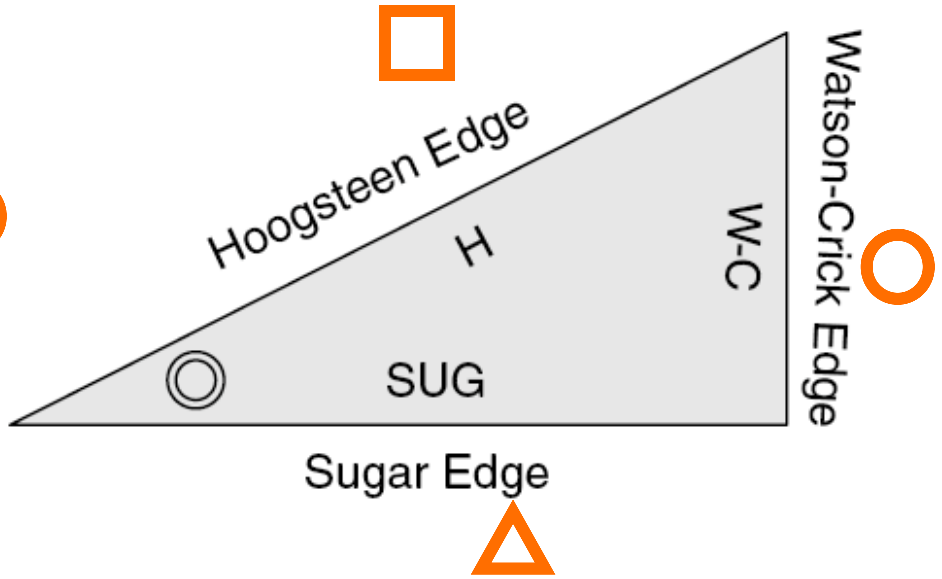


Recurrent Kink-turn, C-loops, & SR Motifs in 23S rRNA



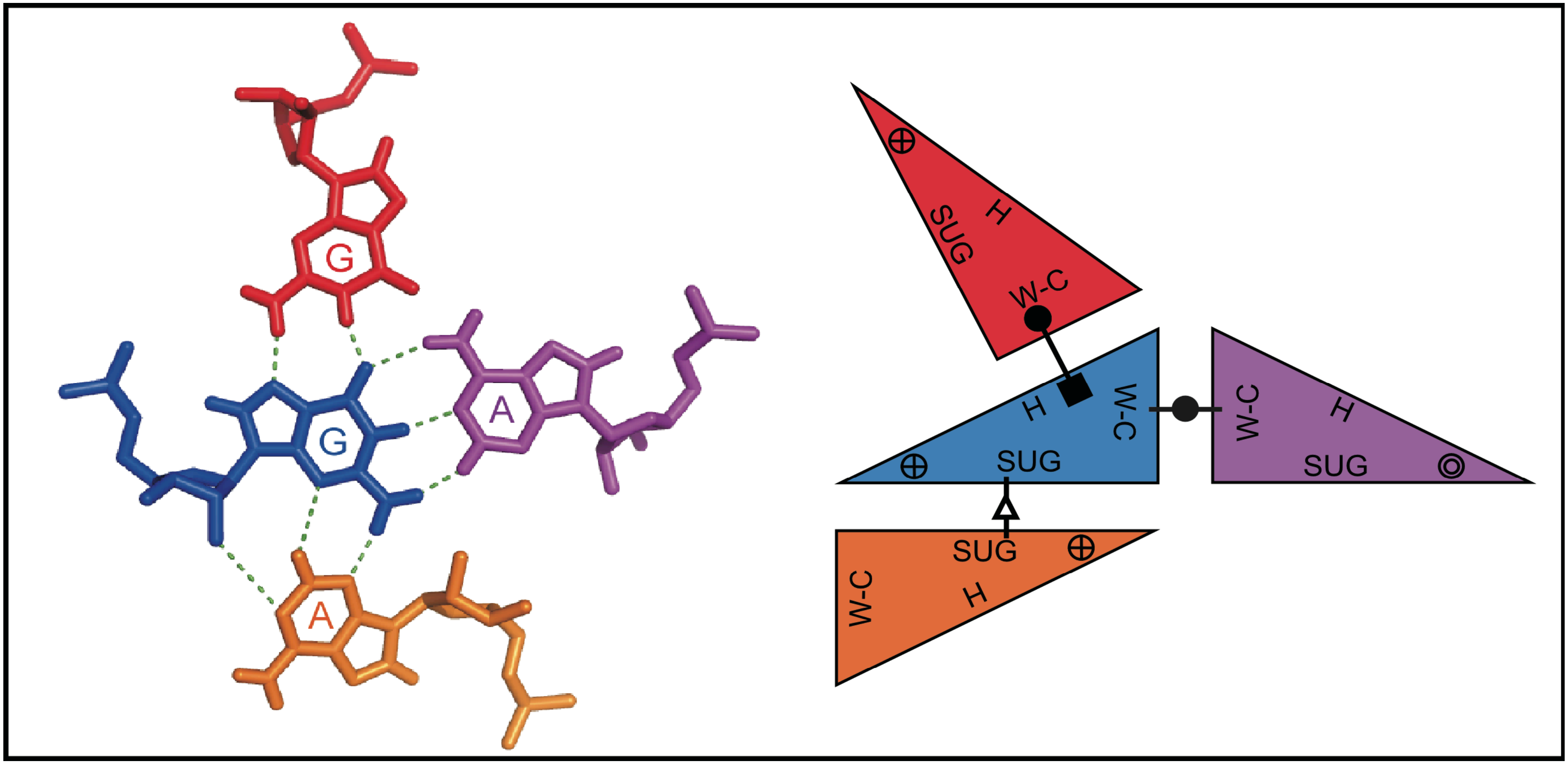
Classification of Base-pairs
(All Edge-to-edge Base Interactions)





Leontis & Westhof, RNA, 2001

Bases as Triangles

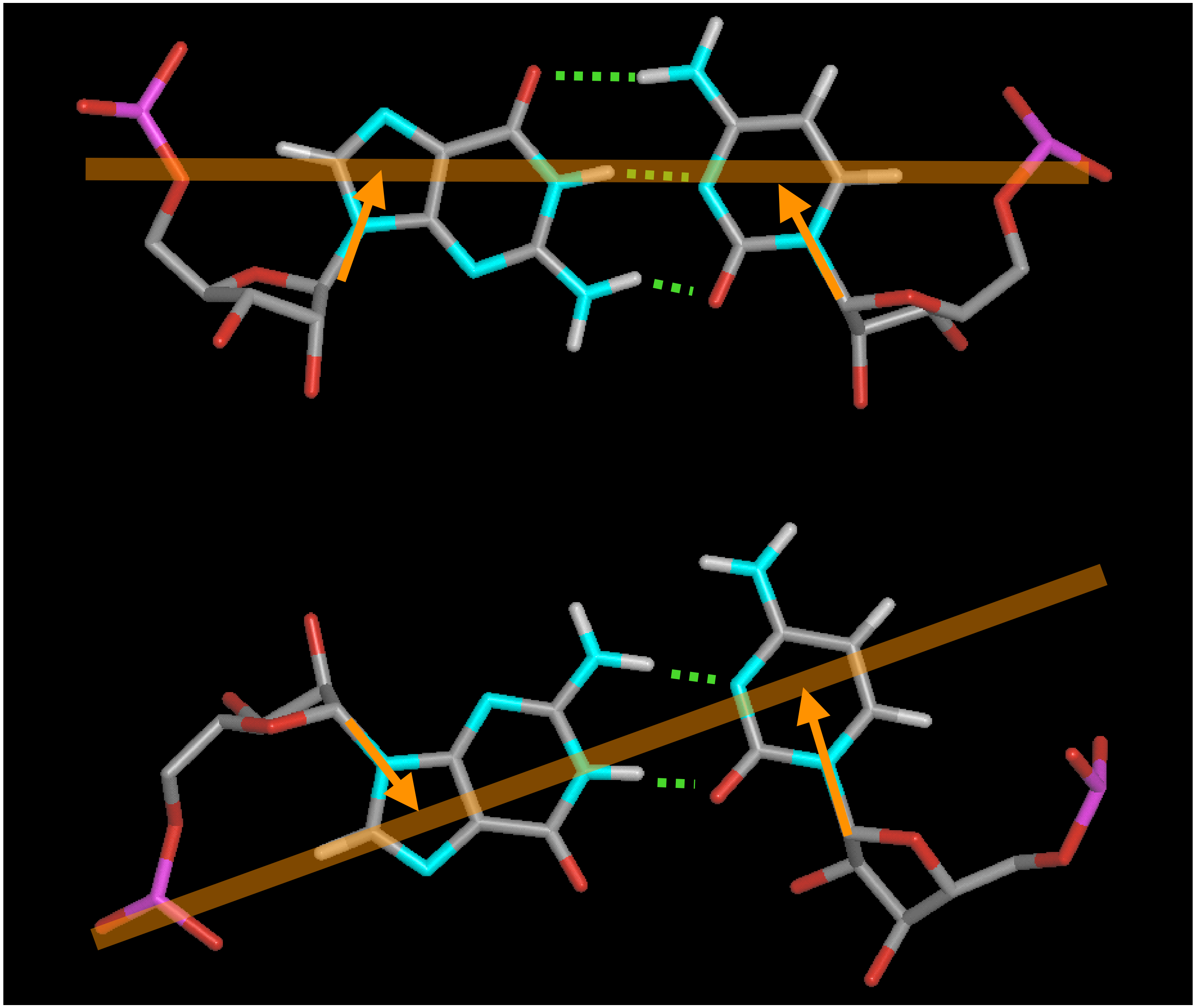


G68/A101/G64/A52 in T. th. 16S: 1j53.pdb

Glycosidic Bond Orientation

- Cis

- Trans



Edge-to-Edge Pairing Types

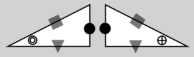
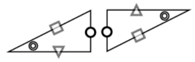
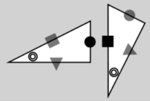
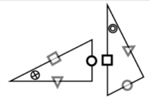
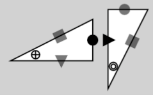
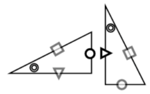
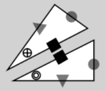
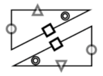
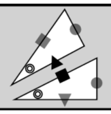
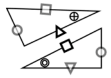
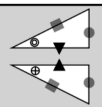
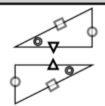
$$\left.\begin{array}{c}\text{Watson-Crick}\\ \text{Hoogsteen}\\ \text{Shallow-Groove}\end{array}\right\}\left\{\begin{array}{c}\text{Watson-Crick}\\ \text{Hoogsteen}\\ \text{Shallow-Groove}\end{array}\right\}\left\{\begin{array}{l}\text{Cis}\\ \text{Trans}\end{array}\right.$$

= 12 Basic Types

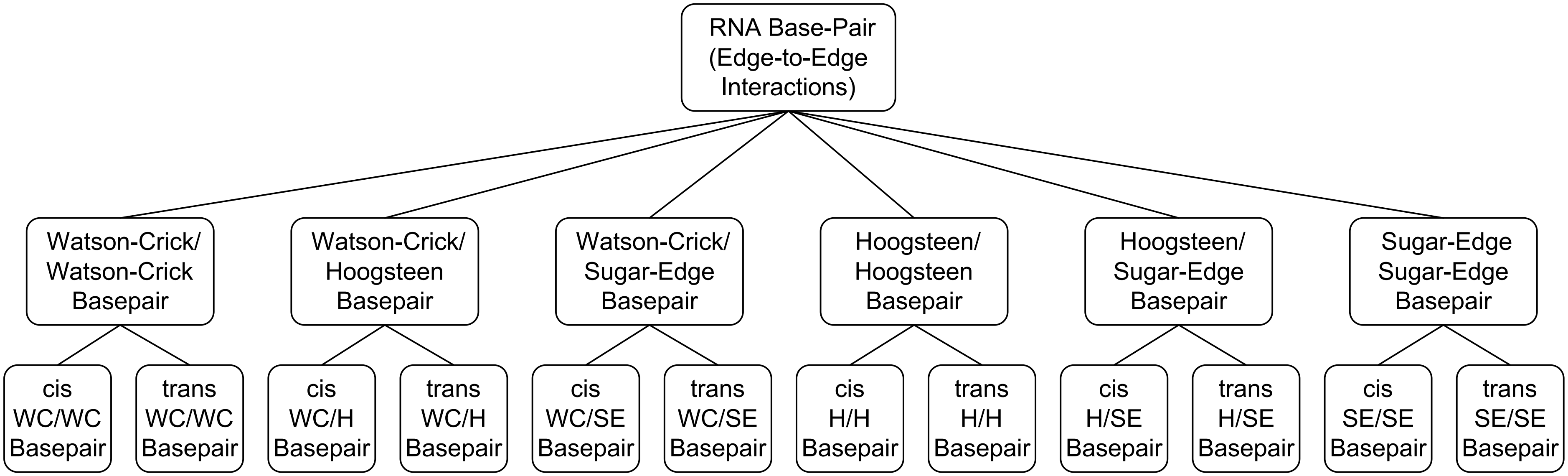
Leontis & Westhof, RNA, 2001

Geometric Basepair Types

Stombaugh et al. 2009

No.	Glycosidic Bond Orientation	Interacting Edges		Abbreviation	Symbol	Triangle Abstraction	Frequencies in 5S, 16S and 23S rRNA
		NT1	NT2				
1	*Cis*	Watson-Crick	Watson-Crick	cWW	—●—		**67.5%**
2	*Trans*	Watson-Crick	Watson-Crick	tWW	—○—		**1.4%**
3	*Cis*	Watson-Crick	Hoogsteen	cWH	●■		**1.3%**
		Hoogsteen	Watson-Crick	cHW	■●		
4	*Trans*	Watson-Crick	Hoogsteen	tWH	○□		**4.5%**
		Hoogsteen	Watson-Crick	tHW	□○		
5	*Cis*	Watson-Crick	Sugar Edge	cWS	●▶		**1.8%**
		Sugar Edge	Watson-Crick	cSW	◀●		
6	*Trans*	Watson-Crick	Sugar Edge	tWS	○▷		**1.5%**
		Sugar Edge	Watson-Crick	tSW	◁○		
7	*Cis*	Hoogsteen	Hoogsteen	cHH	—■—		**0.1%**
8	*Trans*	Hoogsteen	Hoogsteen	tHH	—□—		**1.3%**
9	*Cis*	Hoogsteen	Sugar Edge	cHS	■▶		**1.6%**
		Sugar Edge	Hoogsteen	cSH	◀■		
10	*Trans*	Hoogsteen	Sugar Edge	tHS	□▷		**7.3%**
		Sugar Edge	Hoogsteen	tSH	◁□		
11	*Cis*	Sugar Edge (priority)	Sugar Edge	cSs	—▶—		**6.6%**
		Sugar Edge	Sugar Edge (priority)	csS	—◀—		
12	*Trans*	Sugar Edge (priority)	Sugar Edge	tSs	—▷—		**5.1%**
		Sugar Edge	Sugar Edge (priority)	tsS	—◁—		

RNA Base-Pair Families:
Disjoint and Mutually Exhaustive


RNA Base-Pair (Edge-to-Edge Interactions)

Watson-Crick/ Watson-Crick Basepair

Watson-Crick/ Hoogsteen Basepair

Watson-Crick/ Sugar-Edge Basepair

Hoogsteen/ Hoogsteen Basepair

Hoogsteen/ Sugar-Edge Basepair

Sugar-Edge Sugar-Edge Basepair

cis WC/WC Basepair

trans WC/WC Basepair

cis WC/H Basepair

trans WC/H Basepair

cis WC/SE Basepair

trans WC/SE Basepair

cis H/H Basepair

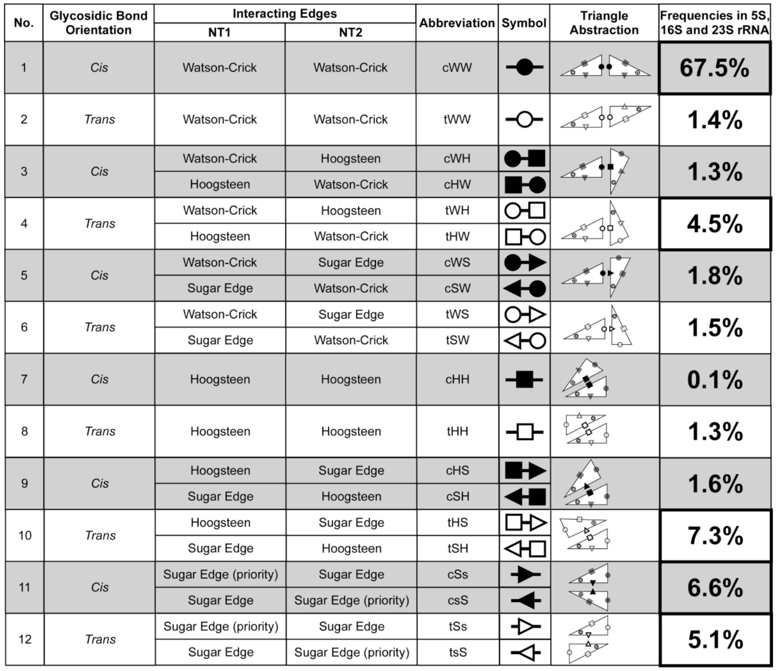
trans H/H Basepair

cis H/SE Basepair

trans H/SE Basepair

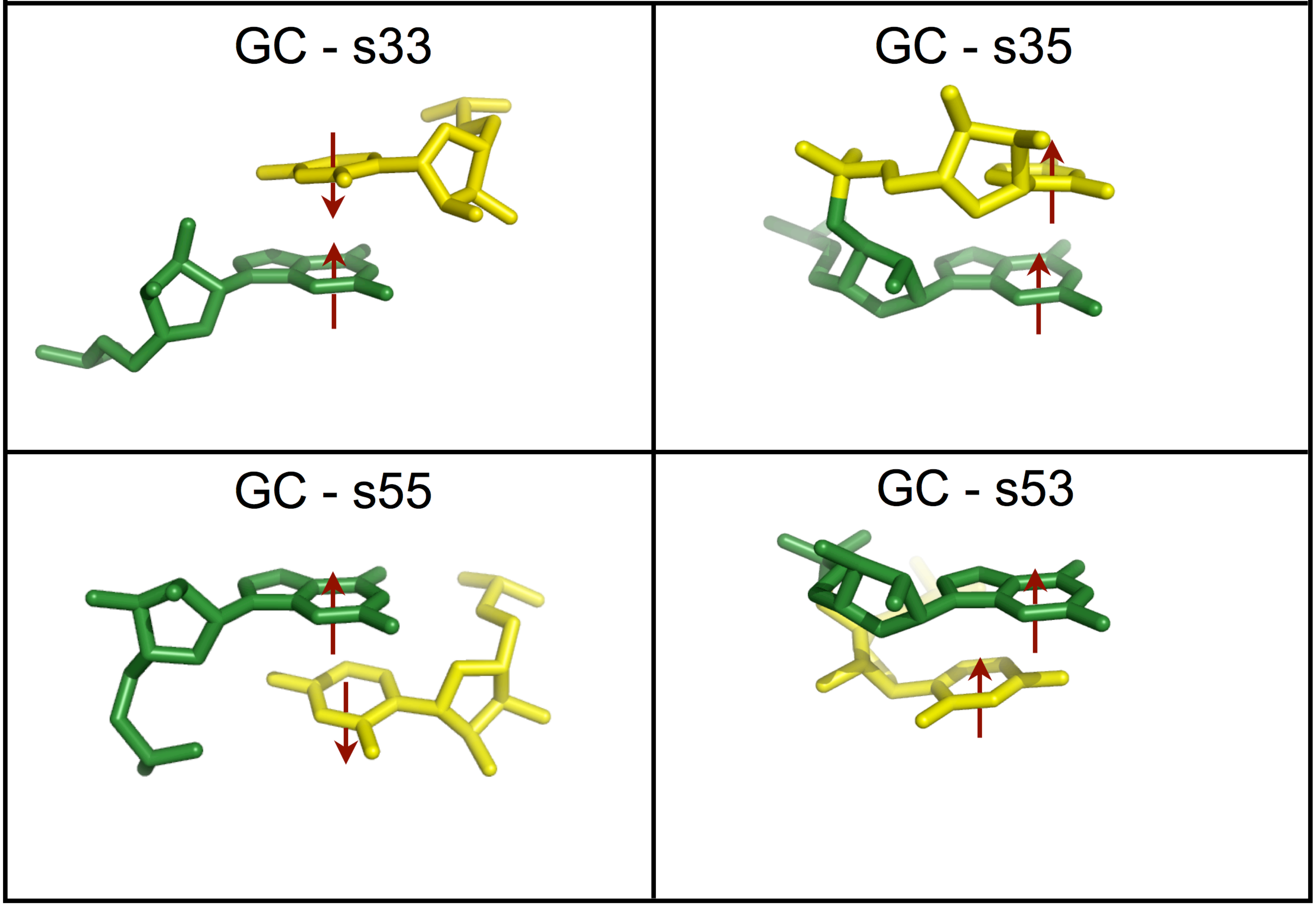
cis SE/SE Basepair

trans SE/SE Basepair



No.	Glycosidic Bond Orientation	Interacting Edges		Abbreviation	Symbol	Triangle Abstraction	Frequencies in 5S, 16S and 23S rRNA
		NT1	NT2				
1	*Cis*	Watson-Crick	Watson-Crick	cWW	—●—		**67.5%**
2	*Trans*	Watson-Crick	Watson-Crick	tWW	—○—		**1.4%**
3	*Cis*	Watson-Crick	Hoogsteen	cWH	●■		**1.3%**
		Hoogsteen	Watson-Crick	cHW	■●		
4	*Trans*	Watson-Crick	Hoogsteen	tWH	○□		**4.5%**
		Hoogsteen	Watson-Crick	tHW	□○		
5	*Cis*	Watson-Crick	Sugar Edge	cWS	●▶		**1.8%**
		Sugar Edge	Watson-Crick	cSW	◀●		
6	*Trans*	Watson-Crick	Sugar Edge	tWS	○▷		**1.5%**
		Sugar Edge	Watson-Crick	tSW	◁○		
7	*Cis*	Hoogsteen	Hoogsteen	cHH	—■—		**0.1%**
8	*Trans*	Hoogsteen	Hoogsteen	tHH	—□—		**1.3%**
9	*Cis*	Hoogsteen	Sugar Edge	cHS	■▶		**1.6%**
		Sugar Edge	Hoogsteen	cSH	◀■		
10	*Trans*	Hoogsteen	Sugar Edge	tHS	□▷		**7.3%**
		Sugar Edge	Hoogsteen	tSH	◁□		
11	*Cis*	Sugar Edge (priority)	Sugar Edge	cSs	—▶		**6.6%**
		Sugar Edge	Sugar Edge (priority)	csS	◀—		
12	*Trans*	Sugar Edge (priority)	Sugar Edge	tSs	—▷		**5.1%**
		Sugar Edge	Sugar Edge (priority)	tsS	◁—		

Four Classes of Base-Stacking Interactions
(Disjoint and Mutually Exhaustive)



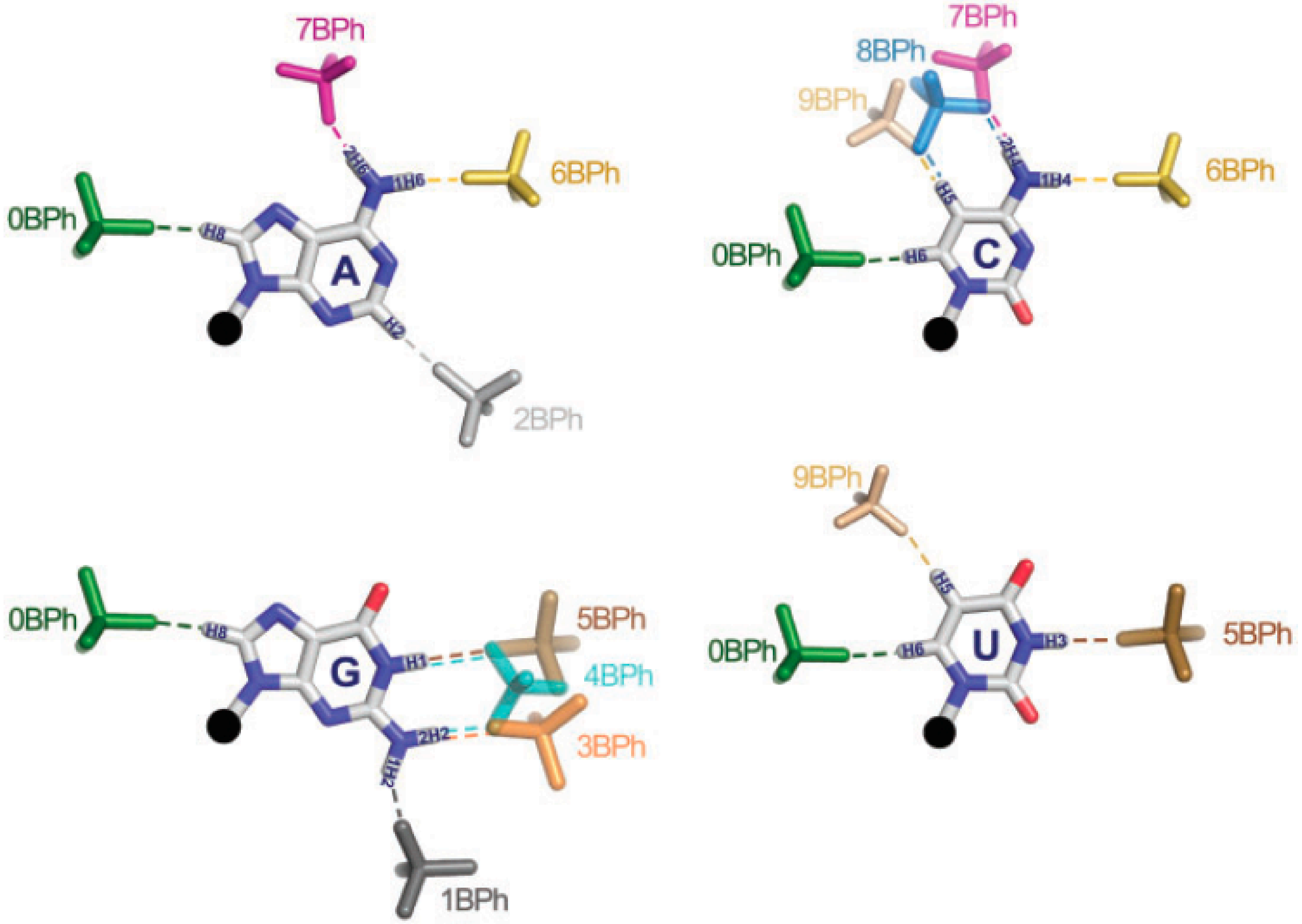
GC - s33

GC - s35

GC - s55

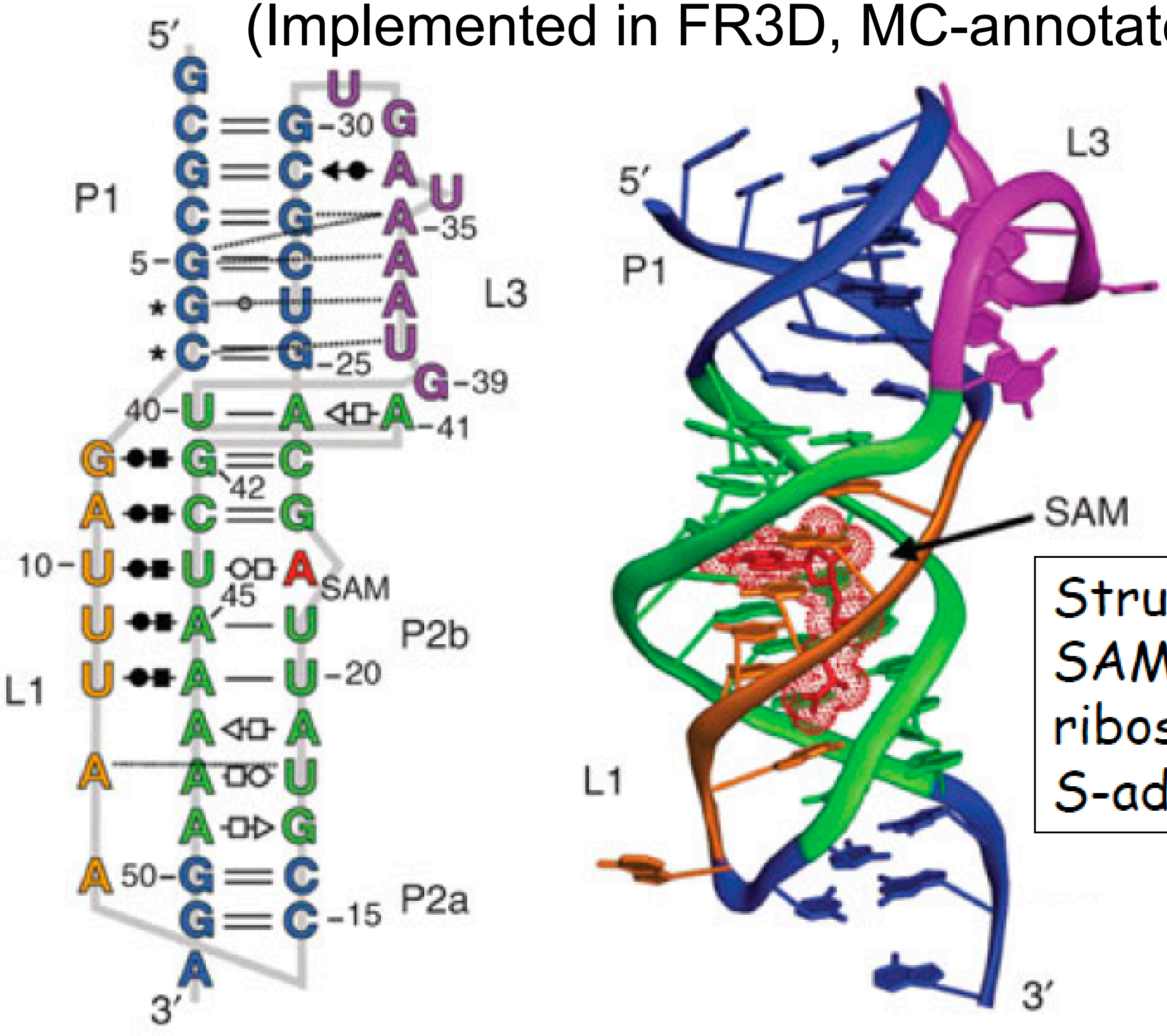
GC - s53

Classification of Base-backbone Interactions into Disjoint and Mutually Exhaustive classes



Zirbel et al. NAR 2009

Automated Annotation of RNA 3D structures (Implemented in FR3D, MC-annotate, RNAView, etc.)



Structure of the SAM-II riboswitch bound to S-adenosylmethionine

Gilbert, Rambo, Van Tyne, & Batey, Nature Struct & Mol Biol 15, 177 (2008)

Formalization

Colin Batchelor, Thomas Bittner, Robert Hoendorf, Neocles Leontis

Aim: First-order logic version based on limited number of primitives

Primitive Entities

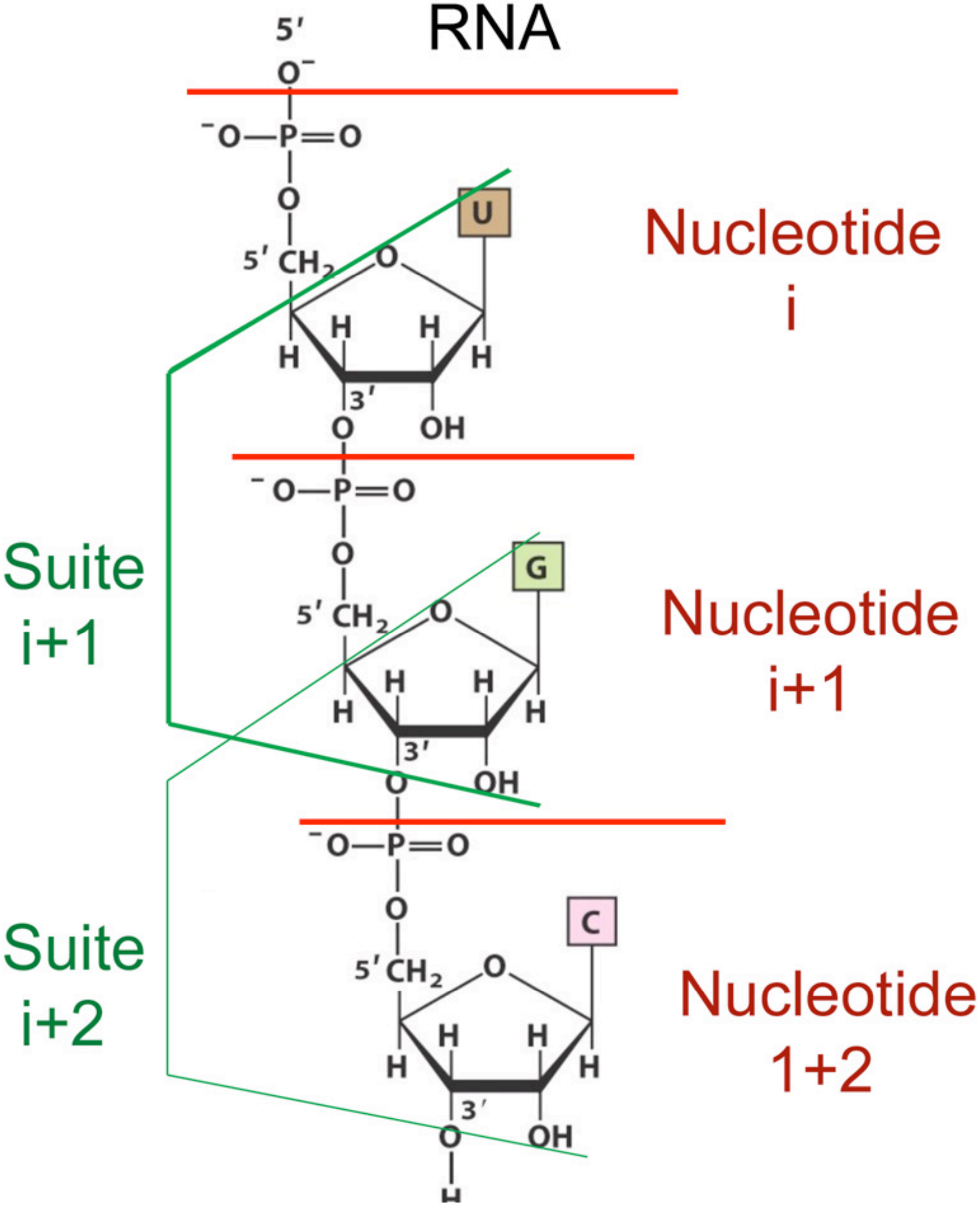
Entity	**Sub-class of**
Atom	Material Entity
Non-covalent Boundary	Boundary
Covalent Boundary	Boundary
Edge	Non-covalent Boundary
Face	Non-covalent Boundary

Covalent Boundaries

Phosphodiester Linkage (centered on the phosphorus atom P)

Links two adjacent nucleotides (i and i+1) into a suite - in this case, spanning bases U and G, their riboses, and the intervening backbone.

Suites overlap; the next suite (i+2) combines parts of nucleotides


RNA

Nucleotide i

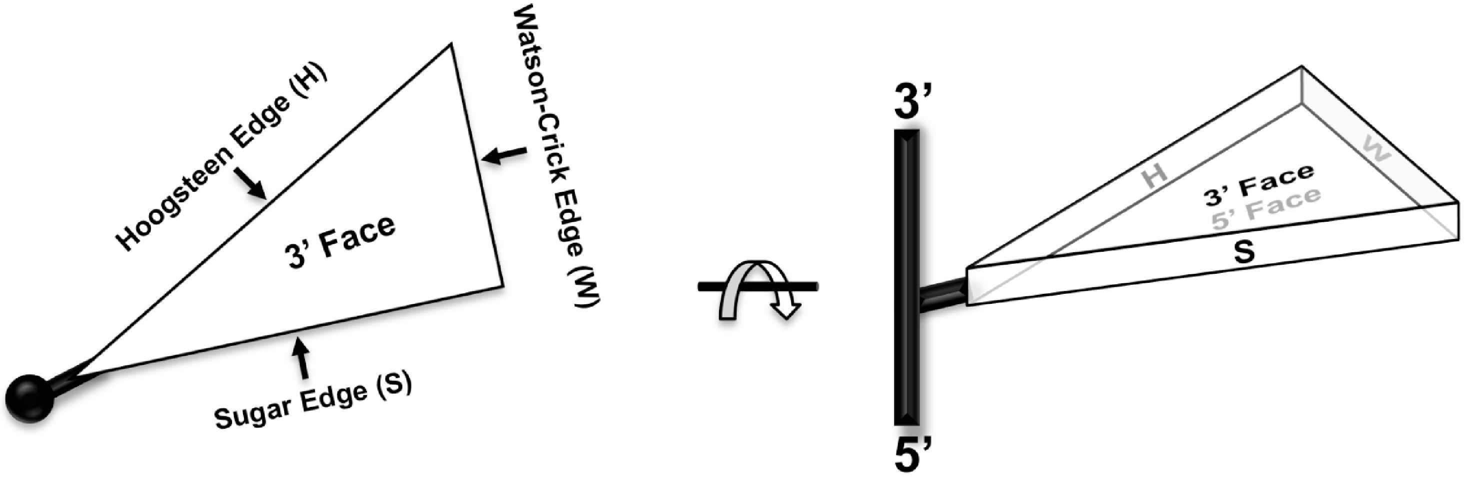
Suite i+1

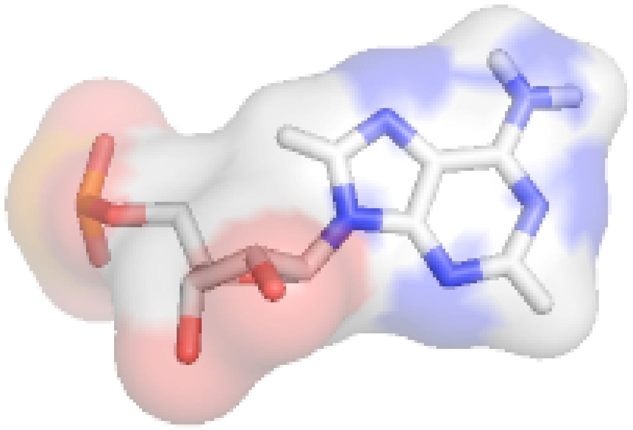
Nucleotide i+1

Suite i+2

Nucleotide 1+2

Non-covalent Boundaries





Defined Entities

Entity	**Sub-class of**
Molecule	Material Entity
Sub-molecule	Material Entity
5’_3’_strand_segment	Sub-molecule
5’_3’_RNA_molecule	Molecule
Base_pair	Collection of nucleotides
Base_stack	Collection of nucleotides
RNA Motif	Collection of nucleotides

Primitive Binary Relations

Relation	Domain	Range
Covalent_bonded_to	Atom	Atom
Weakly_interacting_ with	Atom, Sub-molecule, Molecule, Non-covalent Boundary	Atom, Sub-molecule, Molecule, Non-covalent Boundary
Edge_of	Non-covalent Boundary	Nucleotide
5’_face_of	Non-covalent Boundary	Nucleotide

Defining "Molecule"

D2: **Sum_of_collection**: x is the **sum_of_collection** P iff for all z, z **overlaps_with** x iff (there is a y such that y **member_of** P and y **overlaps_with** z).

D3: **Mereologically_maximal**: x is **mereologically_maximal** iff for all y, if y **overlaps_with** x, then y **part_of** of x.

Similar to the spatial connectedness relation of the RO, **covalently_bonded_to** is not transitive, but unlike spatial connectedness, it is irreflexive. We state this with axiom A1:

A1: For all x, x is not **covalently_bonded_to** x.

Finally, **covalently_bonded_to** is generally not functional. That is, depending on the kind of atom and the bonds a particular atom forms (single, double, triple), atom x can covalently bond to more than one distinct atom (A2).

A2: There exists x, y, z **instance_of** Atom such that [x **covalently_bonded_to** y and x **covalently_bonded_to** z and y not equal z].

D4: **Covalently_bonded_to***: **Covalently_bonded_to*** is the smallest transitive relation including **covalently_bonded_to**.

D5: **Covalently_bonded_sum_of** : x is the **covalently_bonded_sum_of** P iff [P is a **collection_of** atoms and x is the **sum_of_collection** P and (for all y and z in P, y **covalently_bonded_to*** z)].

D6: Molecule: x is a Molecule iff there is a **collection** P of Atoms such that x is the **covalently_bonded_sum_of** P and x is **mereologically_maximal**.

Defining “Base-Pair”

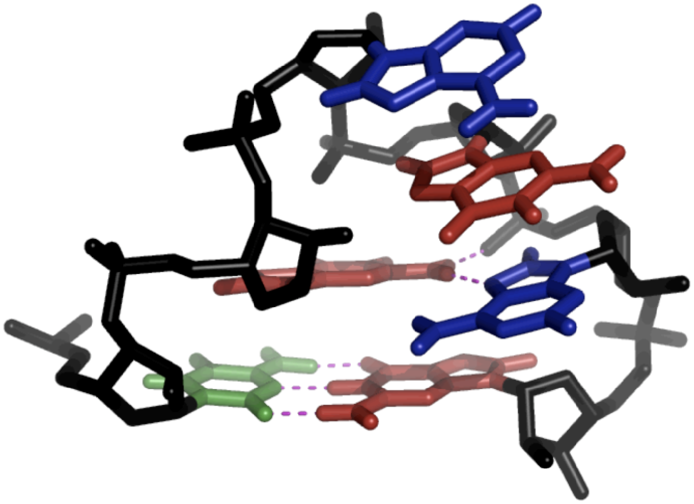
D24: Pairs_with. nt1 **pairs_with** nt2 iff [nt1 and nt2 **instances_of** *Nucleotide* and (nt1 **has_edge** e1 and nt2 **has_edge** e2 and e1 **weakly_interacting_with** e2)].
D25: Pairs_with_WH. nt1 **pairs_with_WH** nt2 iff [nt1 **pairs_with** nt2 and nt1 **has_Watson_Crick_edge** e1 and nt2 **has_Hoogsteen_edge** e2 and e1 **weakly_interacting_with** e2.]

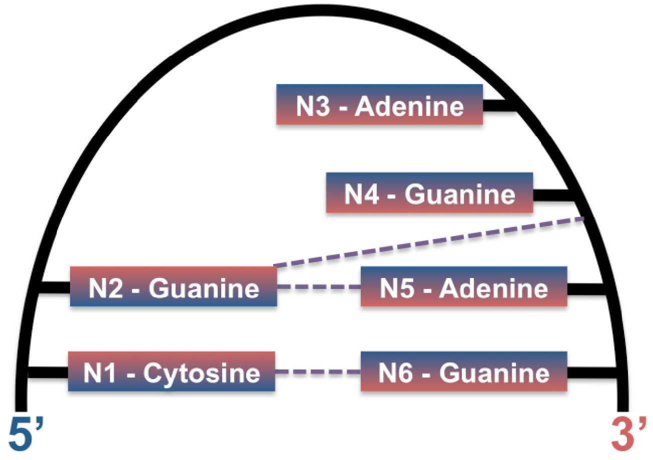
DEF: Pairs_with_cWH. nt1 **pairs_with_cWH** nt2 iff [(nt1 **pairs_with_WH** nt2) and (nt1 **cis_Glycosidic_bond_orientation** nt2)].

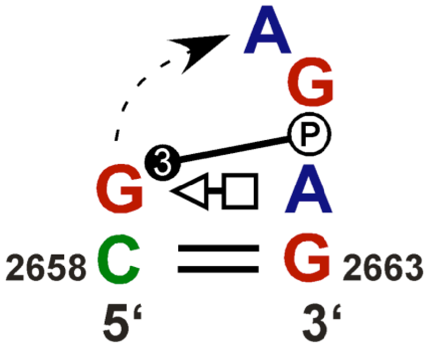
DEF: *Base-pair.* BP **instance_of** *Base-pair* iff there exist nt1 and nt2 such that (nt1 **pairs_with** nt2 and for all p with p **part_of** BP, p **overlaps_with** nt1 or p **overlaps_with** nt2).
Def: Base-pair. BP **instance_of** Base-pair iff there exist nt1 and nt2 such that (nt1 **pairs_with** nt2 and for all p with p **part_of** BP, p **overlaps_with** nt1 or p **overlaps_with** nt2).

Def: cWH_Base-pair. BP **instance_of** cWW_Base-pair iff there exist nt1 and nt2 such that (nt1 **pairs_with_cWH** nt2 and for all p with p **part_of** BP, p **overlaps_with** nt1 or p **overlaps_with** nt2).

Definition of a simple RNA motif (GNRA “tetraloop”



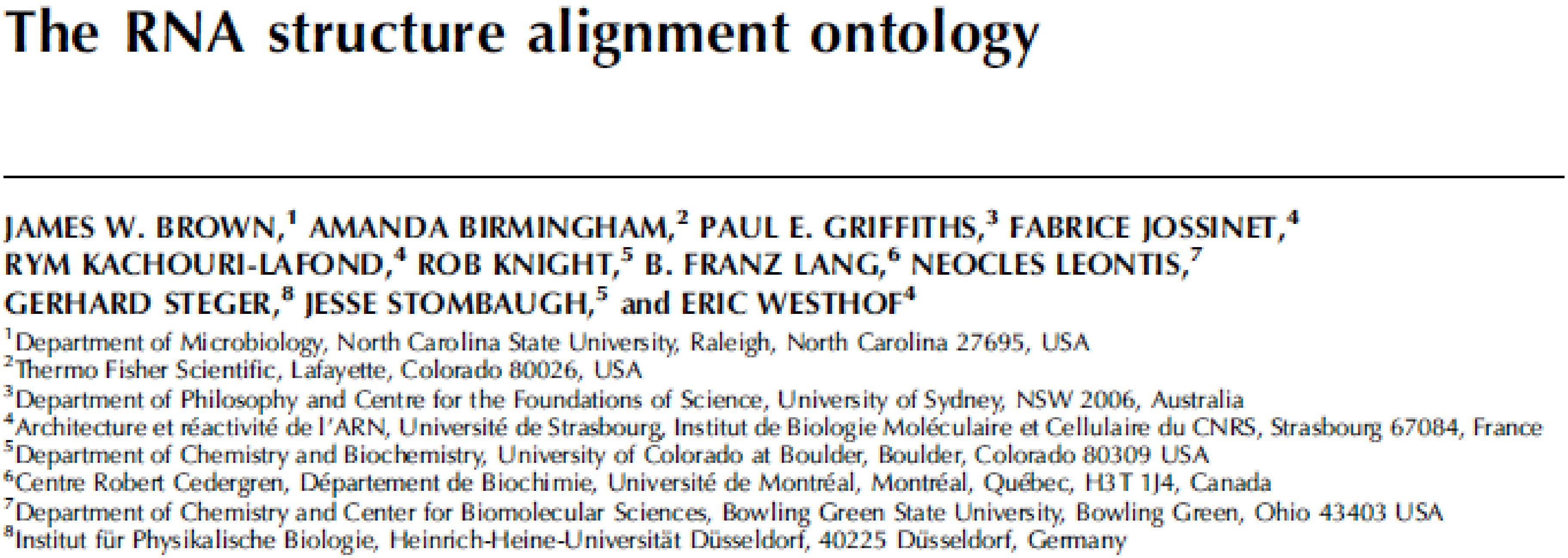




Definition of a simple RNA motif (GNRA “tetraloop”

Def: x **instance_of** GNRA_tetraloop iff There exist nt1, nt2, nt3, nt4, nt5, nt6 such that:
nt1 **cov_conn_3'_5'** nt2
nt2 **cov_conn_3'_5'** nt3
nt3 **cov_conn_3'_5'** nt4
nt4 **cov_conn_3'_5'** nt5
nt5 **cov_conn_3'_5'** nt6
nt1 **pairs_with_cWW** nt6
nt2 **pairs_with_tSH** nt5
nt1 **stack_3'_5'** nt2
nt3 **stack_3'_5'** nt4
nt5 **stack_3'_5'** nt6
nt2 **instance_of** Guanine
nt4 **instance_of** Purine
nt5 **instance_of** Adenosine
x **sum_of_collection** (nt1,nt2,nt3,nt4,nt5,nt6)

RNA Multiple Sequence Alignments



The RNA structure alignment ontology

JAMES W. BROWN,[1] AMANDA BIRMINGHAM,[2] PAUL E. GRIFFITHS,[3] FABRICE JOSSINET,[4] RYM KACHOURI-LAFOND,[4] ROB KNIGHT,[5] B. FRANZ LANG,[6] NEOCLES LEONTIS,[7] GERHARD STEGER,[8] JESSE STOMBAUGH,[5] and ERIC WESTHOF[4]

[1]Department of Microbiology, North Carolina State University, Raleigh, North Carolina 27695, USA
[2]Thermo Fisher Scientific, Lafayette, Colorado 80026, USA
[3]Department of Philosophy and Centre for the Foundations of Science, University of Sydney, NSW 2006, Australia
[4]Architecture et réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, Strasbourg 67084, France
[5]Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, Colorado 80309 USA
[6]Centre Robert Cedergren, Département de Biochimie, Université de Montréal, Montréal, Québec, H3T 1J4, Canada
[7]Department of Chemistry and Center for Biomolecular Sciences, Bowling Green State University, Bowling Green, Ohio 43403 USA
[8]Institut für Physikalische Biologie, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany

RNA, 2009

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.

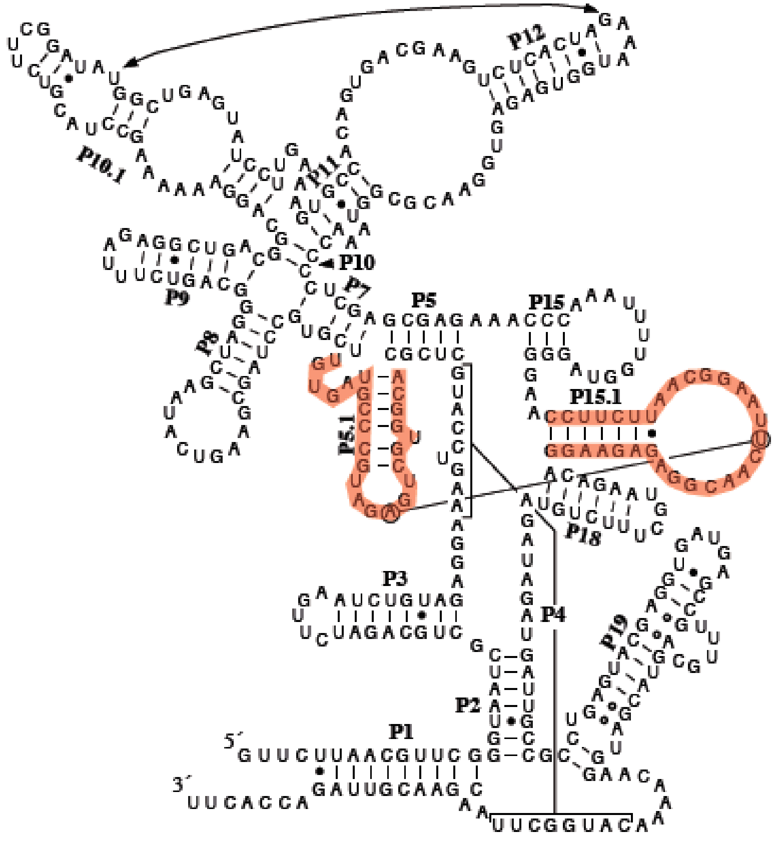
Problems with alignments

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)

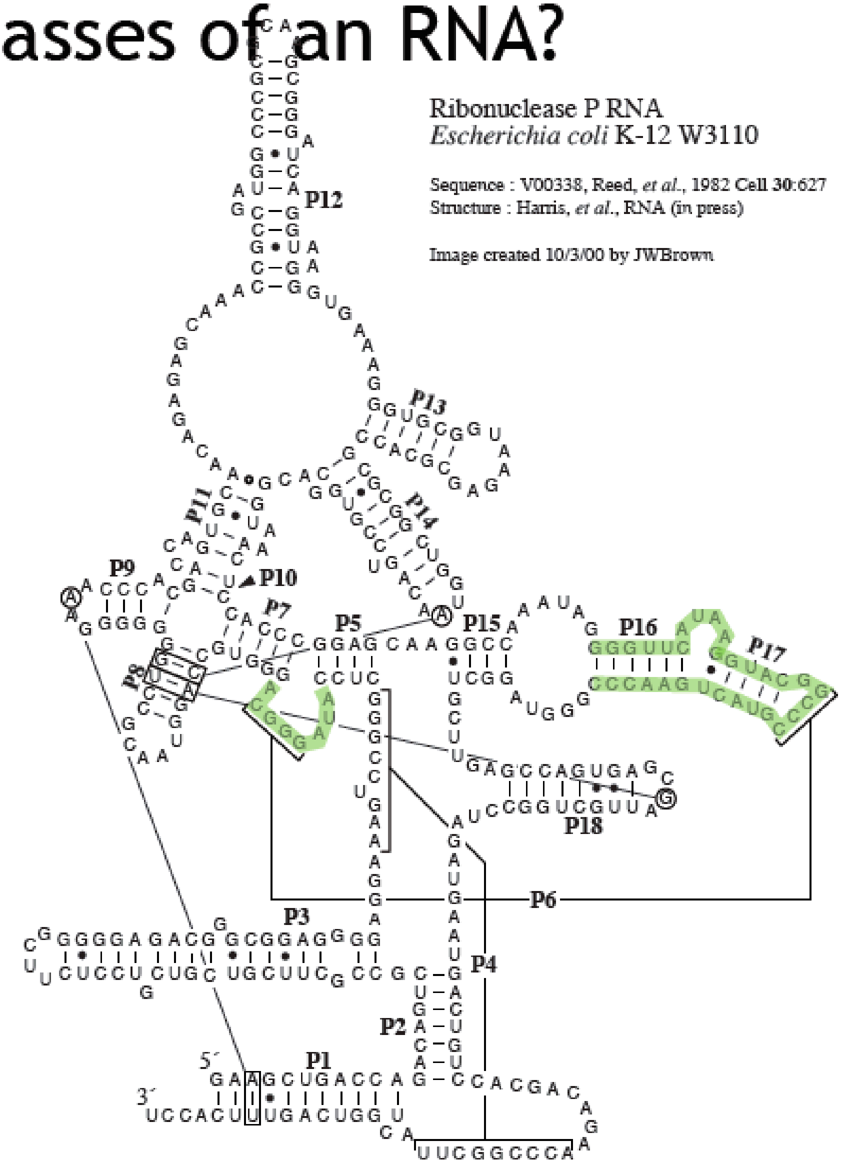
Image created 10/3/00 by JWBrown



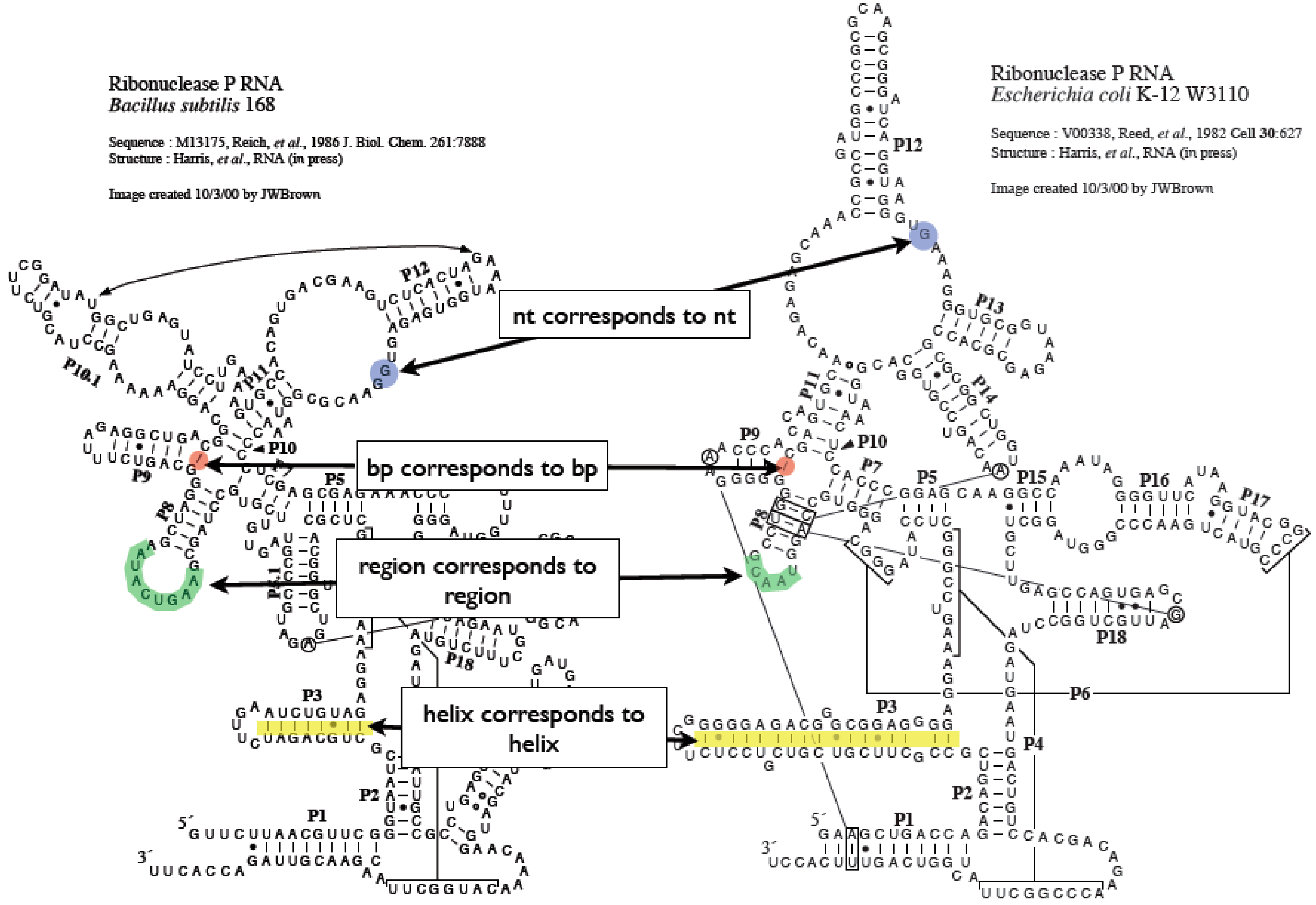
Ribonuclease P RNA
Escherichia coli K-12 W3110

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

Image created 10/3/00 by JWBrown



Capture Correspondences between Homologous RNA molecules


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
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P10.1
P11
P12
P13
P14
P15
P15.1
P16
P17
P18
5′
3′

nt corresponds to nt

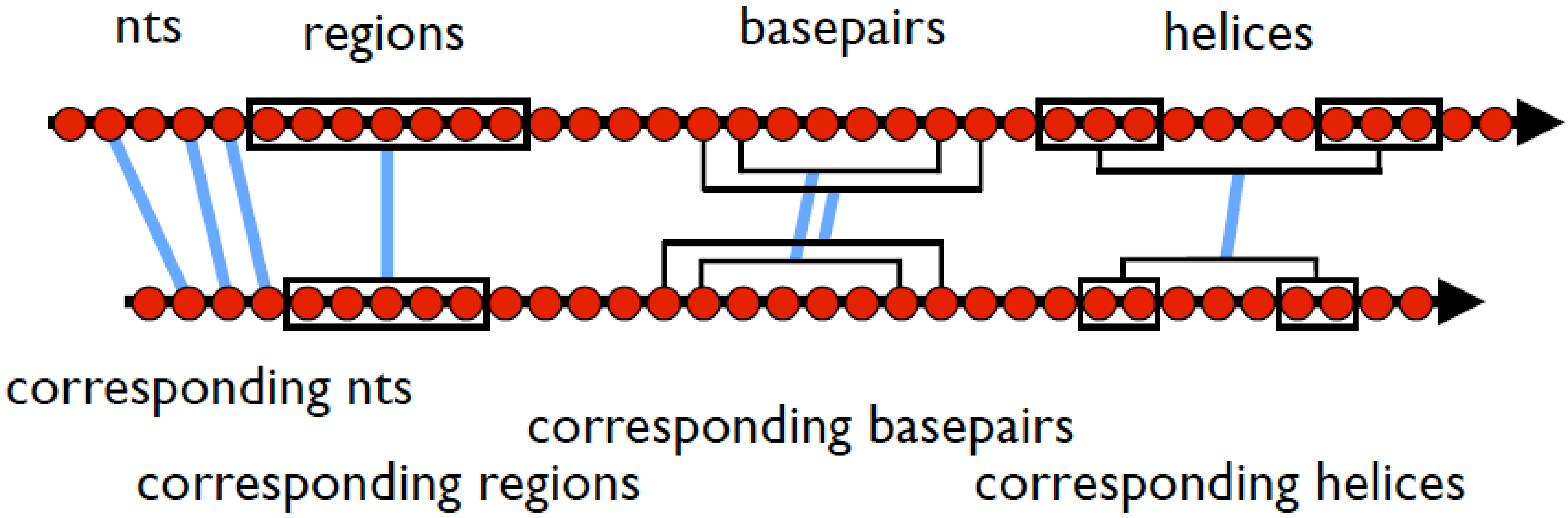
bp corresponds to bp

region corresponds to region

helix corresponds to helix

Brown et al. RNA (In press)

Solution: Define Types of RNA Sequence/Structure Elements and Correspondence Relations



nts
regions
basepairs
helices
corresponding nts
corresponding basepairs
corresponding regions
corresponding helices