ABA :
ARNomique et bioinformatique de l'ARN

Master BIBS 2e année

Daniel Gautheret

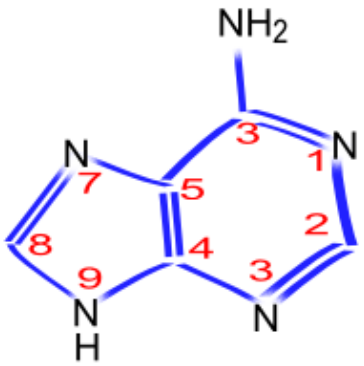
V.2009.2 rna.igmors.u-psud.fr/gautheret/cours/

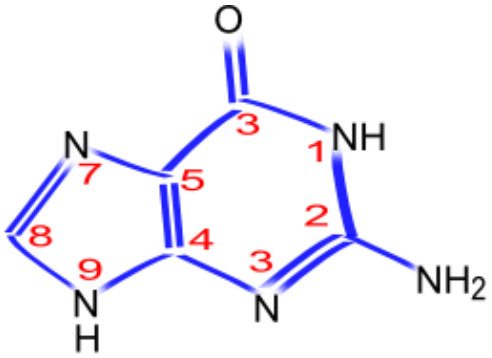
Plan du cours

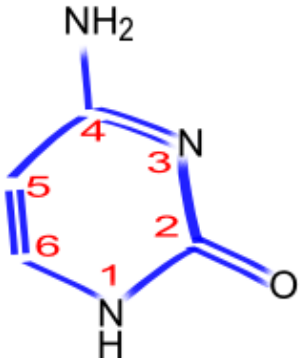
- Séance 1
  - Structure des ARN
    - Secondaire
    - Tertiaire
  - La prédiction de structure 2D
  - La modélisation 3D
- Séance 2
  - La diversité des ARN
  - ARNomique
    - Résultats expérimentaux, RFAM
  - Recherche d'ARN dans les génomes
    - ARN connus
    - ARN inconnus
- Séance 3
  - TD: l'ARNt en 3D / Mfold / Erpin

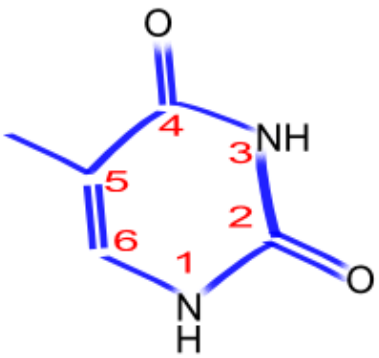
Structure des ARN

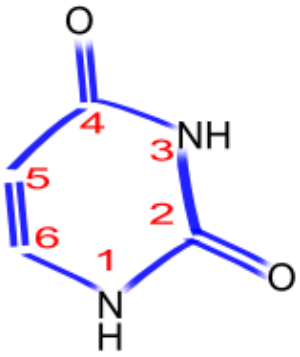
Les bases



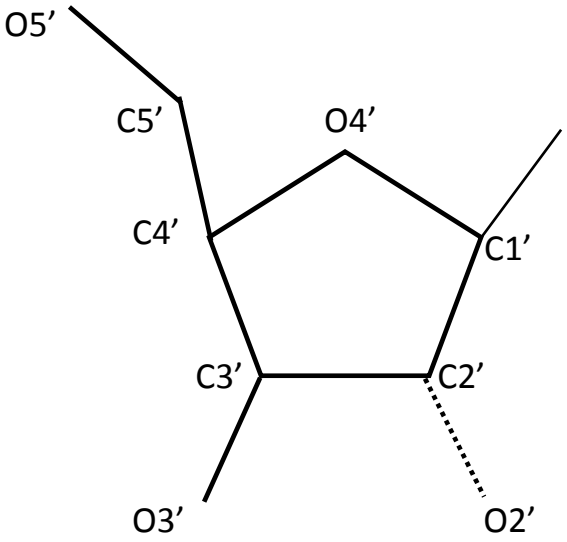


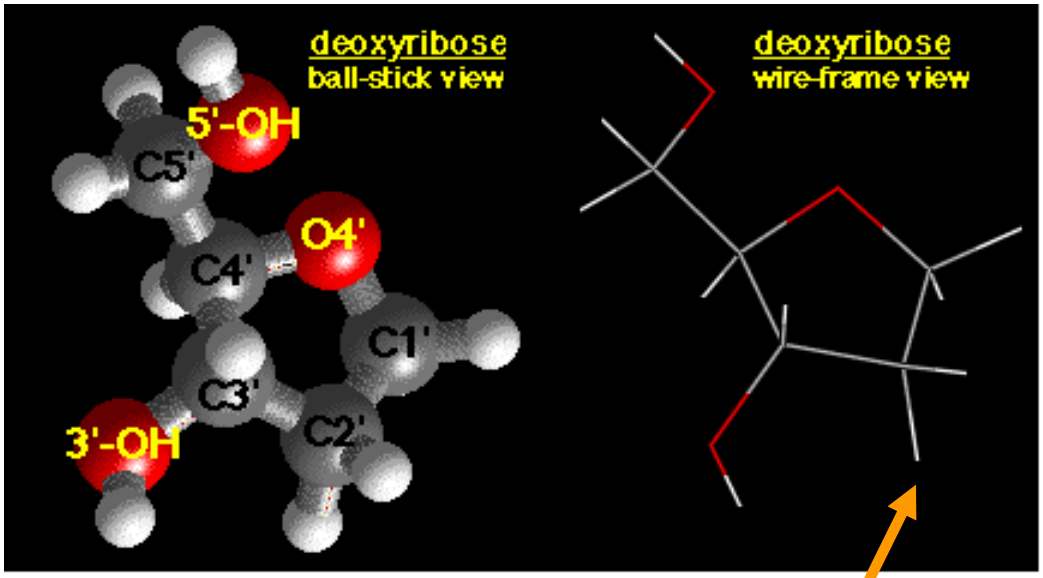






Le ribose ou déoxyribose

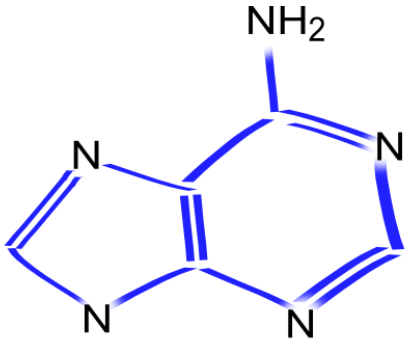




Credit: Richard Hallick. http://www.blc.arizona.edu/Molecular_Graphics/DNA_Structure/DNA_Tutorial.HTML

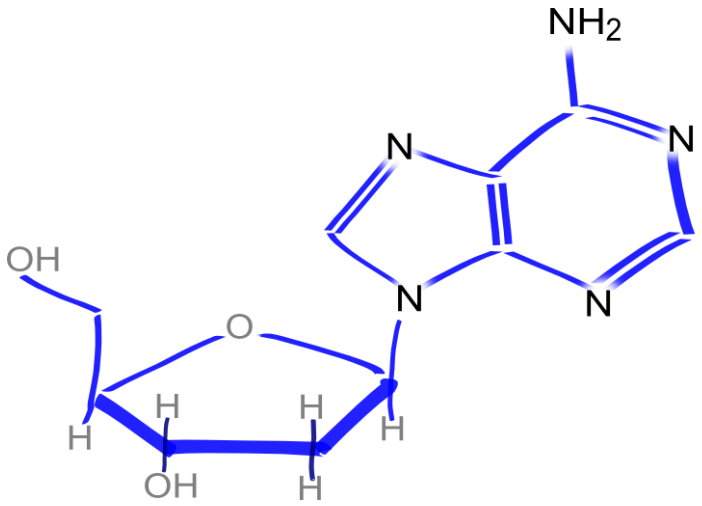
2’OH: ribose

De la base au Nucléotide

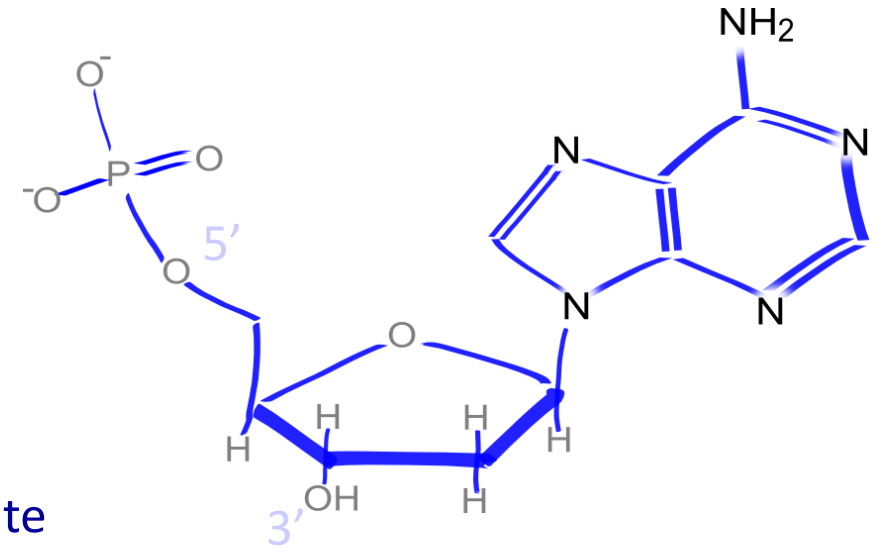


Base

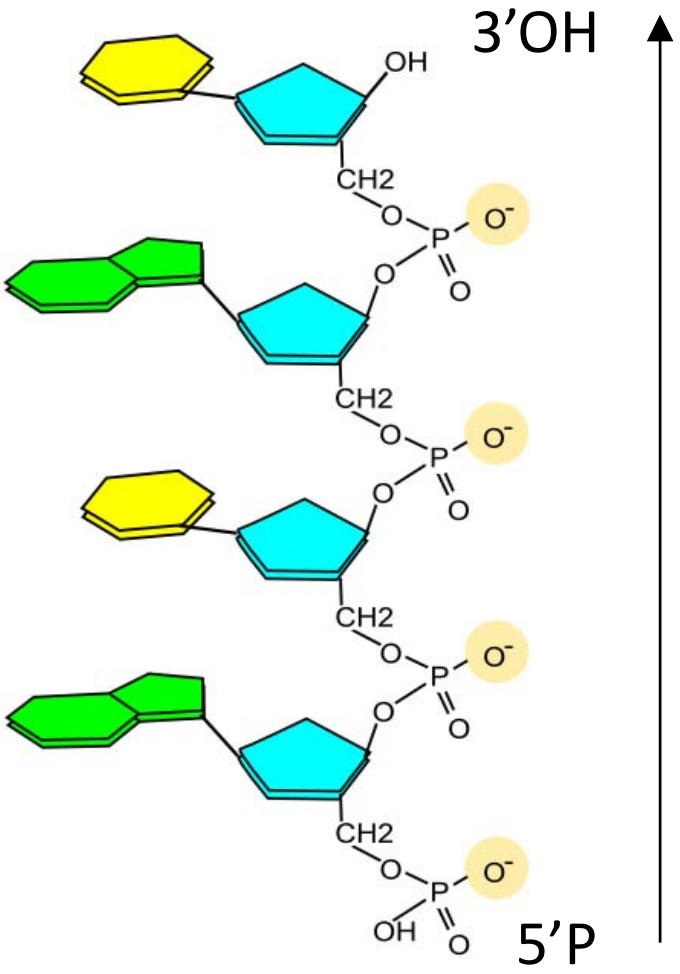
Nucléoside =
sucre + base



Nucléotide =
sucre + base + phosphate



La chaine d'ADN/ARN



Le plissement du sucre (sugar pucker)

En raison des interactions entre non covalentes entre substituants du cycle, les susbstituants ont tendance à se palcer le plus loin possible les uns des autres.
En conséquence, 4 des 5 atomes du ribose sont approximativement dans un plan, mais le 5ème sort du plan de 0,5 Å environ.
4 conformations majeures: C2': endo, C3'-exo, C3'-endo, C2' -exo.
Endo: coté base, exo: loin base

C3’ endo

C2’ endo

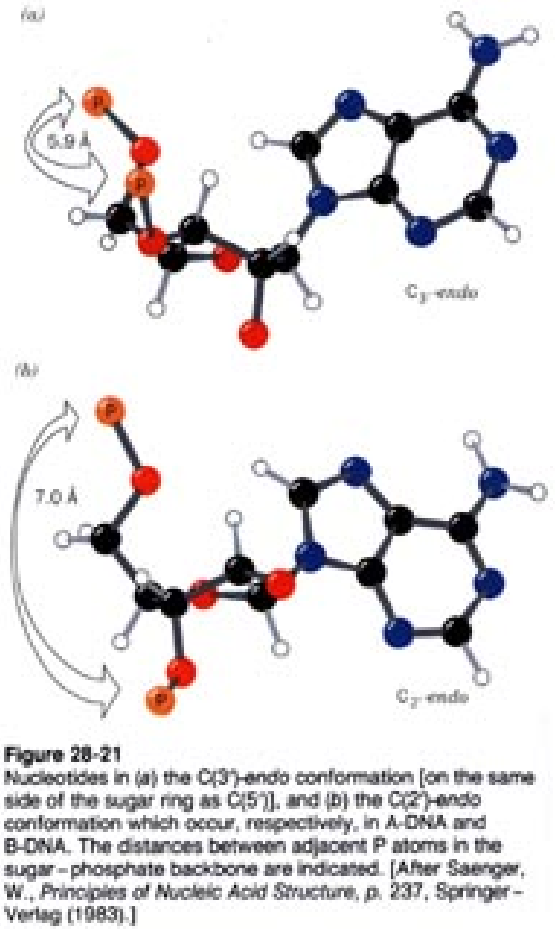


Figure 28-21
Nucleotides in (a) the C(3')-endo conformation [on the same side of the sugar ring as C(5')], and (b) the C(2')-endo conformation which occur, respectively, in A-DNA and B-DNA. The distances between adjacent P atoms in the sugar–phosphate backbone are indicated. [After Saenger, W., *Principles of Nucleic Acid Structure*, p. 237, Springer-Verlag (1983).]

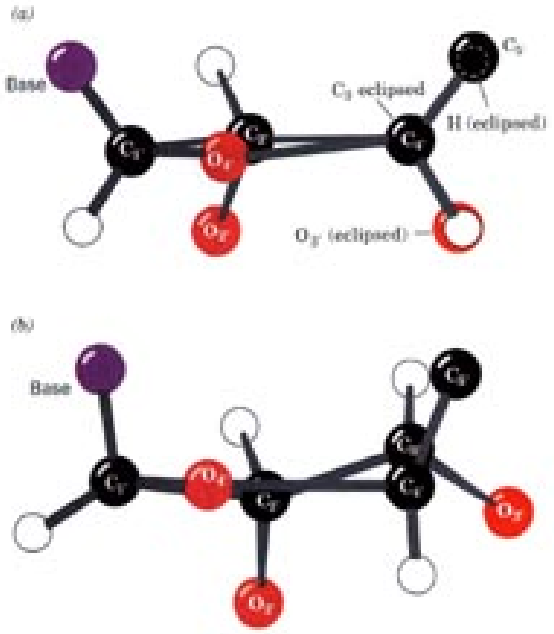
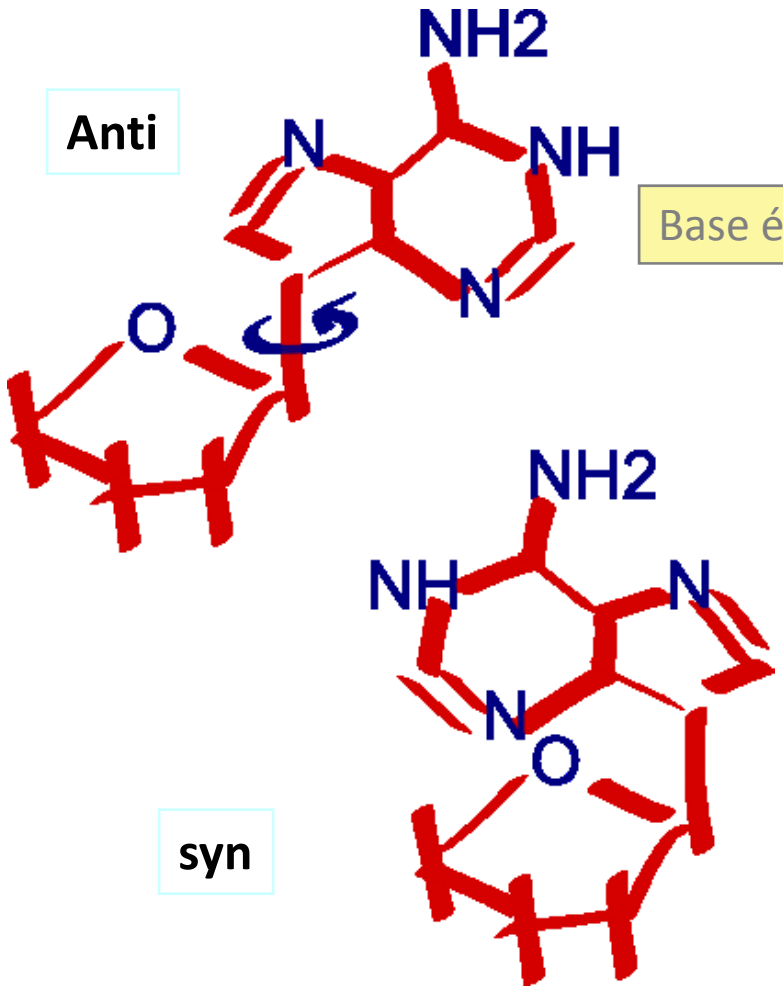


Figure 28-20
The substituents to (a) a planar ribose ring [here viewed down the C(3')—C(4') bond] are all eclipsed. The resulting steric strain is partially relieved by ring puckering such as in (b), a half-chair conformation in which C(3') is the out-of-plane atom.

L’orientation de la base

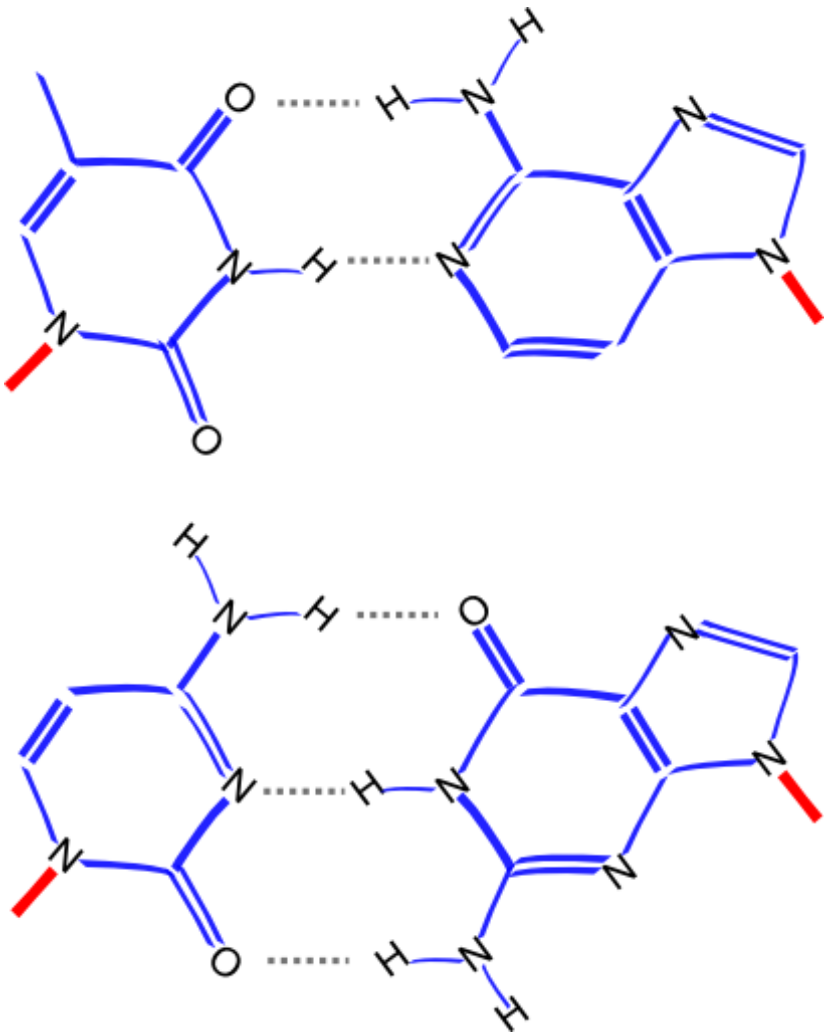


Base éloignée du sucre

Rotation autour de la liaison C’1-N9

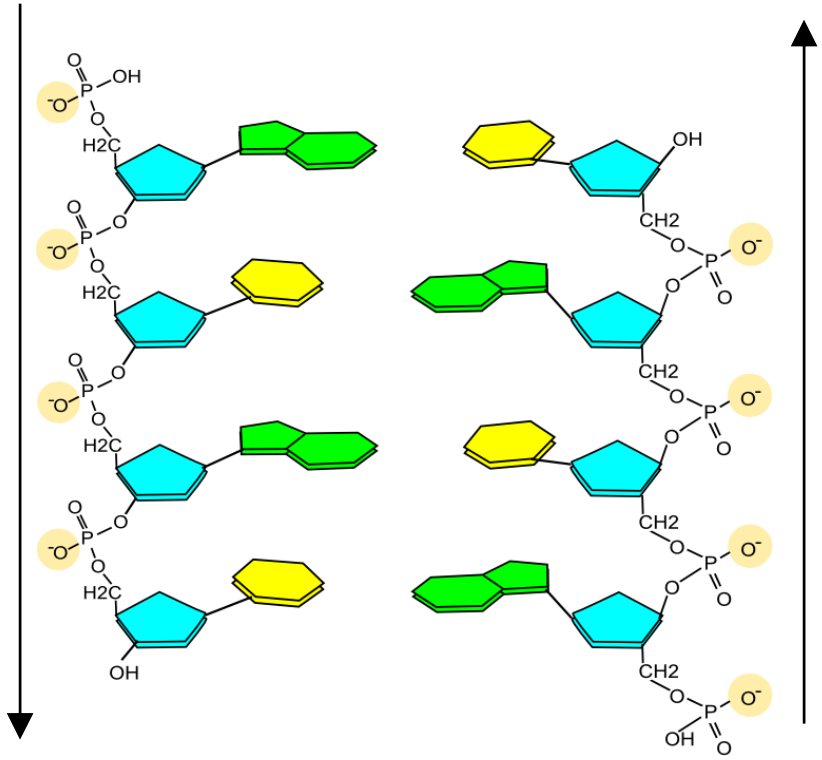
Base proche du sucre

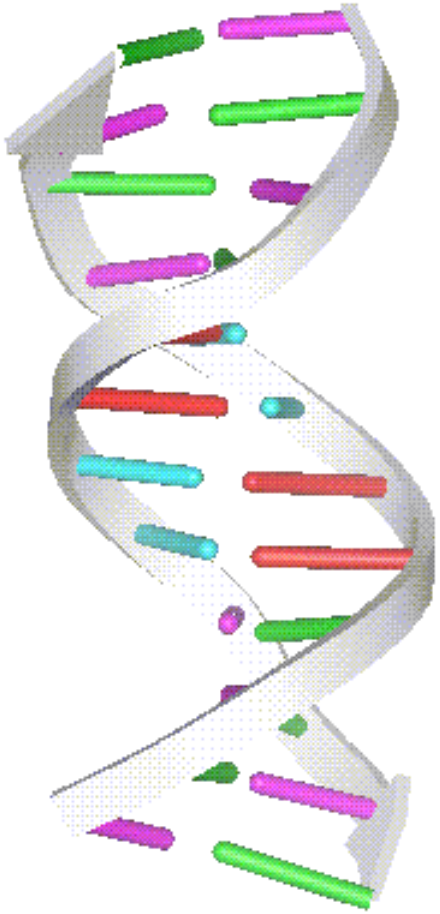
Les paires Watson-Crick



- La distance entre les deux points d'attachement des paires A - T et G - C sont identiques.

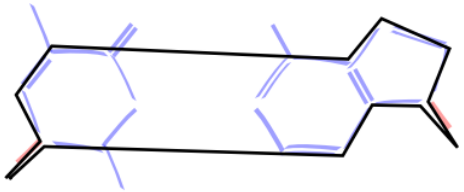
La double-hélice

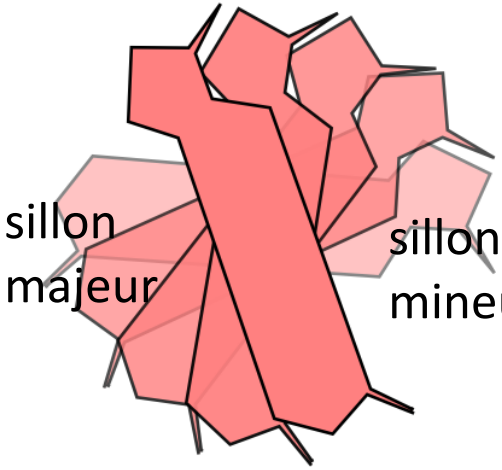


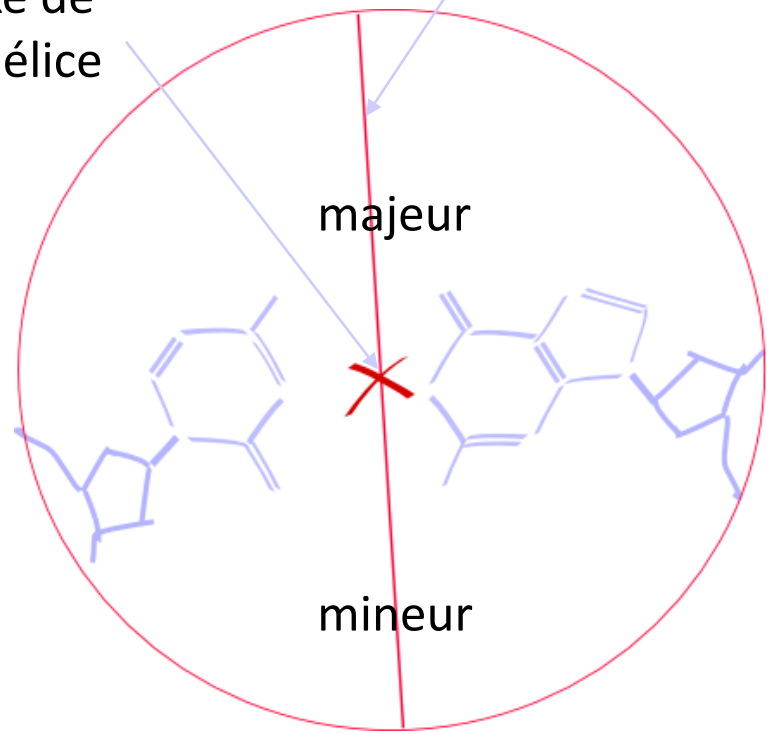


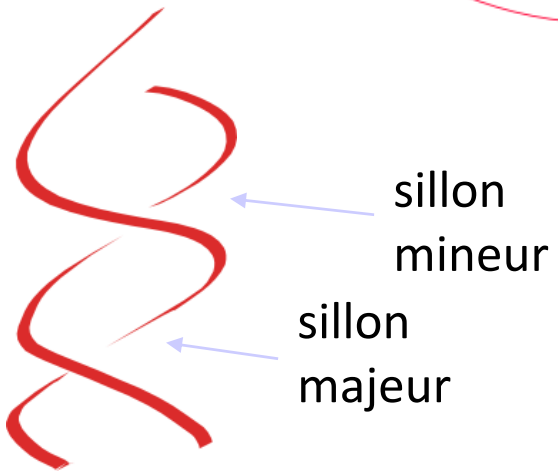
(ADN)

Les sillons



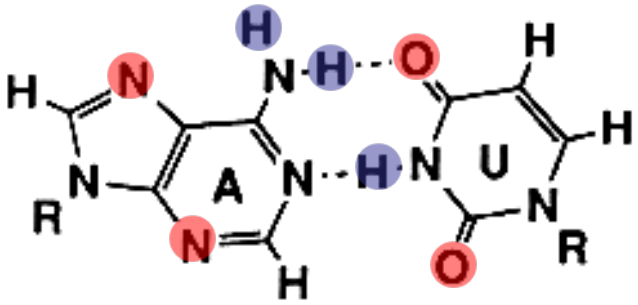

sillon
majeur
sillon
mineur


Axe de pseudo-
symétrie
Axe de
l'hélice
majeur
mineur


sillon
mineur
sillon
majeur

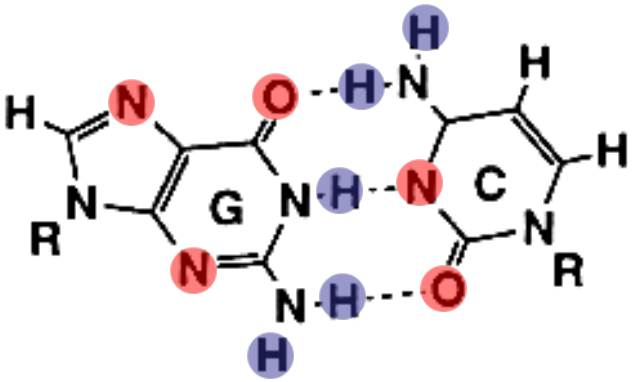
Donneurs et accepteurs de ponts-H: l'identité des paires de bases

majeur

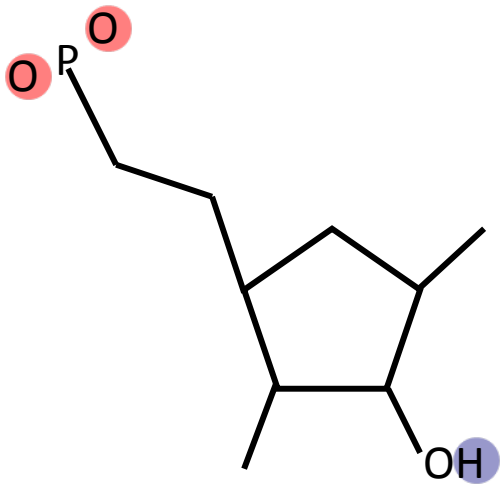


mineur

majeur



mineur



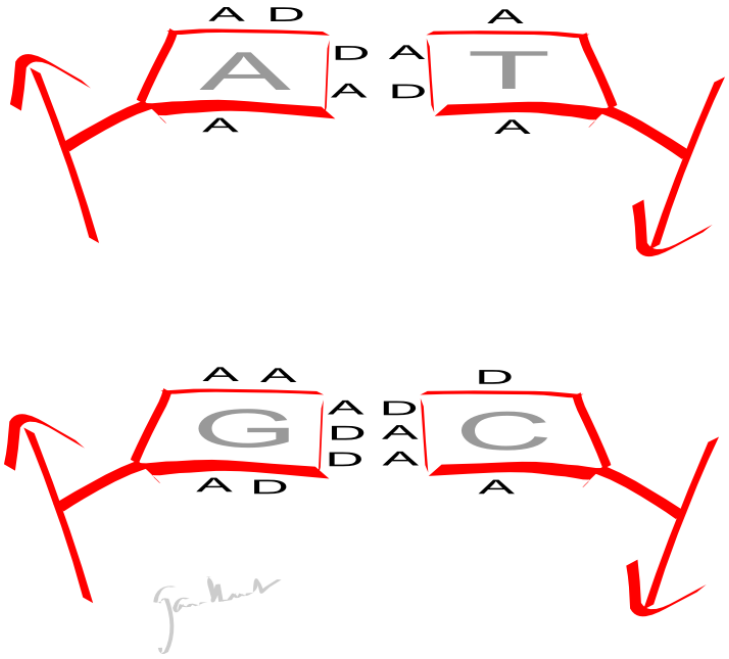
(RNA)

donneur

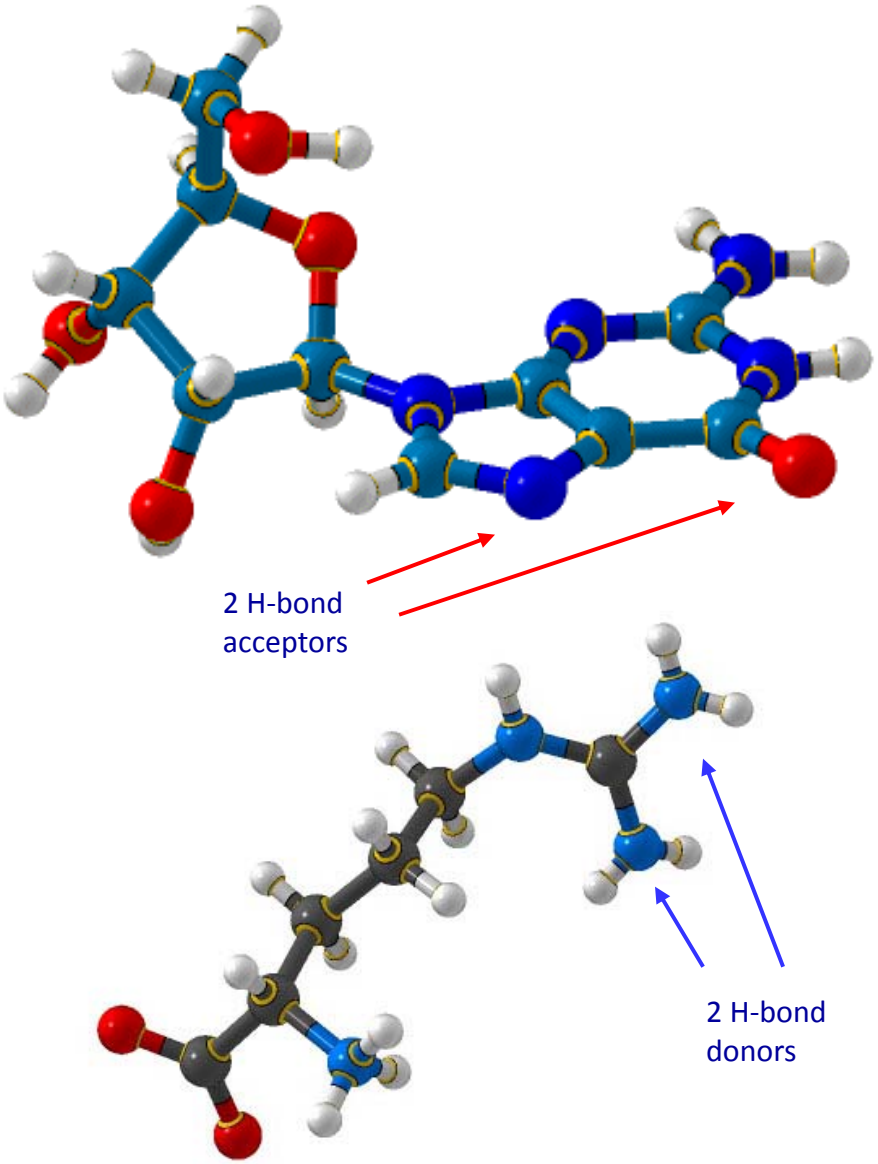
accepteur

Identité des paires de base vue des sillons

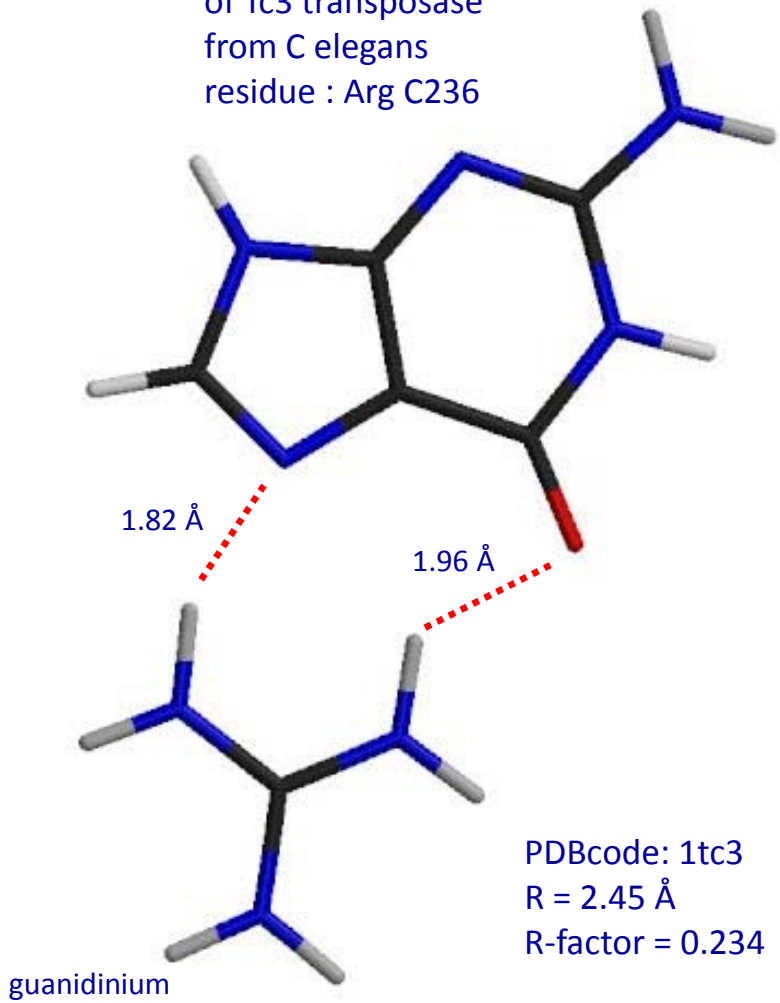
Séquence	petit sillon	grand sillon
GC	ADA	DAA
CG	ADA	AAD
AT	AA	ADA
TA	AA	ADA



Arg – Gua : a perfect H-bonding association
(33% of the total of amino acid-base pair interactions)



DNA-binding domain
of Tc3 transposase
from C elegans
residue : Arg C236



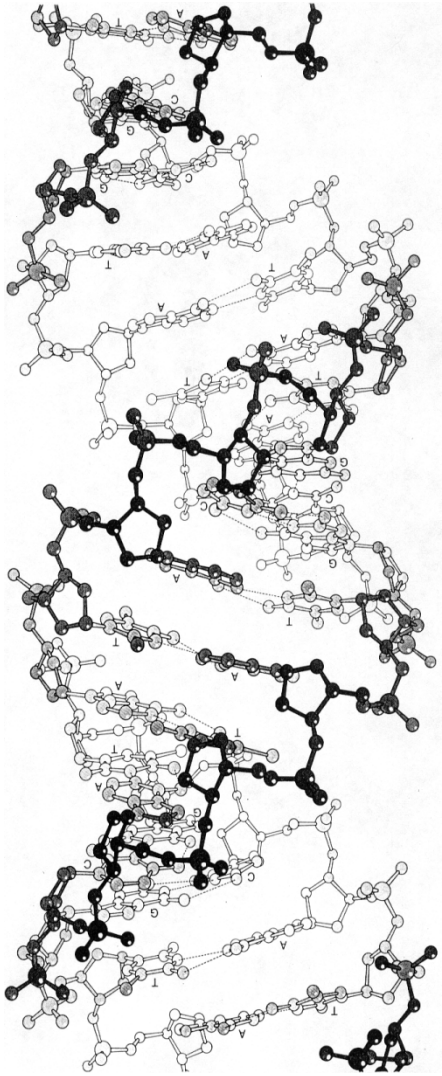
PDBcode: 1tc3
R = 2.45 Å
R-factor = 0.234

Les types d'hélice

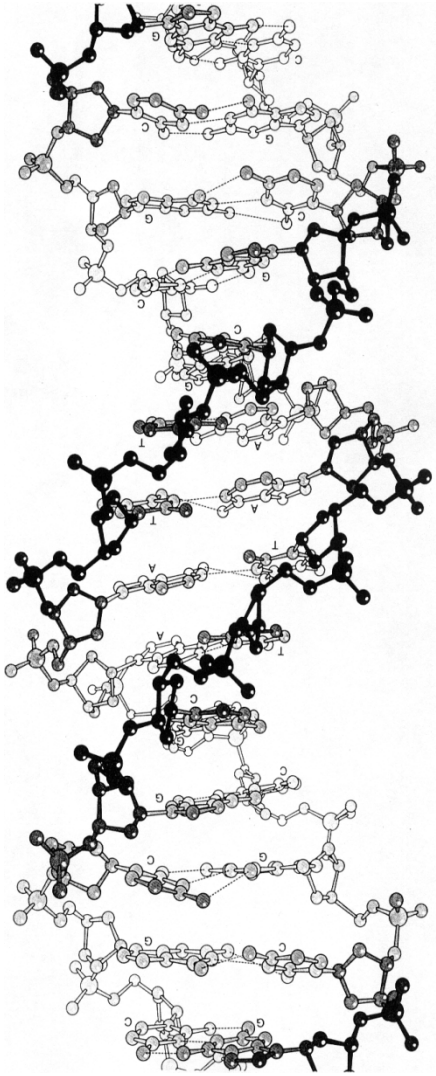
Selon: force ionique, solvents, degré d'hydratation.

	A	B	Z
	ADN/ARN	ADN	ADN/ARN
Nt/tour	11	10	12
Sens hélice	droit	droit	gauche
Confo sucre	3' endo	2'endo	2' endo (py) 3'endo (pu)
Diamètre	26 Å	20 Å	18 Å
Liaison glycosidique	anti	anti	Anti (py) Syn (pu)
Déplacement bp/axe	4 Å	aucun	
Sillon majeur largeur	3 Å	12 Å	plat
Sillon majeur prof	13,5 Å	9 Å	
Sillon mineur largeur	11 Å	6 Å	étroit
Sillon mineur prof	3 Å	7,5 Å	profond

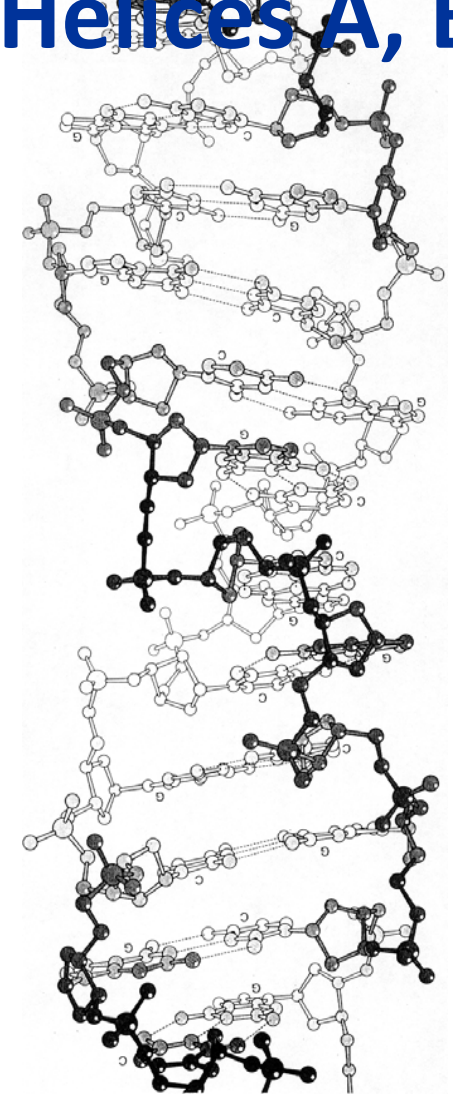
Hélices A, B et Z



A: ADN/ARN

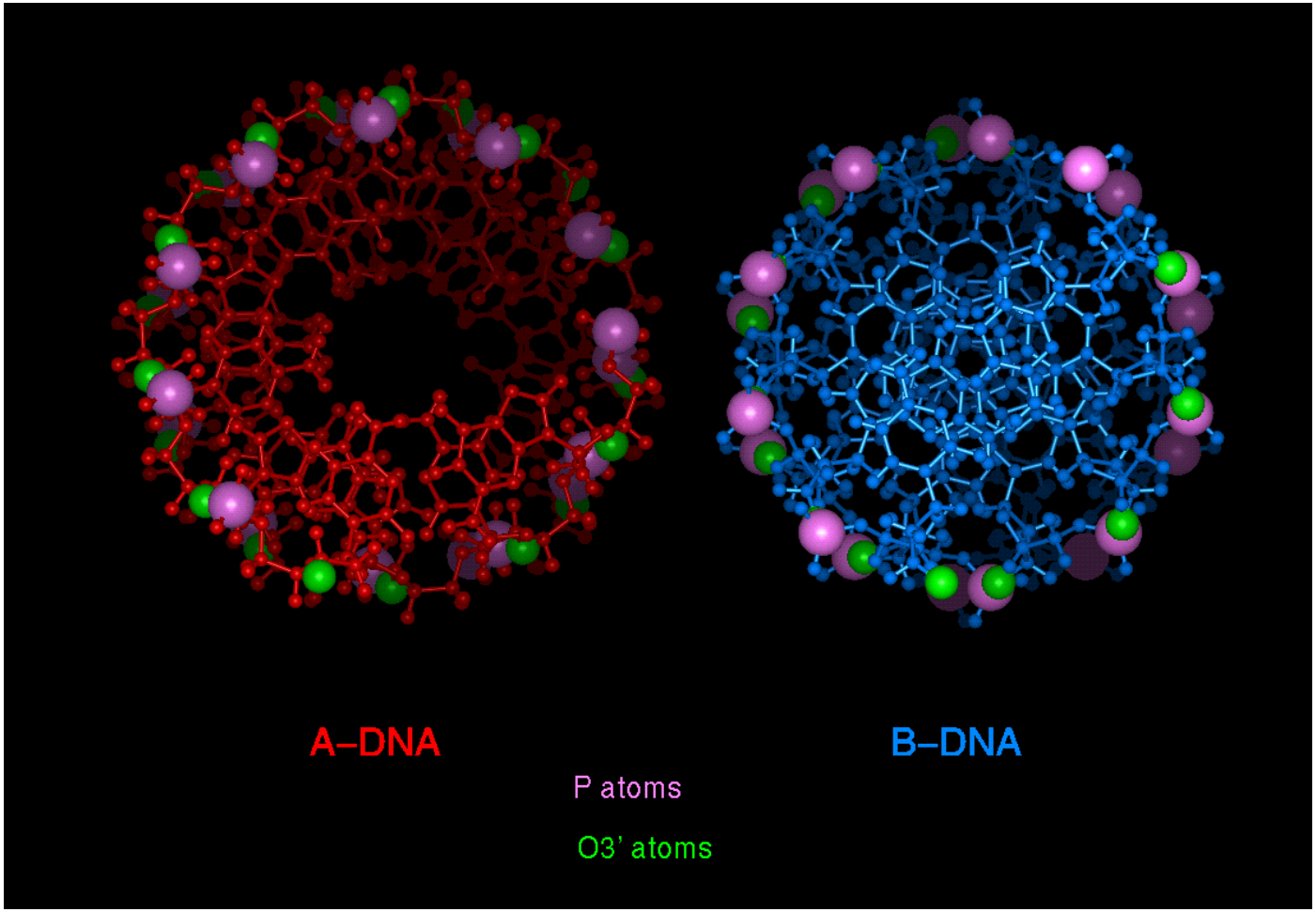


B: ADN



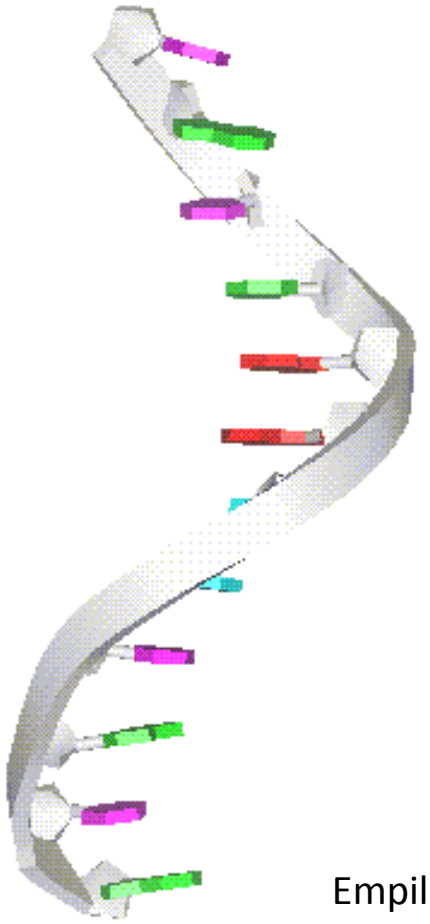
Z: ADN/ARN

Vues axiales des hélices A et B



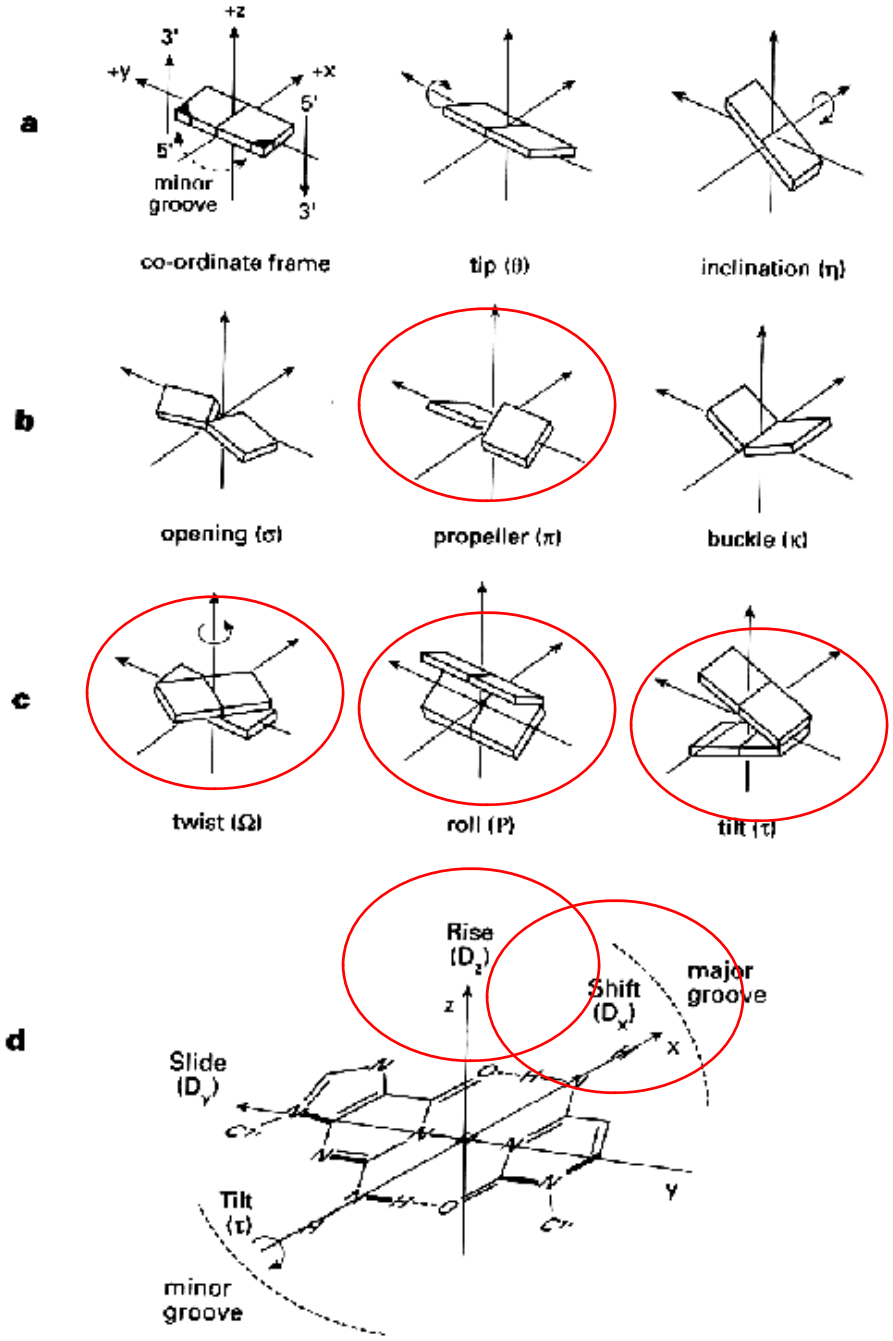
L’empilement des bases (stacking)

- L’empilement n’est pas imposé par la double hélice: il est l’un des principaux facteurs de stabilité des A.N.
- Cycles aromatiques séparés d’environ 4 Å
- Causes: hydrophobicité et interactions VdW
- Les séquences YpR et PrY ont un empilement différent: La séquence influe sur la stabilité et le forme de l’hélice.



Empilement dans un brin de type B

Paramètres des hélices



On peut définir une double hélice avec les paramètres suivants:

Tilt (theta-t): autour de l'axe pseudo-sym

Twist (t): tour/residu autour axe hélice

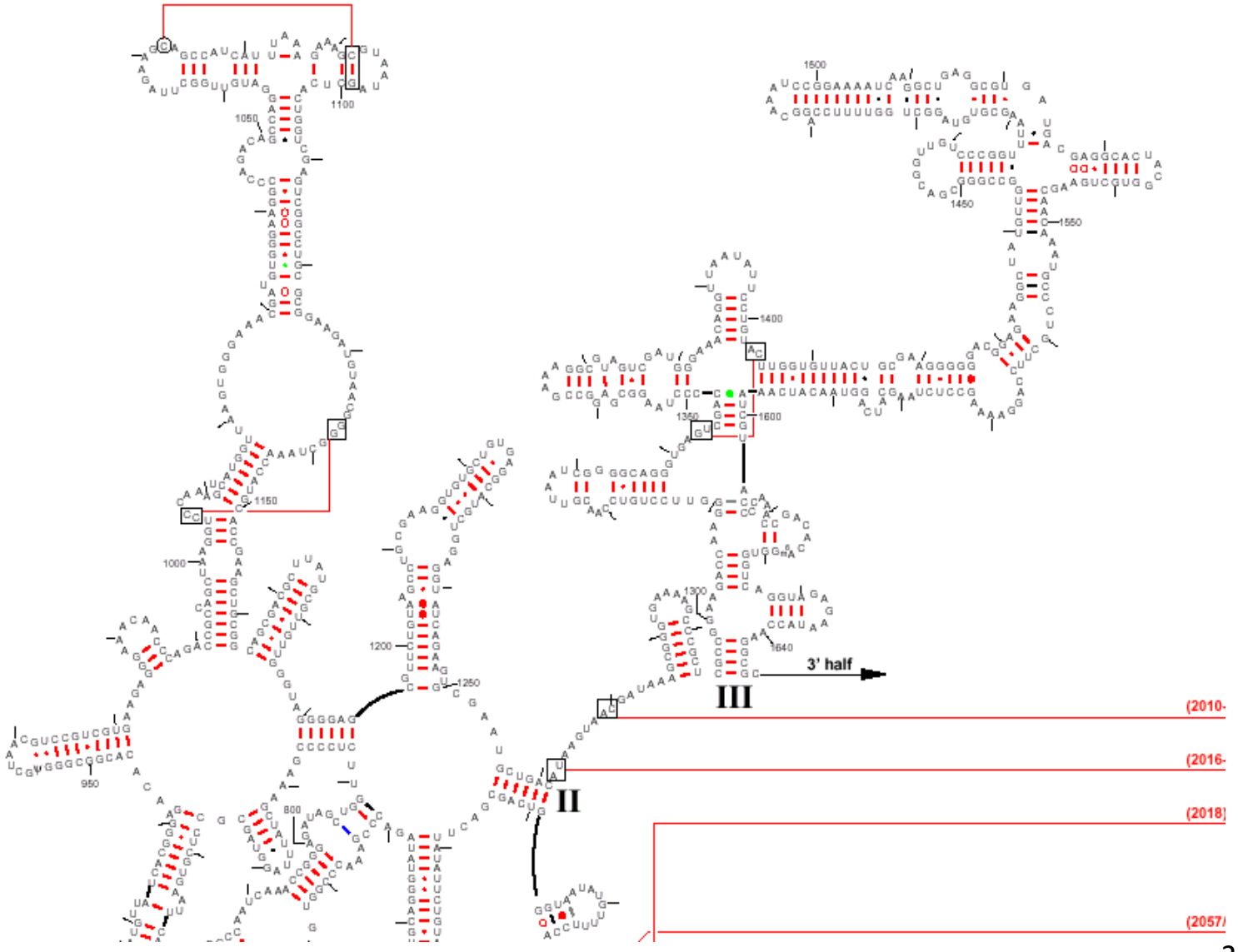
propeller Twist (theta-p): entre une base et son appariée

axial rise (h): élévation/residu

Dislocation (D): distance entre centre bp et axe hélice

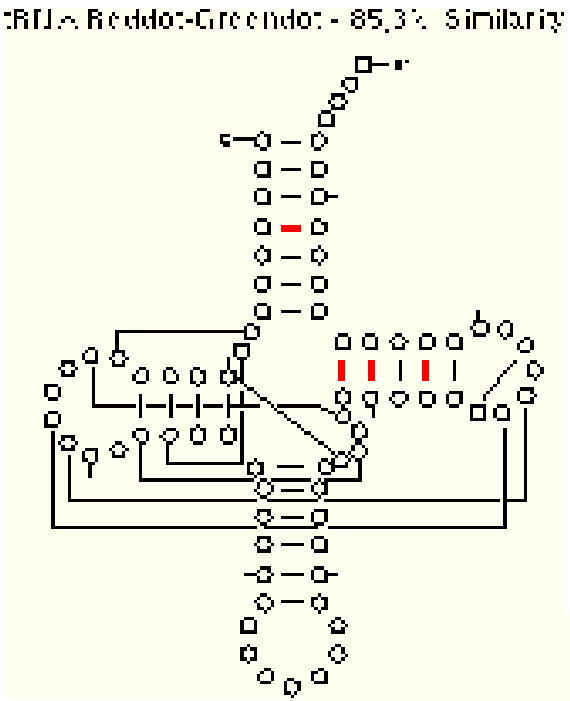
Roll (theta-r): angle autour du 3ème axe (axe C6-C8).

ARN: les structures secondaires

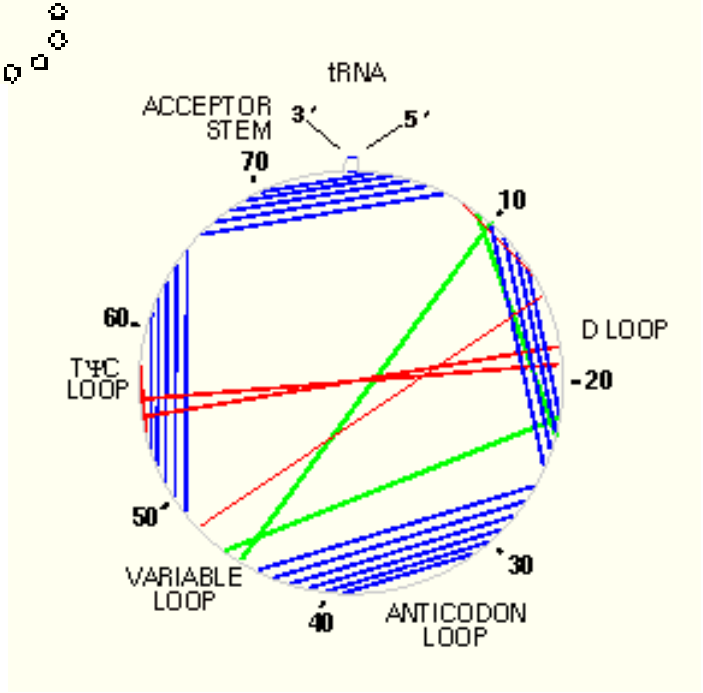


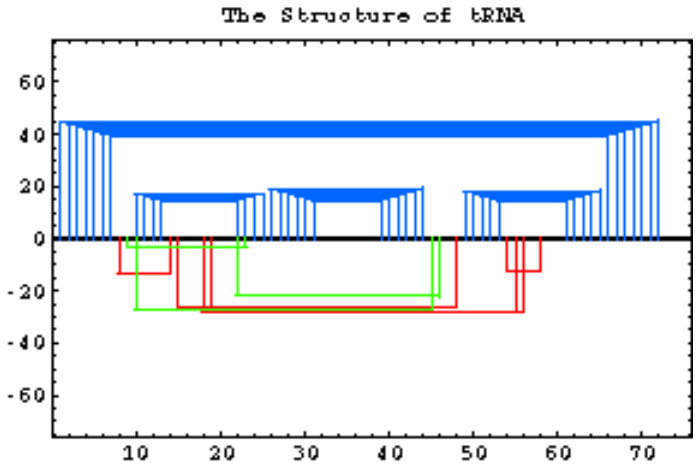
Extrait ARNr 23S (R. Gutell)

Interactions secondaires et tertiaires

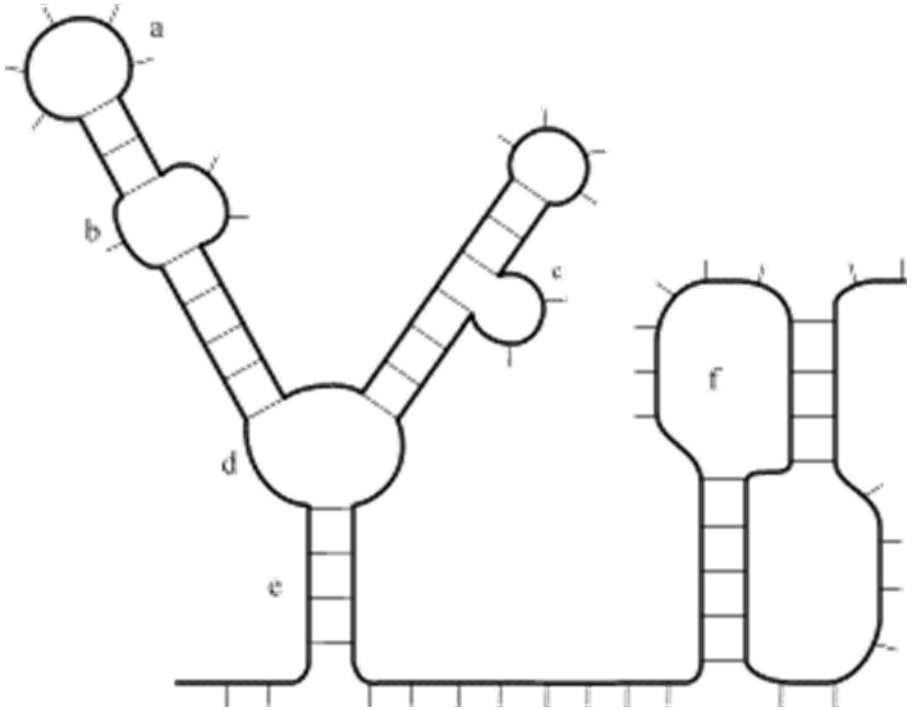


Structure secondaire: sous ensemble maximum des paires de bases imbriquées (sans croisement)





Eléments de structure secondaire



a. épingle à cheveux (hairpin)

b. interne

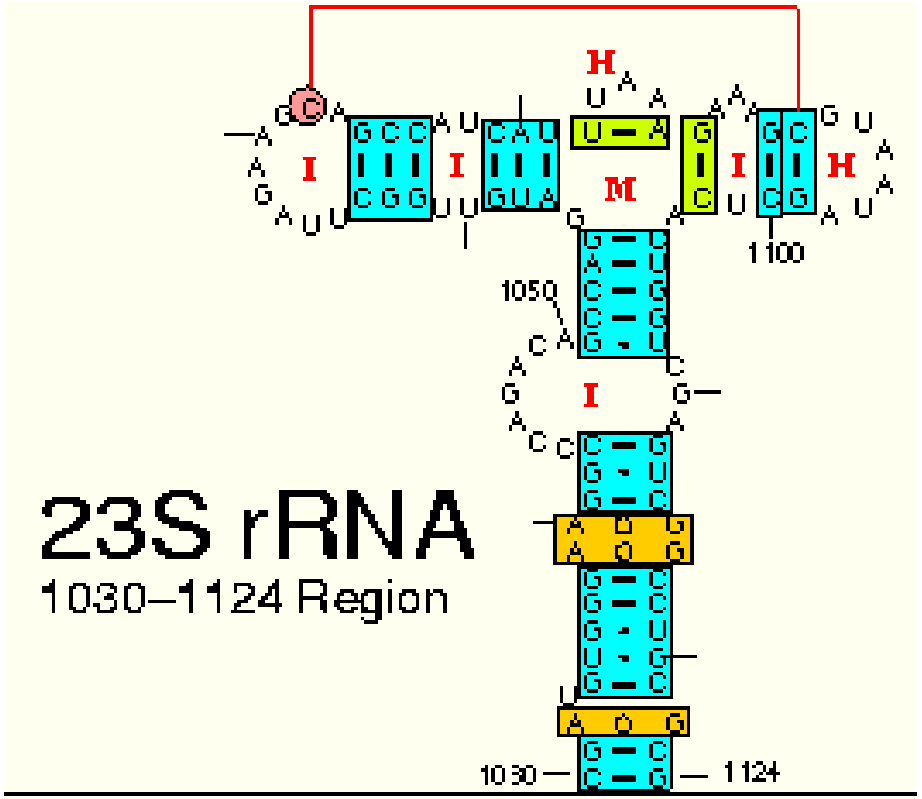
c. bulge

d. multi-branche

e. duplex (longue-distance)

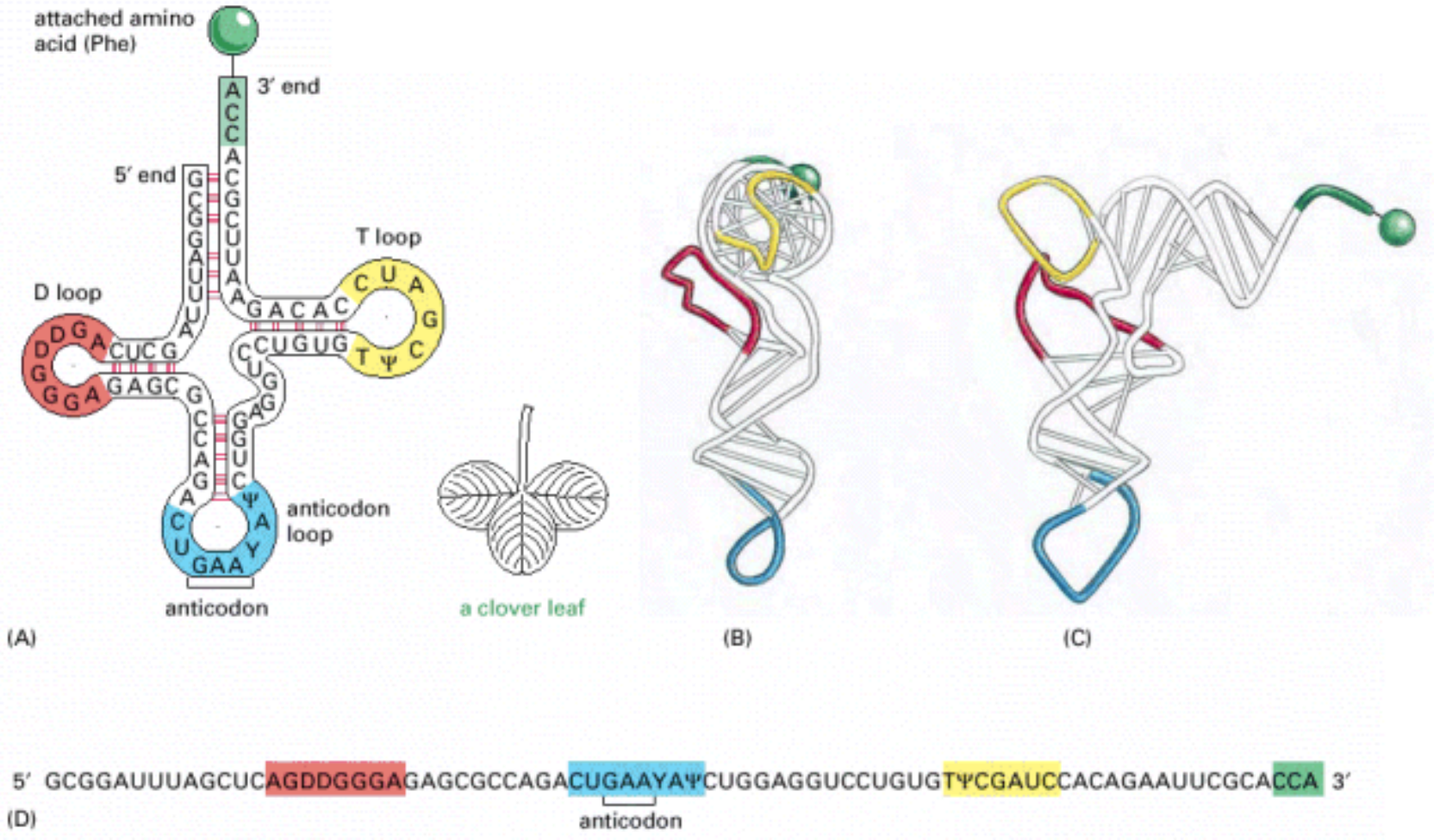
f. pseudonoeud

Secondaire et tertiaire: un exemple…



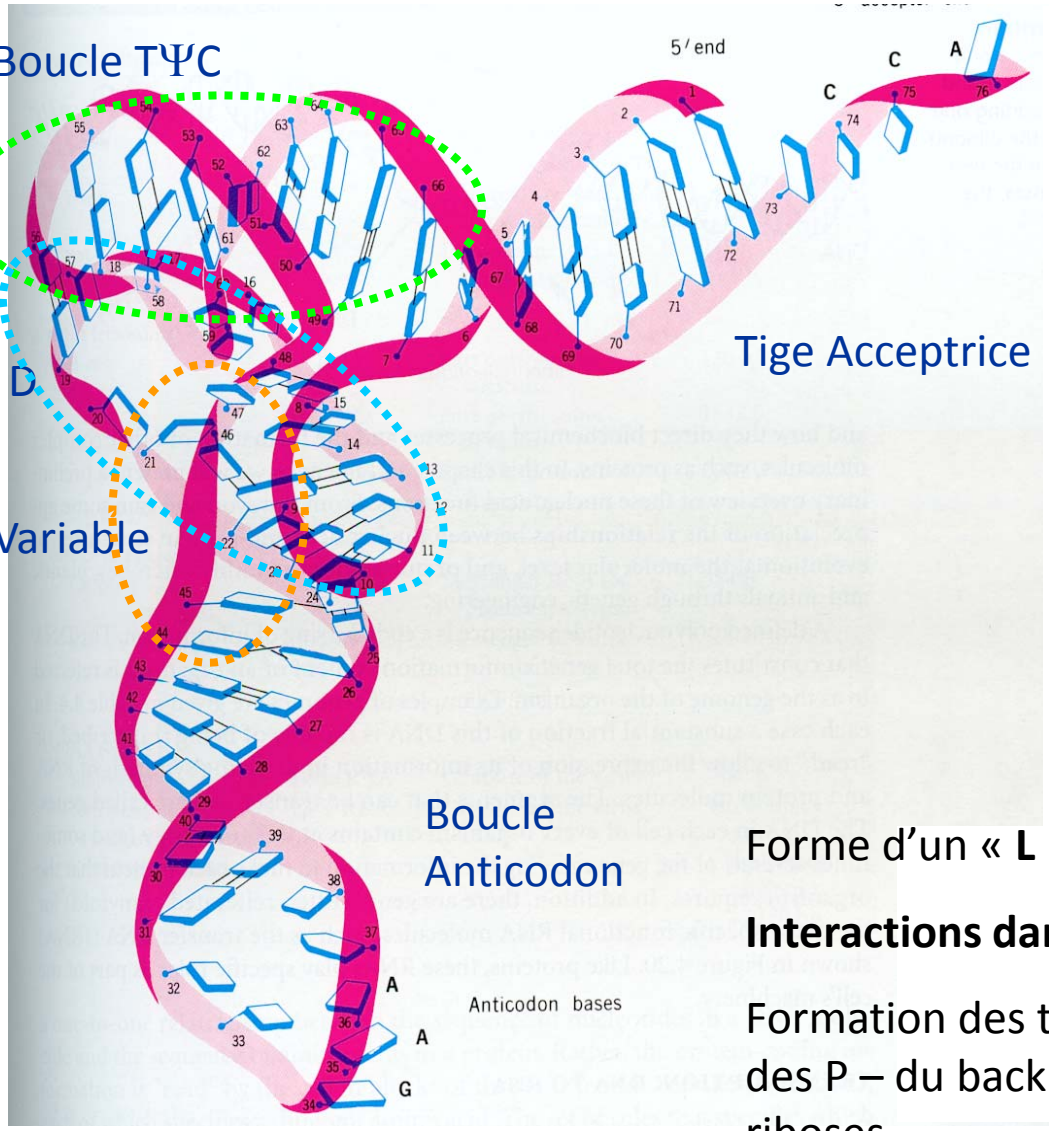
– The 1030-1124 Region of 23S rRNA has seveal tertiary interactions: two canonical lone base pairs (1082:1086 and 1087:1102), a base triple ((1092:1099)1072) and three non-canonical tertiary base pairs (1032:1122, 1039:1116, and 1040:1115)

L'ARNt: du 2D au 3D



Source: Molecular Biology of the Cell (http://www.garlandscience.com/textbooks/0815332181.asp)

L'ARNt



Forme d'un « **L** »

Interactions dans la structure tertiaire:

Formation des triplets; interactions avec des P – du backbone et avec le 2'OH des riboses

ARNt: interactions tertiaires

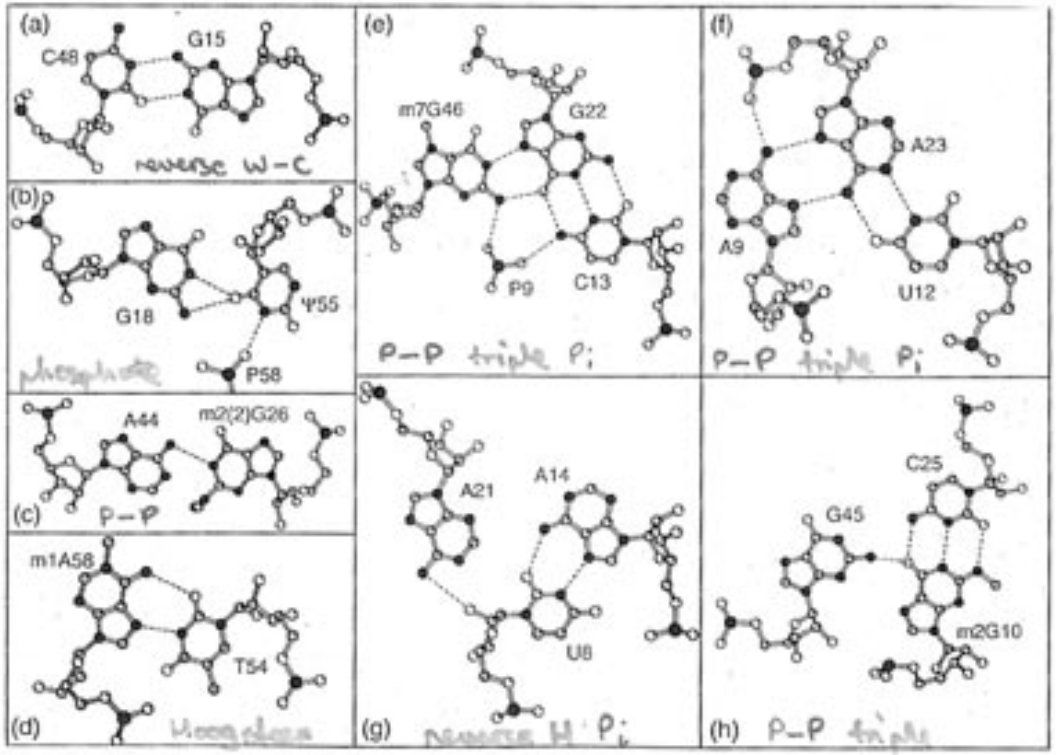
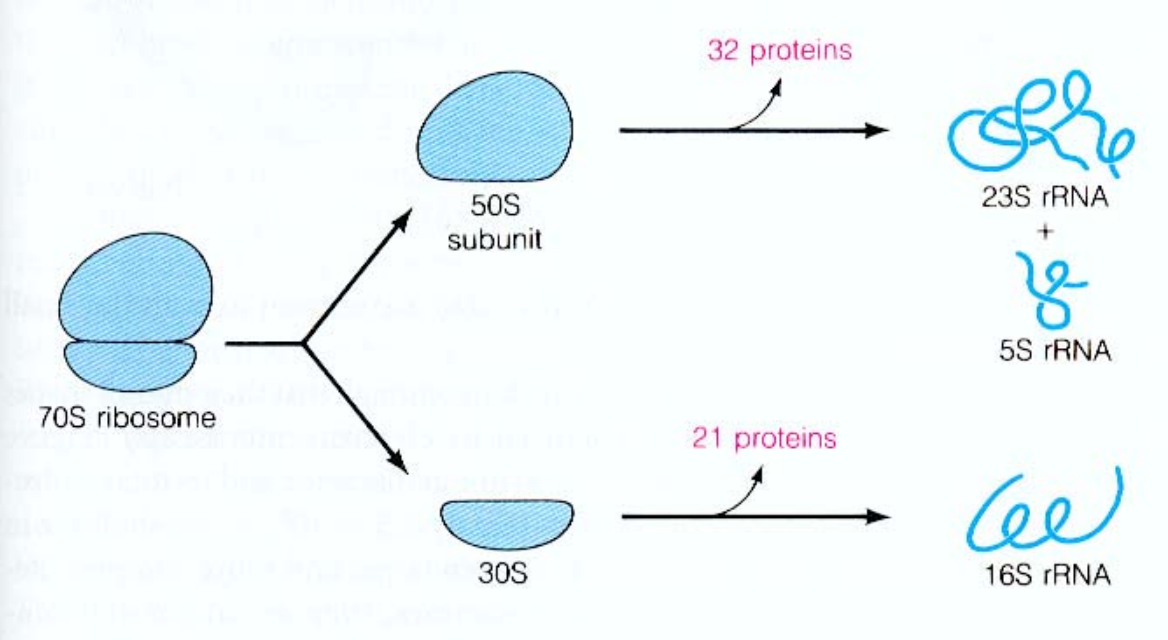


Fig. 2 Tertiary interactions in tRNA^{Phe} (11, 12). The pair G19–C56, a standard Watson–Crick base pair, is not shown. Atoms P–N–C–O are grey-coded in this and subsequent figures with P being the darkest and O the lightest grey.

Image: W. Sanger. Principles of nucleic acid structure. Springer Verlag, 1984.

Le ribosome



Masse totale: 2.6 x 10^3 kDa (100 fois plus que le Lysosyme)

Composition: 1/3 protéine 2/3 nucléotides

Sous-unité 30S : Interaction avec les codons du mRNA et les anticodons du tRNA

Sous-unité 50S : activité peptidyl-transferase et interaction avec le GTP-binding protein.

Structures :

Ribosome 70S de *T. thermophilus* à 7.8 Å (1999)

Sous-unité 30S de *T. thermophilus* à 4.5 Å (1999)

Sous-unité 50S de *H. marsimortui* à 2.4 Å (2000)

"for studies of the structure and function of the ribosome"

Photo: MRC Laboratory of Molecular Biology

Venkatraman Ramakrishnan

1/3 of the prize

United Kingdom

MRC Laboratory of Molecular Biology
Cambridge, United Kingdom

Credits: Michael Marsland/Yale University

Thomas A. Steitz

1/3 of the prize

USA

Yale University
New Haven, CT, USA;
Howard Hughes Medical Institute

Credits: Micheline Pelletier/Corbis

Ada E. Yonath

1/3 of the prize

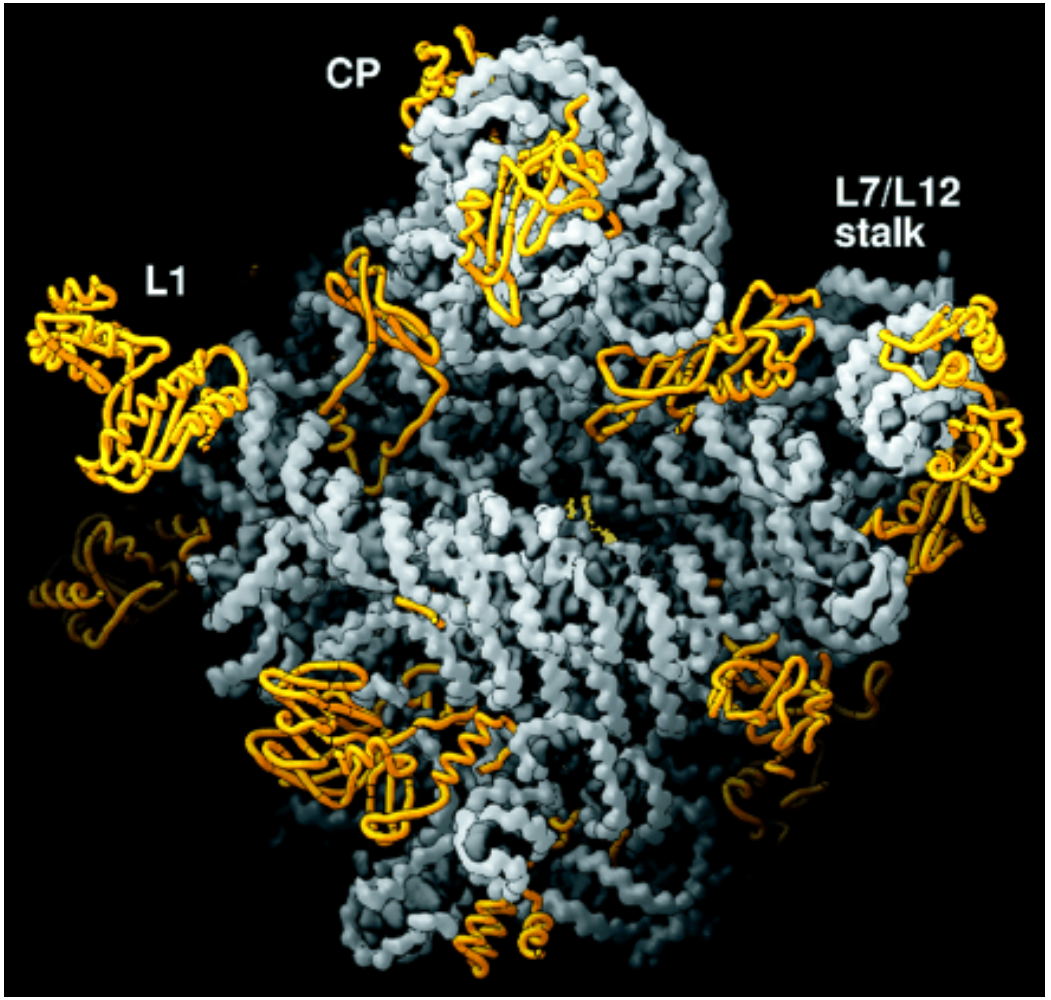
Israel

Weizmann Institute of Science
Rehovot, Israel

La grande sous-unité (ARN 23S)

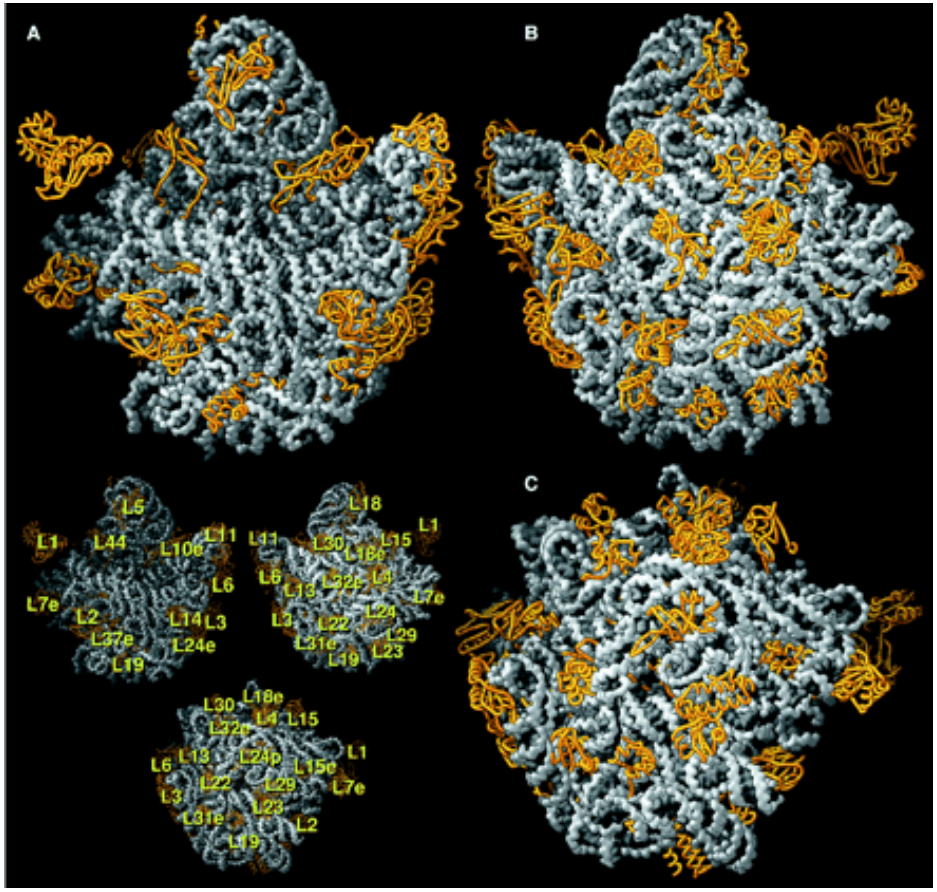
crystal structure of the large ribosomal subunit from Haloarcula marismortui at 2.4 angstrom resolution

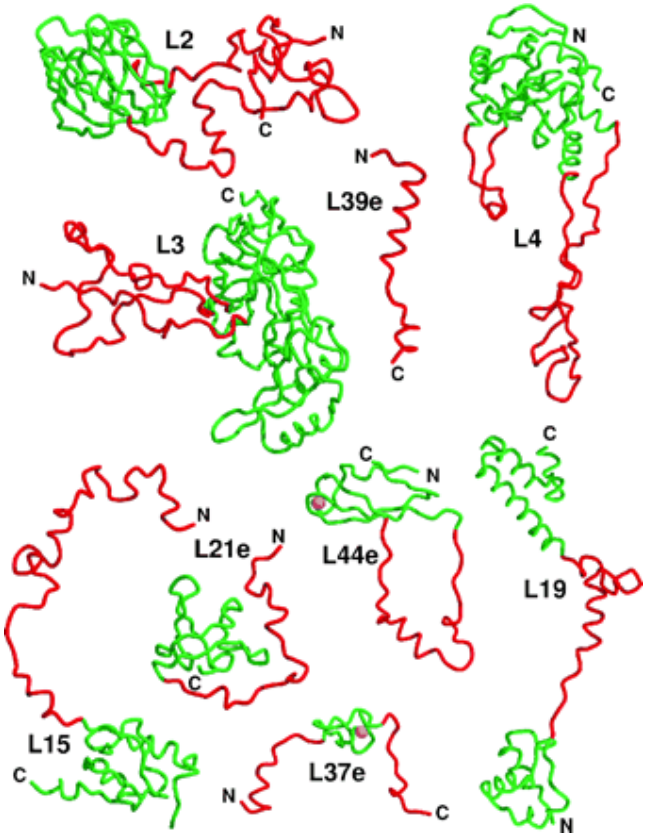
Grande sous Unité: 35 prot + 2 ARN (23S, 5S)



Nenad Ban, Poul Nissen, Jeffrey Hansen, Peter B. Moore, Thomas A. Steitz
Science. 289:878-9, 2000

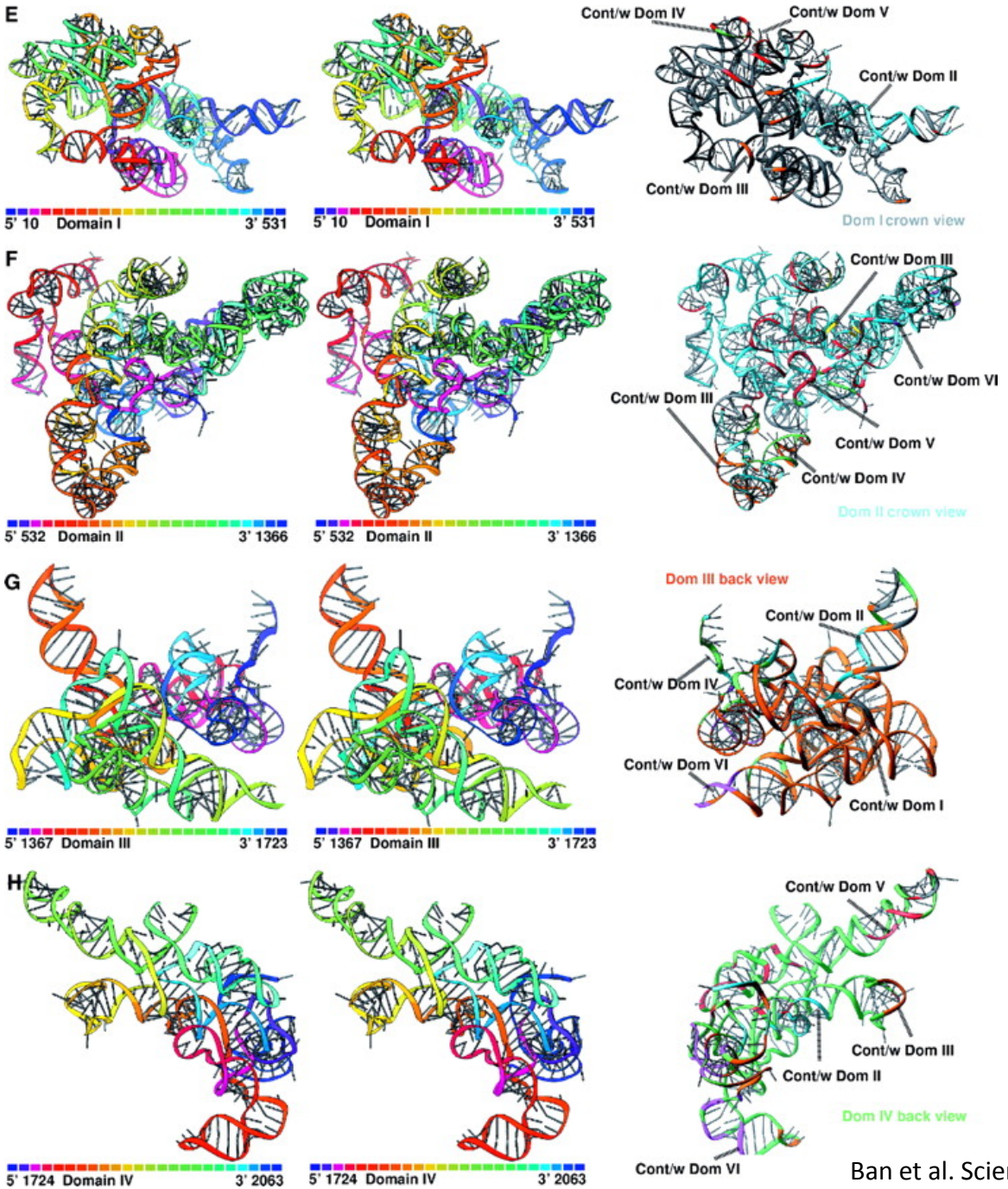
Protéines de la grande sous-Unité





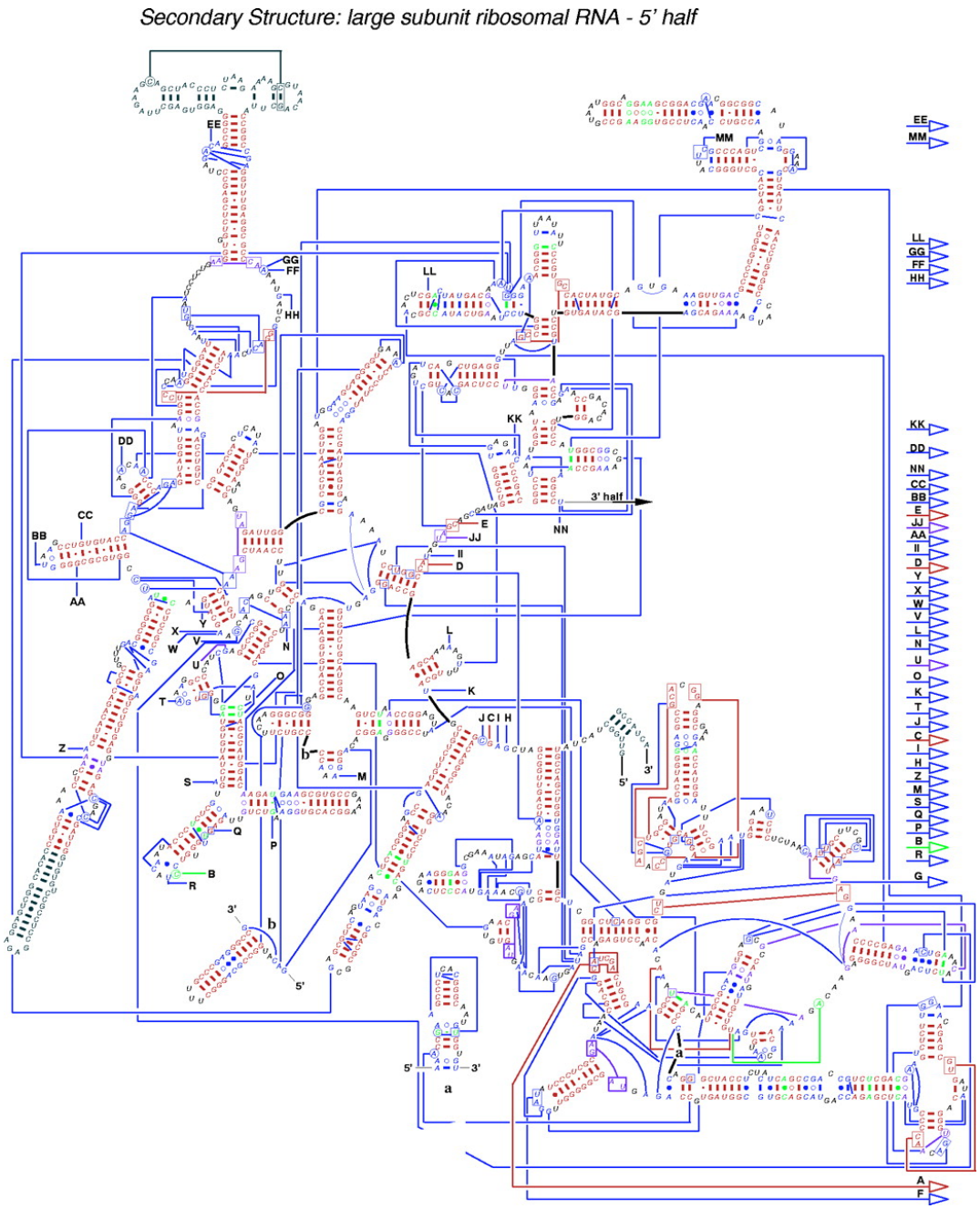
Ban et al. Science. 289:878-9, 2000

Repliement des domaines dominés par les interactions inter-hélices



Ban et al. Science. 289:878-9, 2000

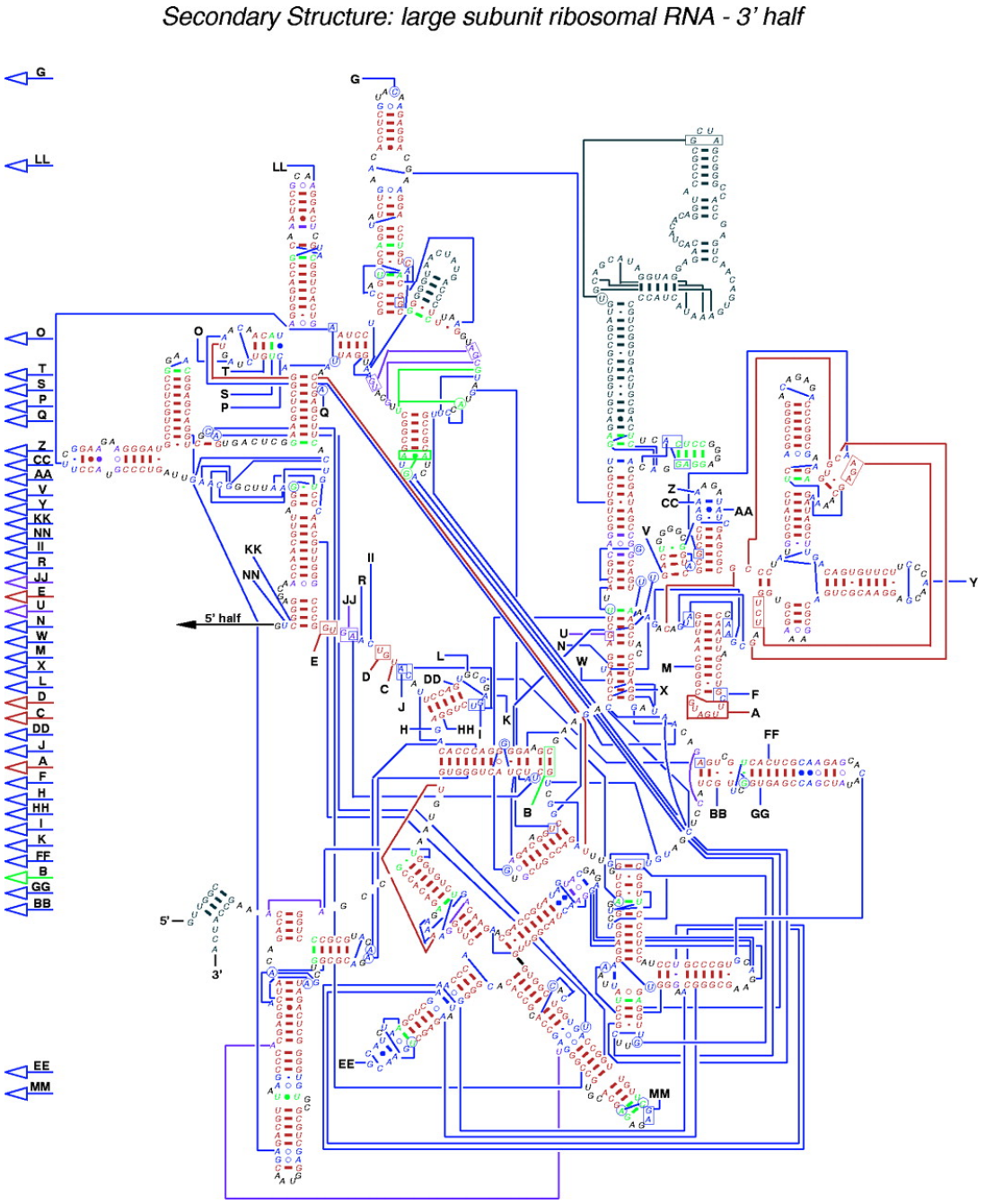
23S: interactions secondaires et tertiaires



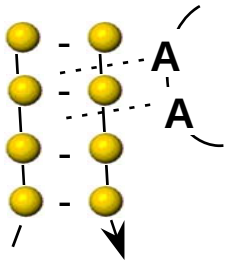
(moitié 5')

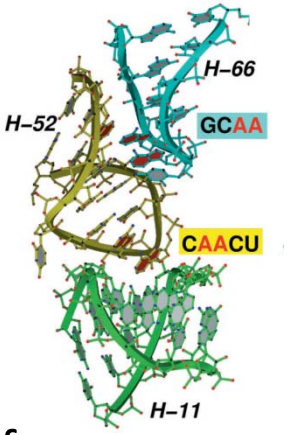
23S: interactions secondaires et tertiaires

(moitié 3’)

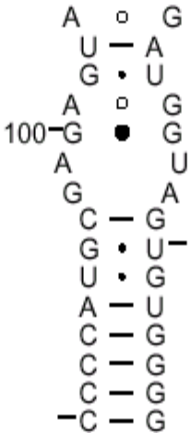


Les motifs ARN

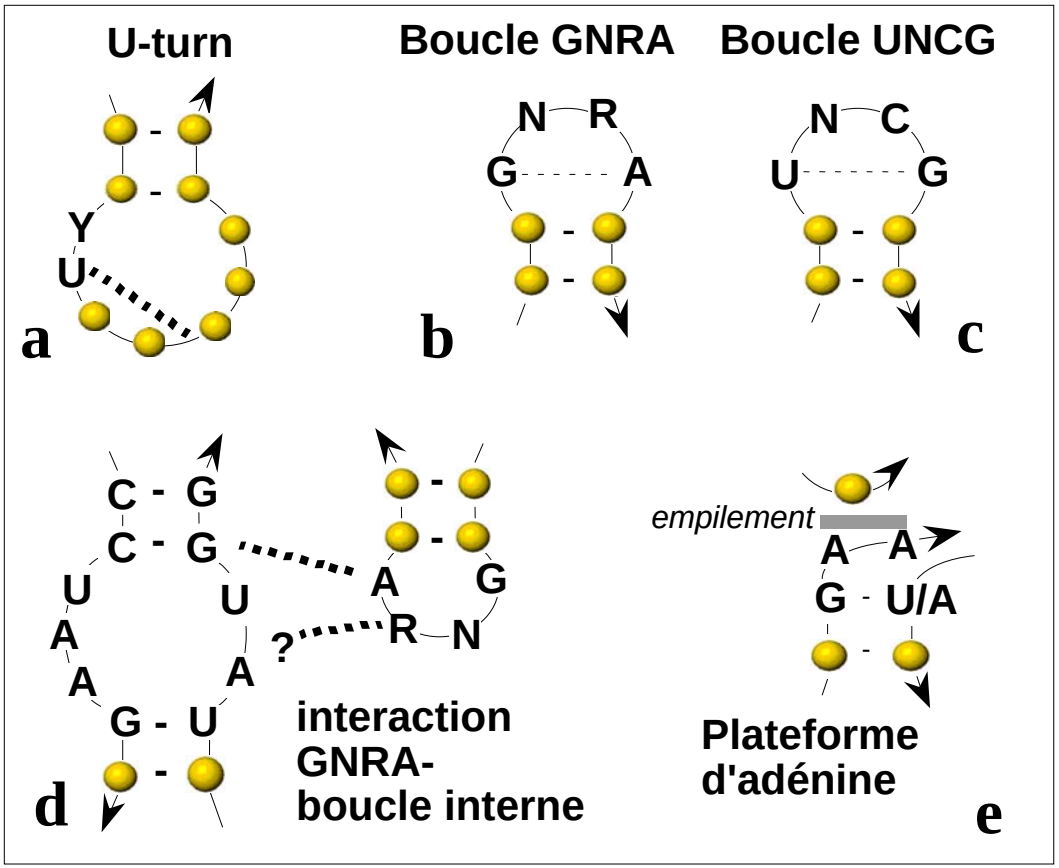




A-minor motif

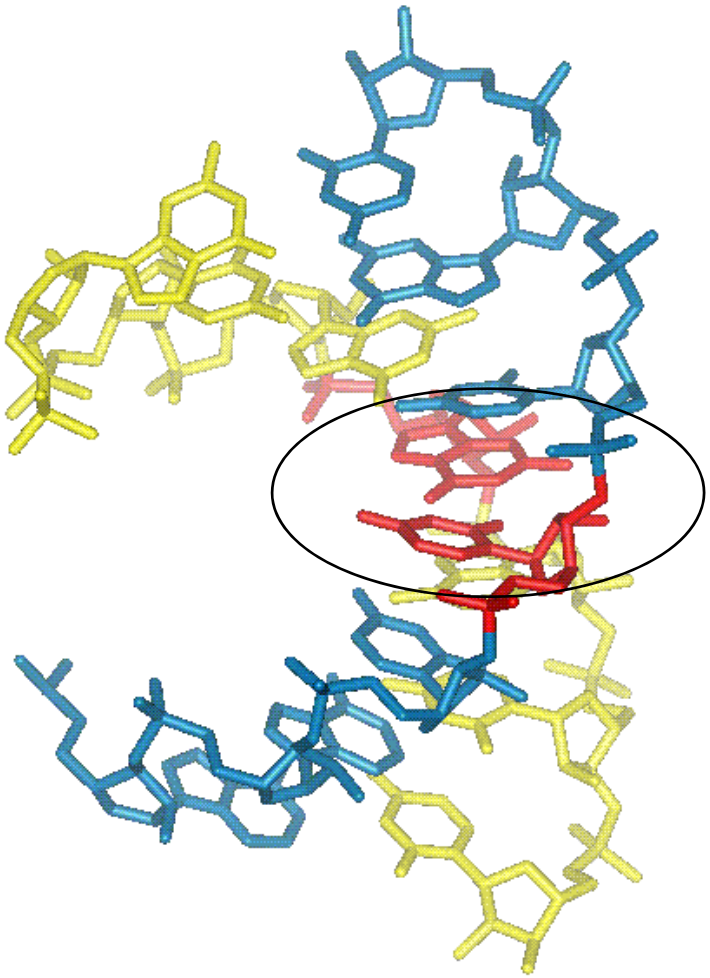


Boucle E

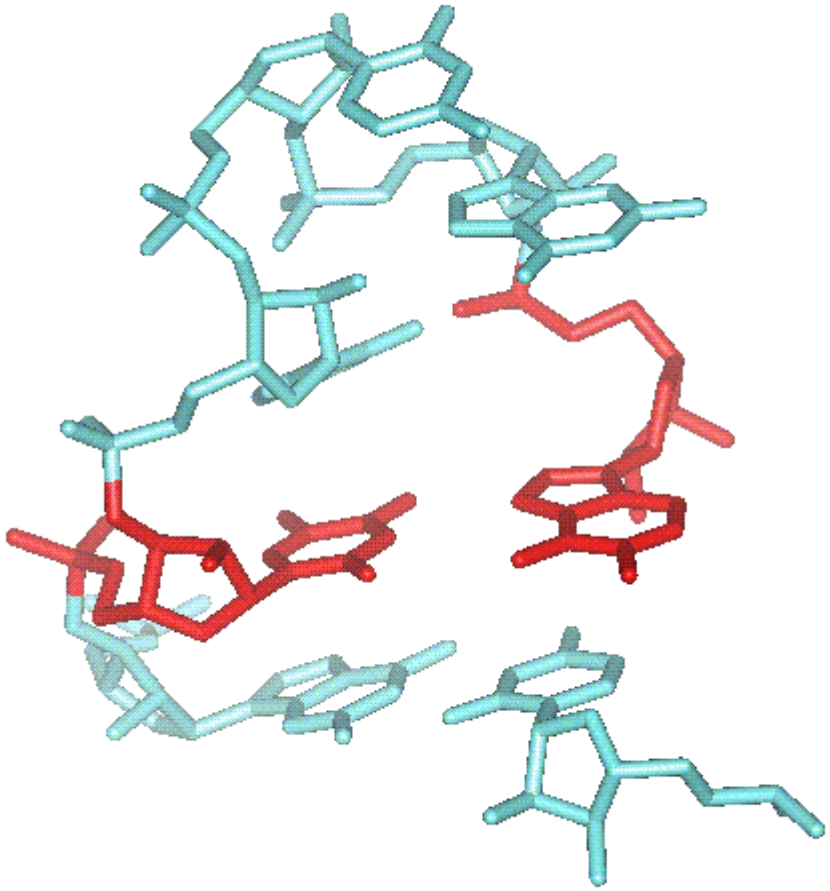


Quelques motifs ARN. Les traits pointillés désignent des interactions tertiaires. Les flèches indiquent le sens 5'->3'. Le point d'interrogation dans le motif *d* indique que le partenaire de cette interaction tertiaire n'est pas connu.

La paire G:U Wobble

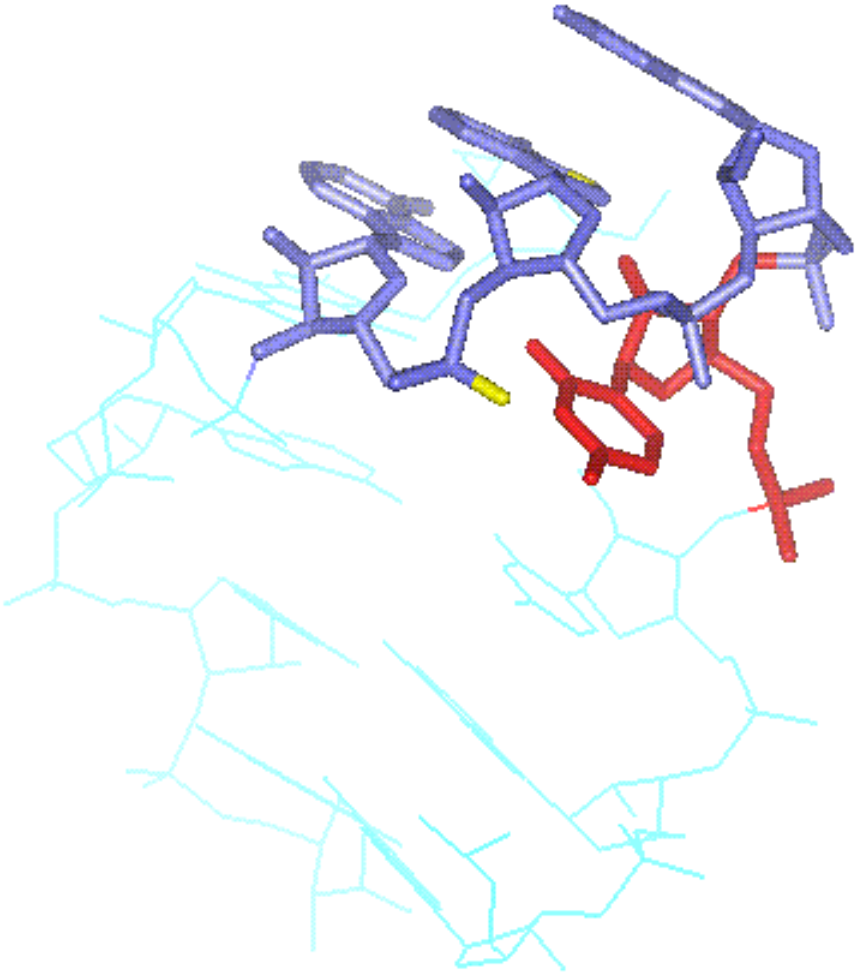


La paire Hoogsteen



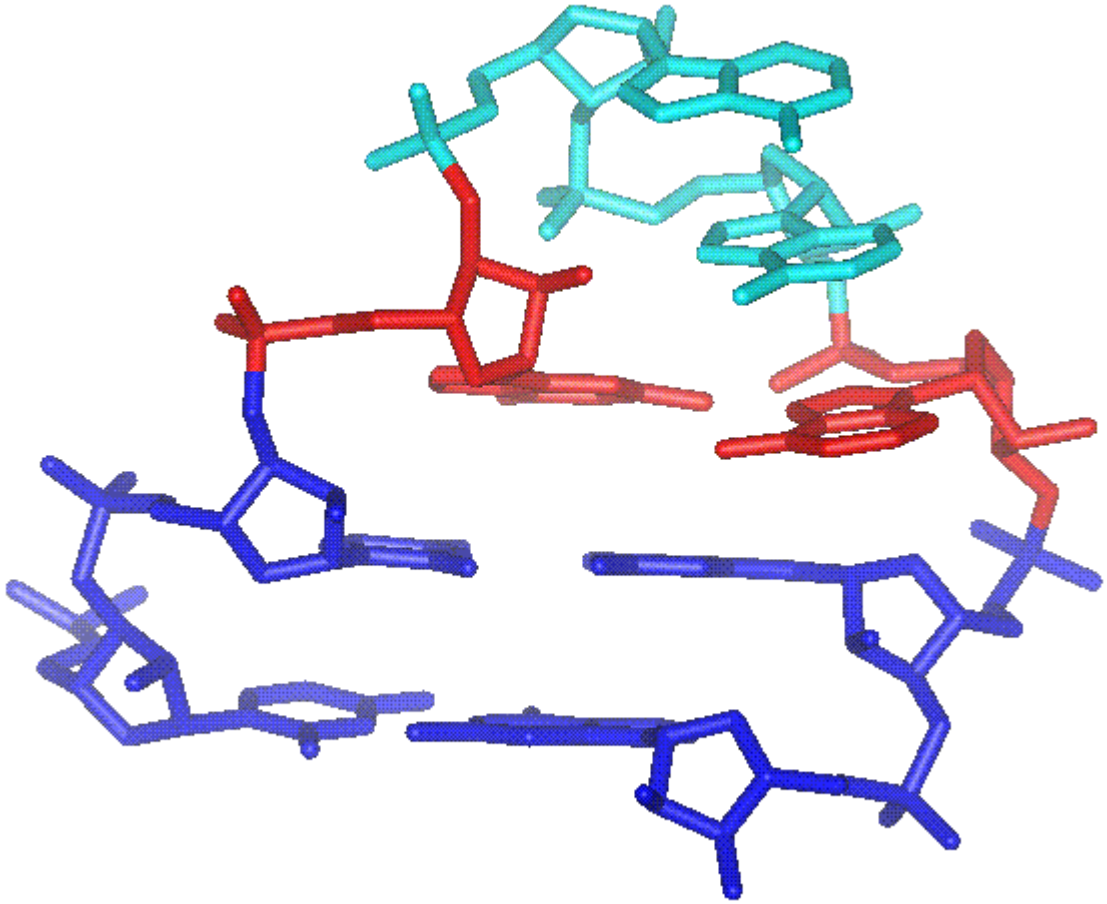
Boucle TTC de tRNA Phe levure

Le U-turn

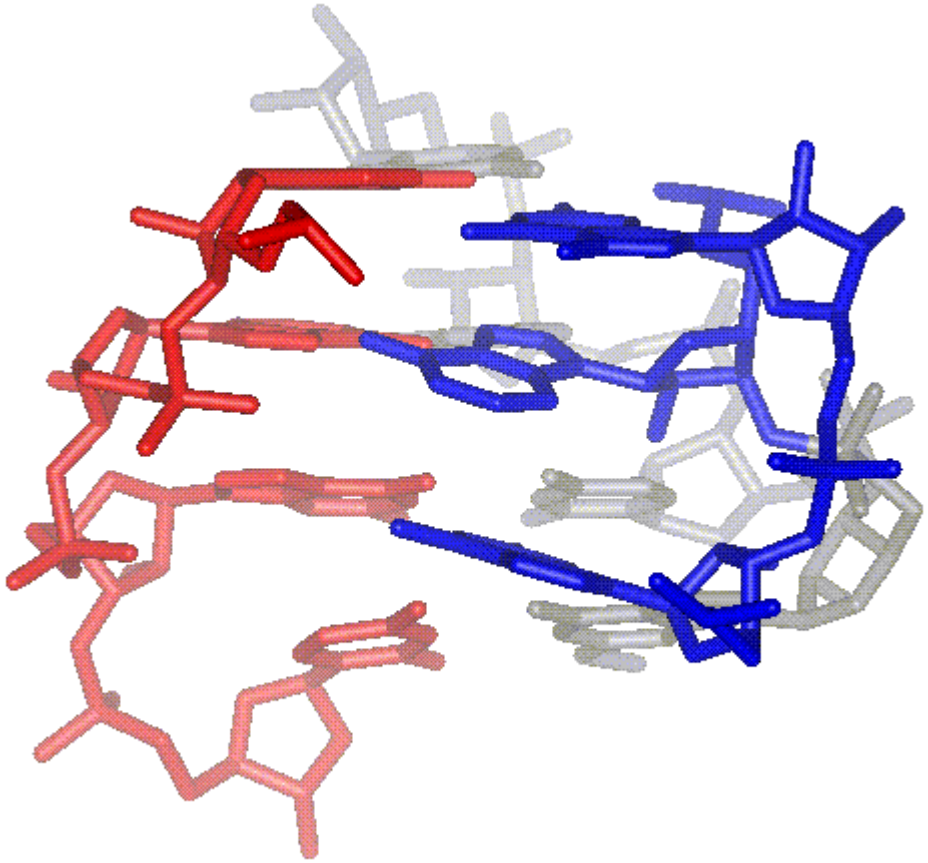


Boucle anticodon de tRNA
Phe levure

La boucle GNRA



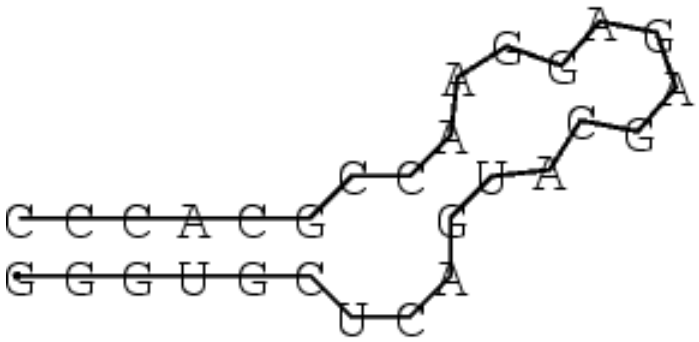
Les triplets de bases

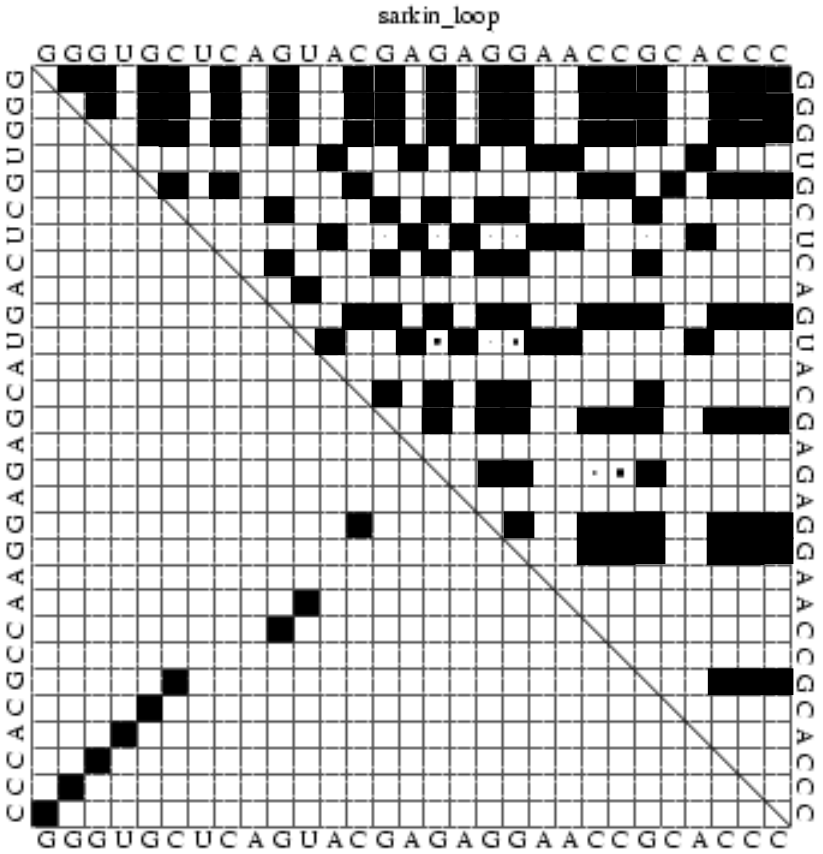


Triple hélice 10-13 + 22-25+ 9+45+46 de tRNA Phe levure

La prédiction des structures secondaires d’ARN

Les Dot-Plots





Algorithme de Zuker et Stiegler

- L'algorithme de programmation dynamique appliqué aux structures secondaires
- recherche les structures de plus basse énergie pour tous les sous-fragments d'une séquence (Zuker and Stiegler, 1981).
- Garantit une solution optimale à l'intérieur du modèle énergétique choisi. N'autorise pas les pseudo-nœuds.

Paramètres d'energie libre d'empilement (stacking free energy) (Turner et al.)

– Dans l'hélice (seulement paires WC et GU):

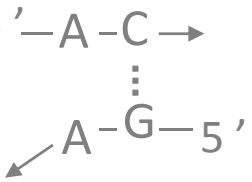
	A	C	G	U	A	C	G	U	A	C	G	U	A	C	G	U
	5' --> 3'				5' --> 3'				5' --> 3'				5' --> 3'			
	UX				UX				UX				UX			
	AY				CY				GY				UY			
	3' <-- 5'				3' <-- 5'				3' <-- 5'				3' <-- 5'			
A	.	.	.	-8.1	.	.	.	.	.	.	.	-4.0	.	.	.	.
C	.	.	-13.3	.	.	.	.	.	.	.	-5.3	.	.	.	.	.
G	.	-10.5	.	-4.0	.	.	.	.	.	-1.8	.	-5.3	.	.	.	.
U	-6.6	.	-3.6	.	.	.	.	.	-3.6	.	-3.6	.	.	.	.	.

– Attention:

Turner's free energies (suite)

– Terminal mismatches:

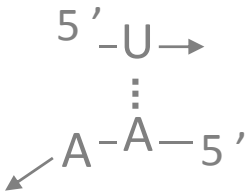
	Y				Y				Y				Y			
	A	C	G	U	A	C	G	U	A	C	G	U	A	C	G	U
	5' --> 3'				5' --> 3'				5' --> 3'				5' --> 3'			
	AX				AX				AX				AX			
	AY				CY				GY				UY			
	3' <-- 5'				3' <-- 5'				3' <-- 5'				3' <-- 5'			
A	.	.	.	-4.0	.	.	.	-4.3	.	.	.	-3.8	-4.3	-6.0	-6.0	-6.0
C	.	.	-5.2	.	.	.	-7.2	.	.	.	-7.1	.	-2.6	-2.4	-2.4	-2.4
G	.	-10.3	.	-7.2	.	-5.2	.	-4.8	.	-9.4	.	-6.6	-3.4	-6.9	-6.9	-6.9
U	-4.3	.	-4.3	.	-2.6	.	-2.6	.	-3.4	.	-3.4	.	-3.3	-3.3	-3.3	-3.3



Turner's free energies (suite)

– Dangling ends:

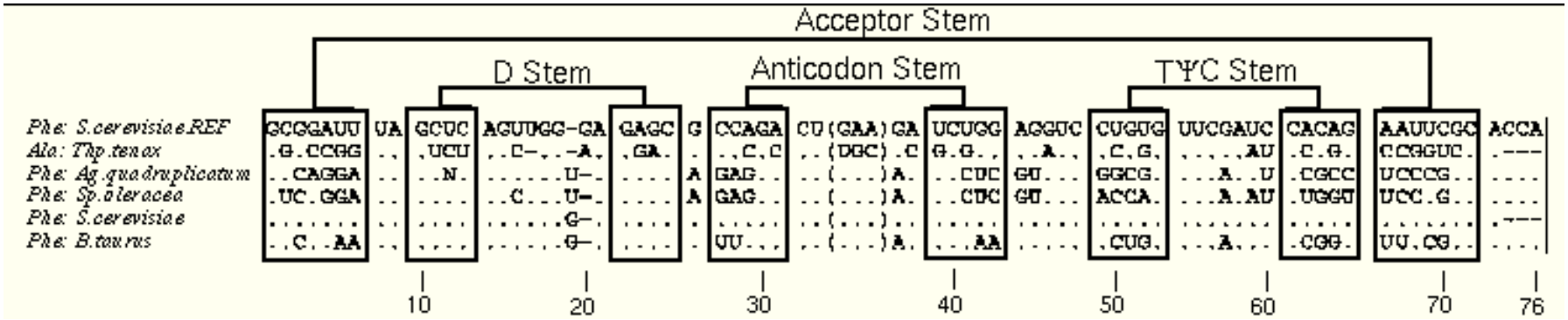
X				X				X				X			
A	C	G	U	A	C	G	U	A	C	G	U	A	C	G	U
	5'	-->	3'		5'	-->	3'		5'	-->	3'		5'	-->	3'
	AX				AX				AX				AX		
	A				C				G				U		
	3'	<--	5'		3'	<--	5'		3'	<--	5'		3'	<--	5'
.	.	.	.	.	.	.	.	.	.	.	.	-4.9	-0.9	-5.5	-2.3



Limites de la prédiction par minimisation d'énergie

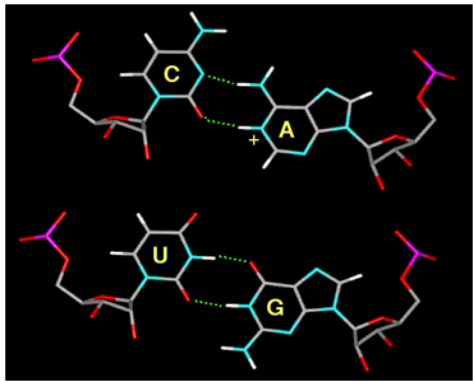
- Influence de la structure tertiaire
  - Triplets de base
  - Interactions sucre-base / sucre-phosphate
  - Méconnaissance de la stabilité des paires non-canoniques
- Même si les paramètres energétiques des paires étaient parfaits, on ne pourrait pas prédire parfaitement la structure secondaire, faute de contexte

Analyse comparative



- L'analyse comparative est la façon la plus fiable de déterminer les structures secondaires
- Elle repose sur la détection de covariations
- Elle nécessite plusieurs séquences homologues alignées

- Fonctionne même pour les paires non Watson-Crick!



Mesures de la covariation

- Table de contingence:

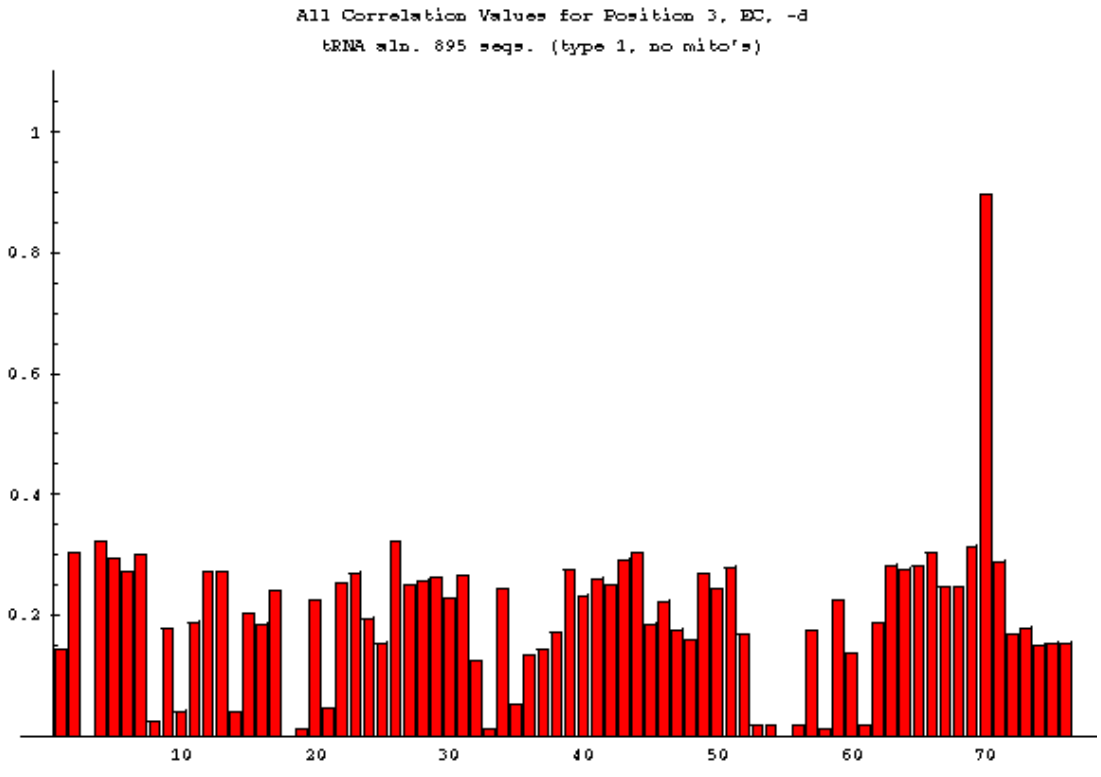
```
table p normal 53 61

            a  c  g  u  -
   53, 61|  0 97  0  3  0 61
   ------+------------------
   a  3  |  0  0  0  3  0
   c  0  |  0  0  0  0  0
   g 97  |  0 97  0  0  0
   u  0  |  0  0  0  0  0
   -  0  |  0  0  0  0  0
   53

   gc=( 29, 28.03, 96.7%)
   au=(  1,  0.03,  3.3%)
```

- Tests:
  - Chi 2
  - Information Mutuelle
  - évènements phylogénétiques

$$\chi^2 = \sum_{M,N} \frac{[no(Mi,Nj) - ne(Mi,Nj)]^2}{ne(Mi,Nj)}$$

Exemple: covariations d'un nucléotide avec tous les autres

- Position 1 du tRNA contre toutes les autres positions:

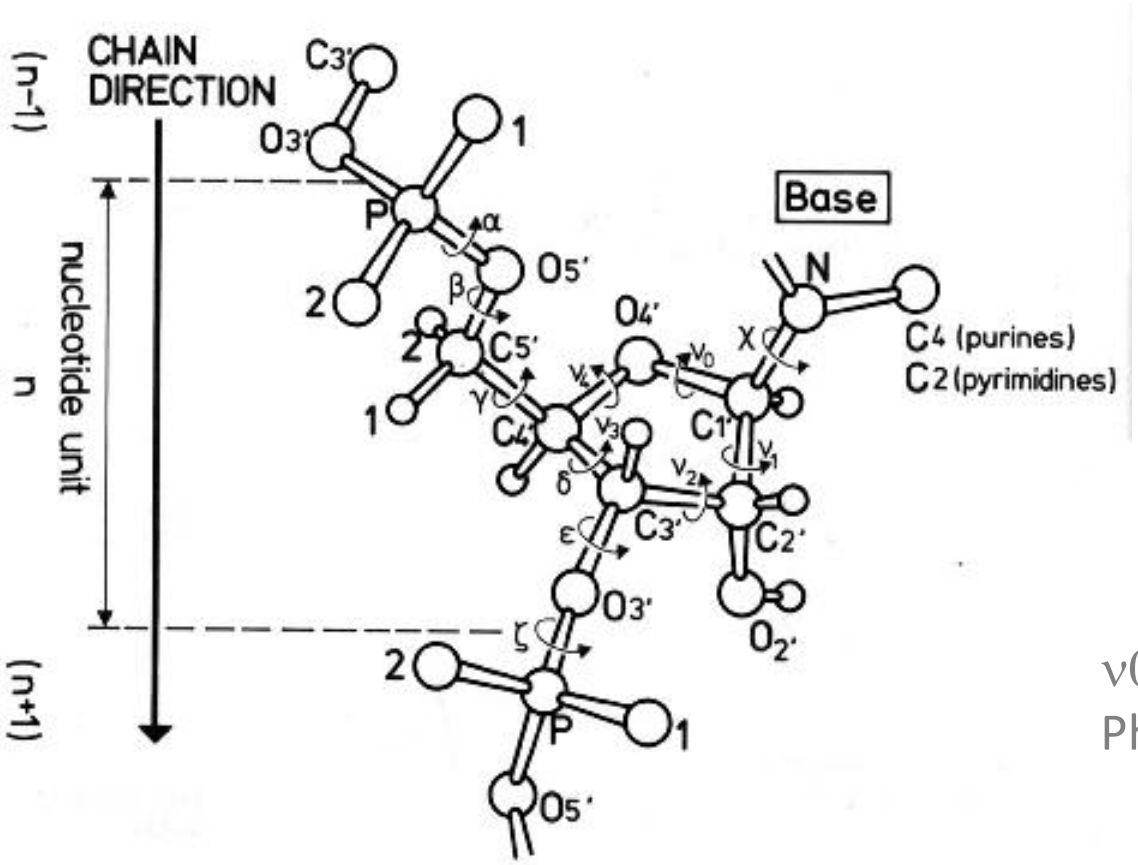


La modélisation 3D des ARN

De 2D à 3D

- Si l'on connaît la bonne structure 2D, peut-on prédire la structure 3D?
  - NON! On ne peut prédire que les hélices
  - Il reste trop d'inconnues dans les régions simple-brin
  - Une modélisation 3D est nécessaire

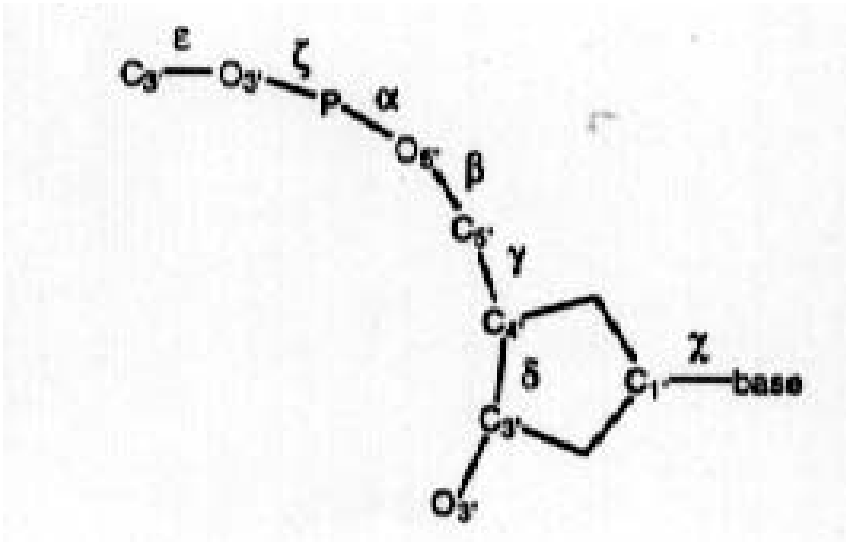
Les degrés de liberté du nucléotide



ν0 à ν4 résumés par:
Phase+amplitude

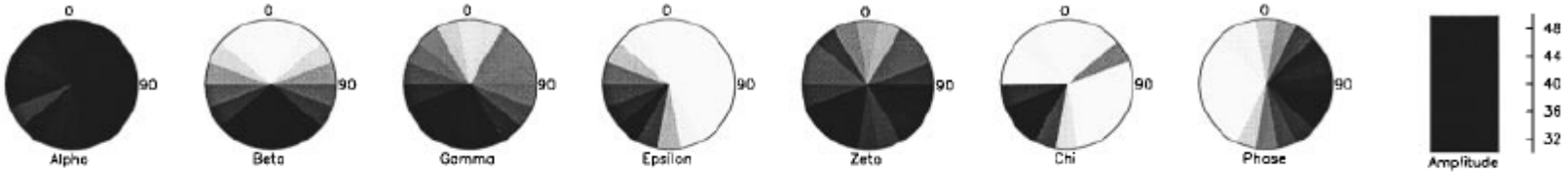
Image: W. Sanger. Principles of nucleic acid structure. Springer Verlag, 1984.

Contraintes sur les angles de torsion

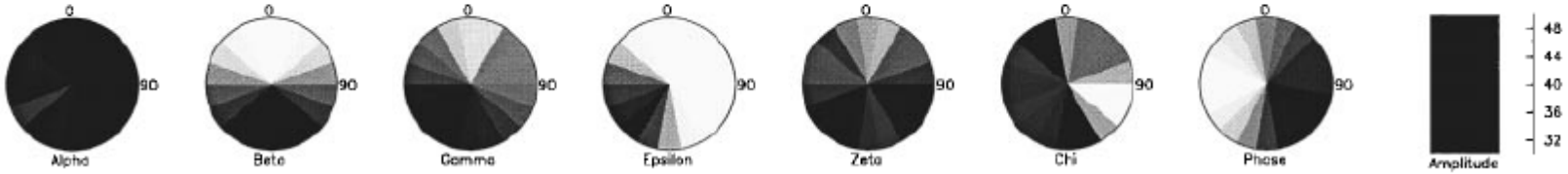


Roues conformationnelles:

CDP



GDP



Murty et al. J Mol Biol 1999 vol 291

Mécanique moléculaire

- Champ de force (Charmm, Amber etc.)
- Minimisation, dynamique, recuit simulé
- Comme pour les protéines, la modélisation *de novo* n'est pas satisfaisante
- Ces méthodes sont utilisées principalement pour le raffinement
  - Cristallographie, RMN
  - Modélisation par homologie avec ARN connus

Mécanique moléculaire

- Champ de force (Charmm, Amber etc.)
- Minimisation, dynamique, recuit simulé
- Comme pour les protéines, la modélisation *de novo* n'est pas satisfaisante
- Ces méthodes sont utilisées principalement pour le raffinement
  - Cristallographie, RMN
  - Modélisation par homologie avec ARN connus

Assembly Approach

Assemble RNA components using construction operators to position and orient them in 3-D space.

- *MANIP*, an interactive system,
- *MC-Sym*, an automated procedure

Credit: François Major

Modélisation par contraintes

- Principes du programme MC-SYM

MC-Sym (Macromolecular Conformations by SYMbolic programming) builds RNA tertiary structure models from syntactic descriptions of RNAs.

Implements a CSP solver with backtracking, and thus generates models that are consistent with constraints.

Credit: François Major

The Constraint Satisfaction Problem (CSP)

The CSP can be described by three finite sets:

1. the **variables** $V = \{v_1, v_2, \ldots, v_n\}$
2. the **domains** $D = \{d_1, d_2, \ldots, d_n\}$ and
3. the **constraints** $C = \{c_1, c_2, \ldots, c_m\}$.

A variable v_i is assigned values from domain $d_i = \{d_{i,1}, d_{i,2}, \ldots, d_{i,|di|}\}$.

Search Space

The search space of a CSP is the Cartesian product of all d_i :

$$\prod_{i=1}^{n} |d_i|$$

The size of a CSP search space is exponential in the number of domain sizes. The solutions of a CSP are found by exploring the variable assignments of its search space and verifying if they satisfy the constraints. **Backtracking** is the classical search algorithm to solve a CSP deterministically and exhaustively.

Credit: François Major

Backtracking

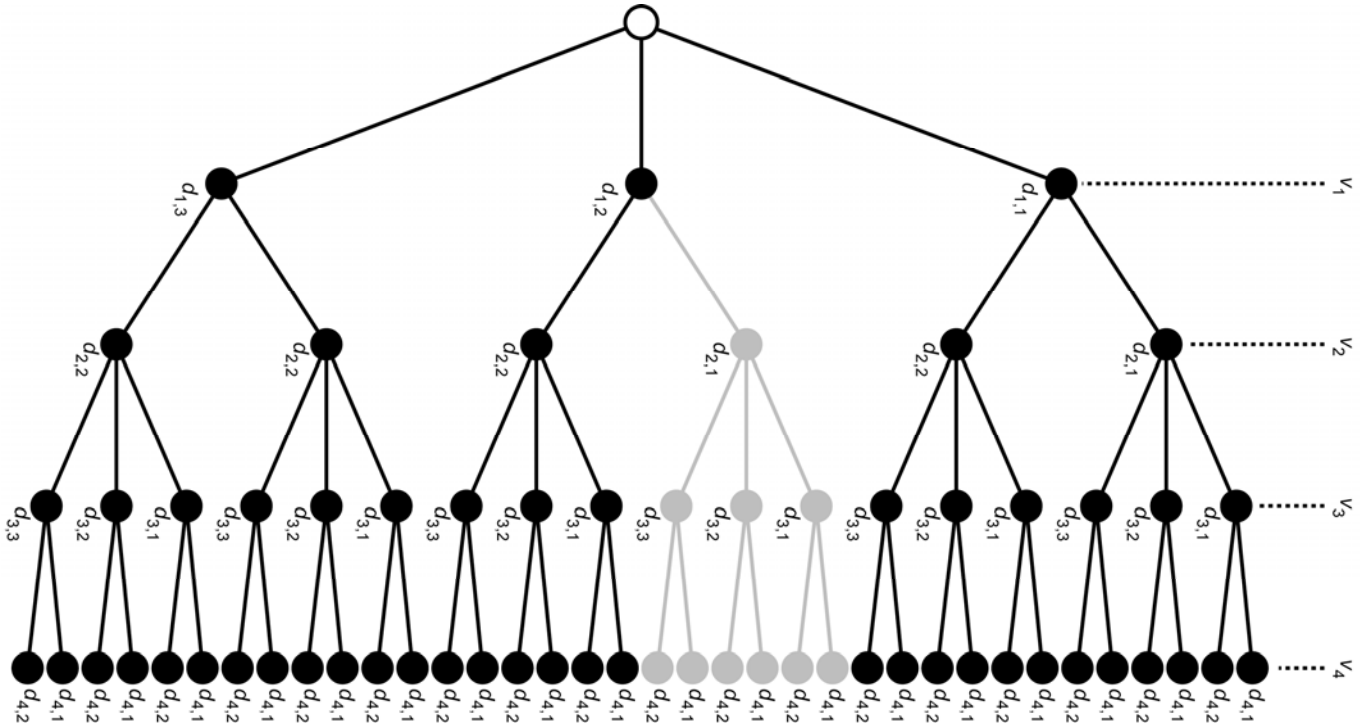
–Variables are assigned values systematically, one at a time and to the next available value from its associated domain.

– When all the values of a domain have been tried, the domain is reset and backtracking moves to the previous variable, assigning its next value, before continuing.

–The search finishes when the domain of the first variable is reset, indicating that all possible assignments have been tried.

Credit: François Major

Search Tree & Pruning



Backtracking search tree. The search space is defined by $V = \{v_1, v_2, v_3, v_4\}$ and $D = \{d_1, d_2, d_3, d_4\}$ where $d_1 = \{d_{1,1}, d_{1,2}, d_{1,3}\}$, $d_2 = \{d_{2,1}, d_{2,2}\}$, $d_3 = \{d_{3,1}, d_{3,2}, d_{3,3}\}$ and $d_4 = \{d_{4,1}, d_{4,2}\}$. Any path from the root (empty circle) to a leaf assigns each variable to a value from its respective domain. A verification of the constraint $\{(v_1, v_2) \in d_1 \times d_2 \mid (v_1, v_2) \neq (d_{1,2}, d_{2,1})\}$ prunes the sub-tree in grey, as soon as the backtracking assigns v_2 to $d_{2,1}$, and then the search jumps to the next assignment for v_2.

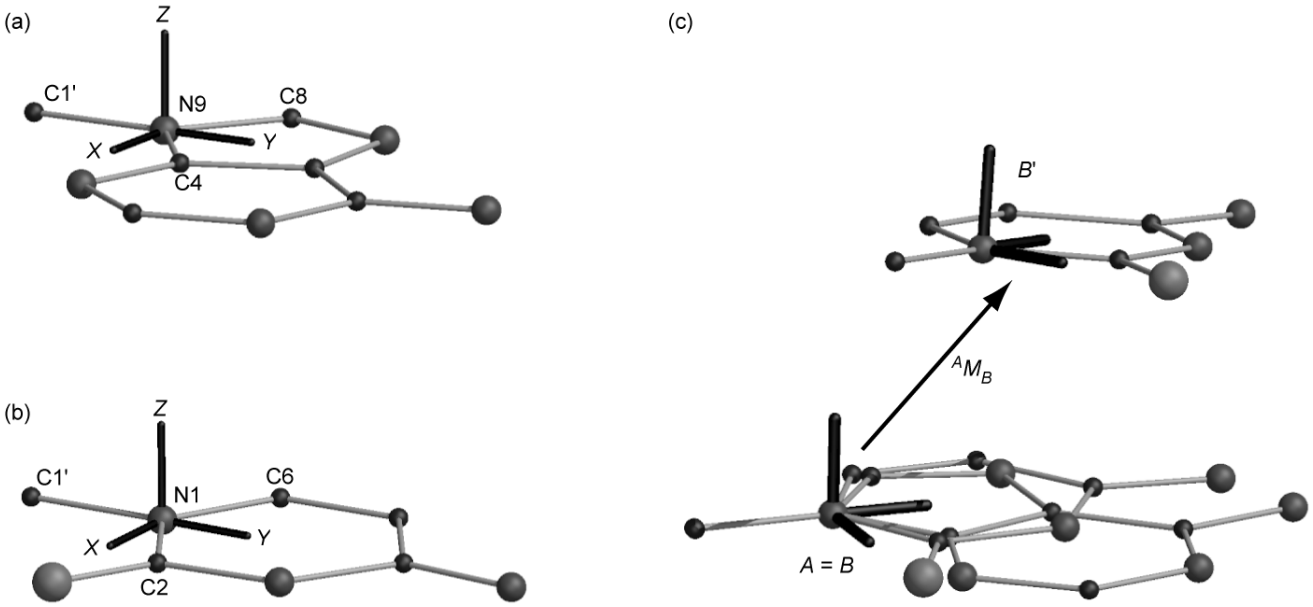
Credit: François Major

MC-SYM conformational space

- For each nucleotide, combine:
  - Predefined conformations (ex: 5 confs for a A-helix nt; or 30 confs for an unknown nt).
  - A set of transformation matrices, corresponding to known nt-to-nt transformations (in helices, base-pairs, etc.)
- This defines the set possible values for conformations and positions of one nt
- No need to sample torsion angles

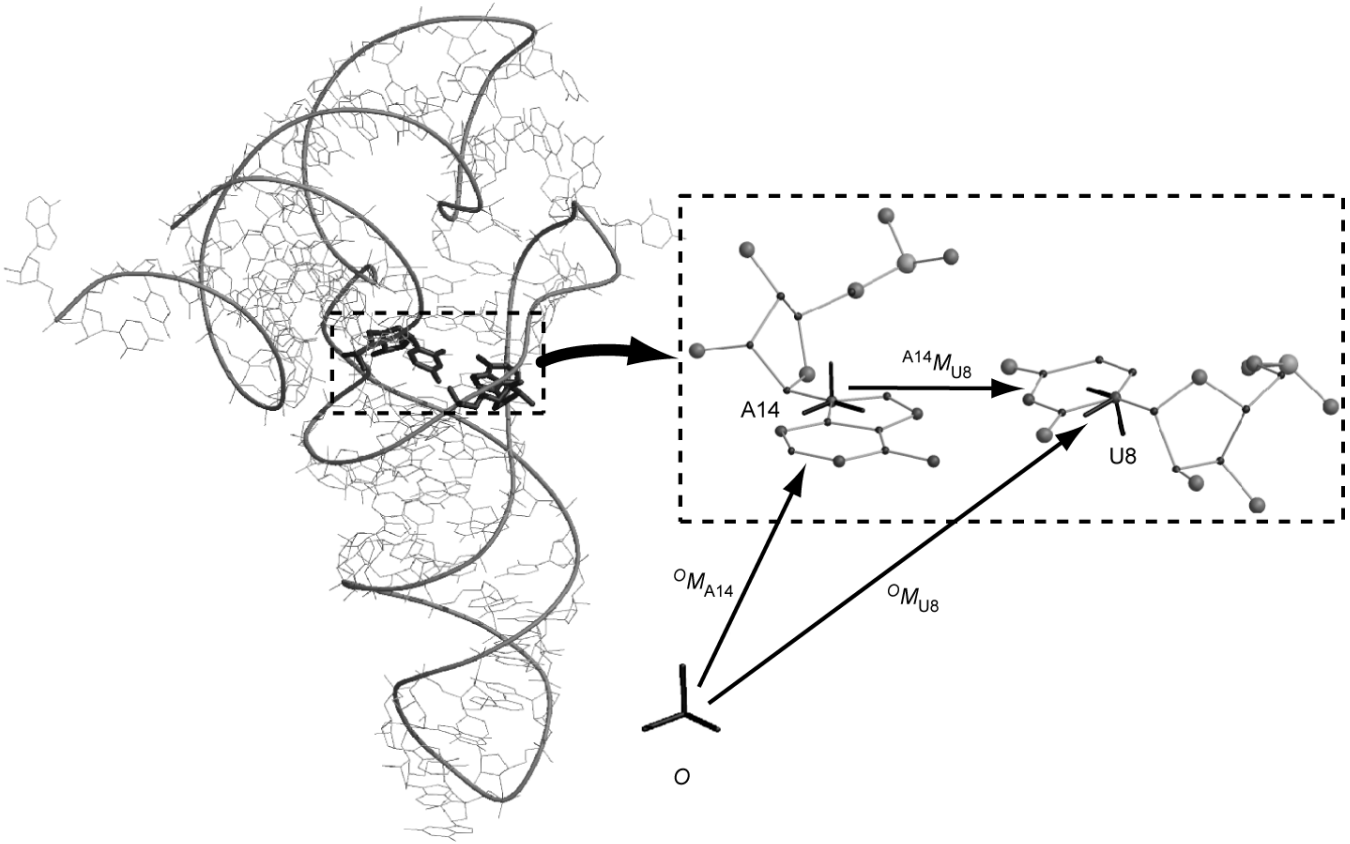
Linear Transformation Matrices

The linear transformation matrices combine nucleotide rotations and translations in 3-D coordinates, which represent the spatial relation between two stacked or paired bases.



Credit: François Major

Transformation Matrices



Toutes les matrices de transformations observées entre nucléotides en contact dans la banque pdb sont sauvegardées dans la banque de données de MC-SYM.

Credit: François Major

MC-Sym conformation and transformation Database

The *MC-Sym* database contains near 3000 nucleotide conformations and near 20 000 base interactions, hence the domain size argument next to each conformation and interaction.

It is the task of the modeler to assign domain sizes so that the conformational space of a given tertiary structure is correctly addressed; not too small to miss valid models and not too large to avoid prohibitive search space sizes.

Credit: François Major

RNA Constraints

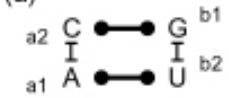
Two types of constraints need to be verified at each variable assignment: the atomic clashes and the O3'-P covalent bond distances. The scope of the atomic clash constraints is all nucleotide pairs, and they are needed to insure that any pair of atoms from both nucleotides is not overlapping. A threshold inter-atomic distance, typically 1 Å, implements the steric clash constraints.

Credit: François Major

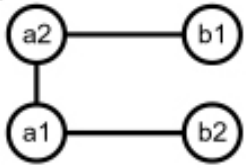
MC-SYM script Example: bp tandem

Script MC-SYM:

(a)



(b)



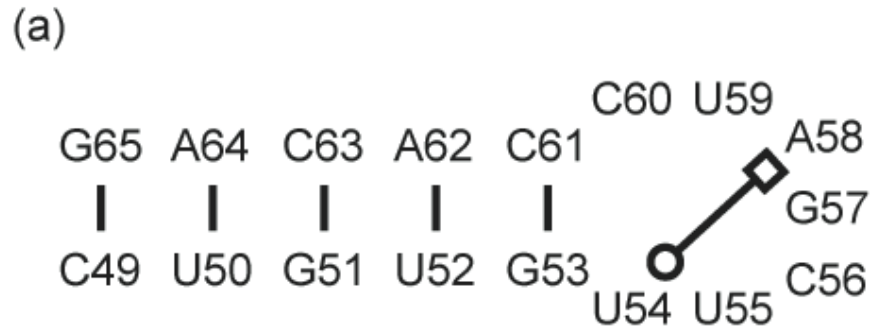
(c)

```
sequence ( r  a1  AC )
sequence ( r  b1  GU )

residue (
  a1  { C3'_endo && anti }  10
  a2  { C3'_endo && anti }  10
  b1  { C3'_endo && anti }  10
  b2  { C3'_endo && anti }  10
)

connect (
  a1 a2  { stack }  5
)

pair (
  a1 b2  { W / W && cis } 7
  a2 b1  { W / W && cis } 7
)

myRNA = backtrack (
  ( a1 b2 )
  ( a1 a2 b1)
)

[...]
```

(a) RNA graph. (b) Spanning tree. (c) *MC-Sym* input. The "stack" keyword is used as a wildcard matching any of the four stacking types: upward, downward, inward or outward. the **`sequence`** section defines the two strands and introduces a global numbering system for the nucleotides. The **`residue`** section defines the nucleotide conformations and sampling sizes (here, ten different C3'-*endo anti* conformations). The nucleotide interactions are defined in the **`connect`** and **`pair`** sections, specified using the *LW+* nomenclature. **`Connect`** is used for adjacent nucleotides in the sequence. In the example, one of the two bs interactions is included in the spanning tree (five stacking transformations). The **`pair`** section is used for the two Watson-Crick bps (here, seven different Watson-Crick transformations).
The **`backtrack`** section defines the order of the nucleotides. Here a1 is selected as the global referential, then b2 Watson-Crick to a1, a2 stacked with a1, and finally b1 Watson-Crick to a2.

Credit: François Major

Yeast tRNA-Phe T-Stem-Loop

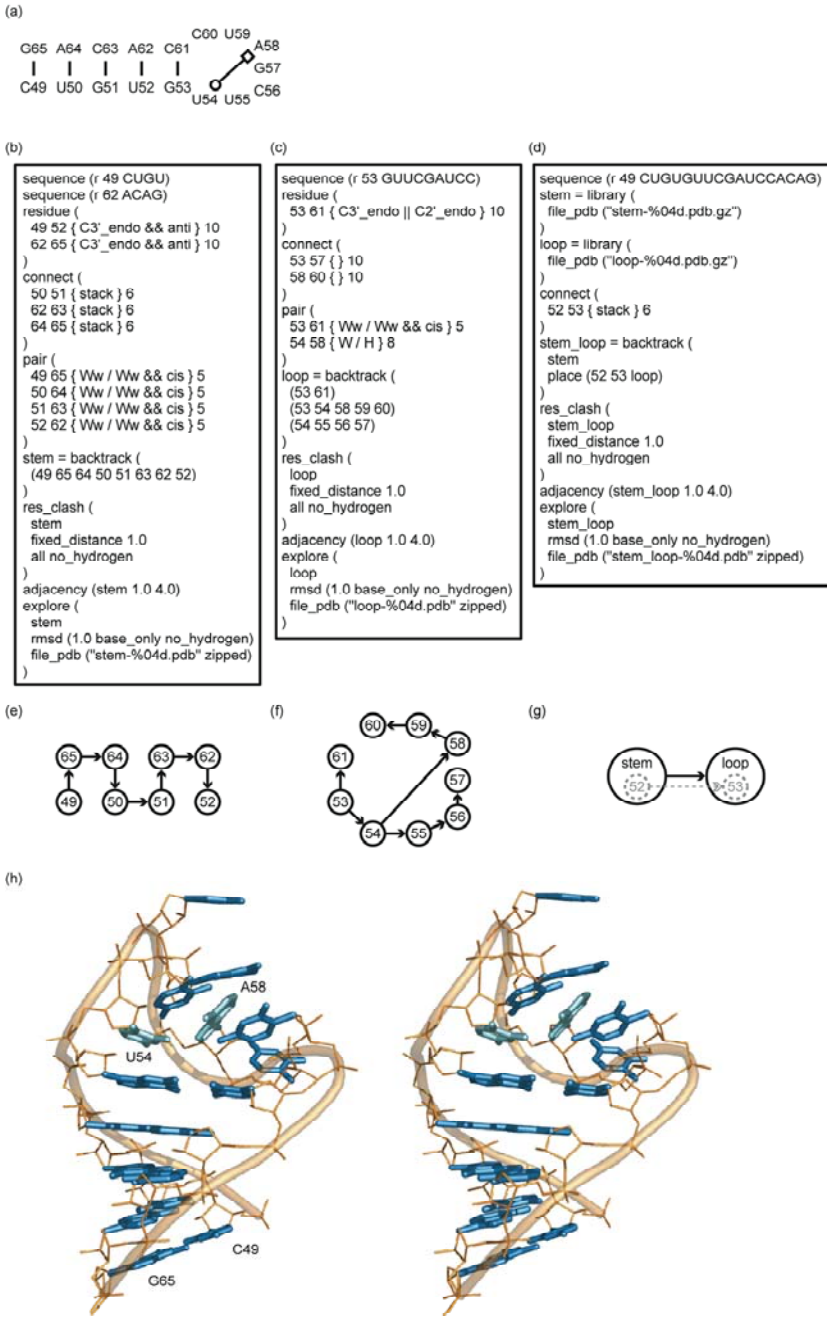


The stem and the loop are modeled independently.

Three input descriptors are needed:

1. first four bps of the stem;
2. hairpin loop, closed by the last bp of the stem
3. Merge of the two others.

Credit: François Major



(a) Secondary structure. (b) *MC-Sym* input for the stem fragment. (c) *MC-Sym* input for the loop fragment. (d) *MC-Sym* input for merging both fragments. (e) Spanning tree of (b). (f) Spanning tree of (c). (g) Spanning tree of (d). (h) Stereoview of one model generated by *MC-Sym*. The bases are shown in blue; the U54-A58 *W/H* bp in lighter blue. The backbone is shown in yellow. The thread follows the phosphodiester chain. Hydrogen atoms are not shown.

Credit: François Major