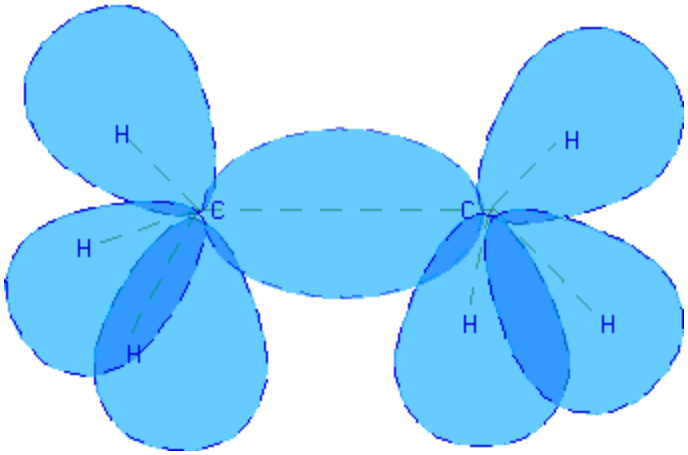
Vendredi 27 février 2004

Démo 7 – Introduction à la prédiction et la recherche de structures

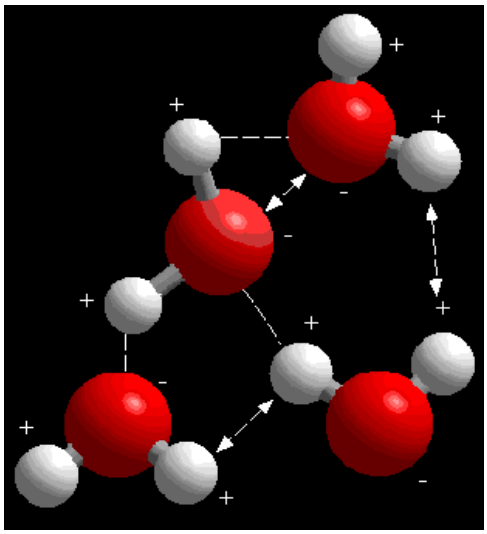
Jean-Eudes Duchesne

1. **Liens chimiques**.

Dans un lien covalent, deux atomes partagent un électron (les électrons aiment bien faire des paires !). Ainsi, un lien covalent va créer un nuage électrique entre deux atomes :

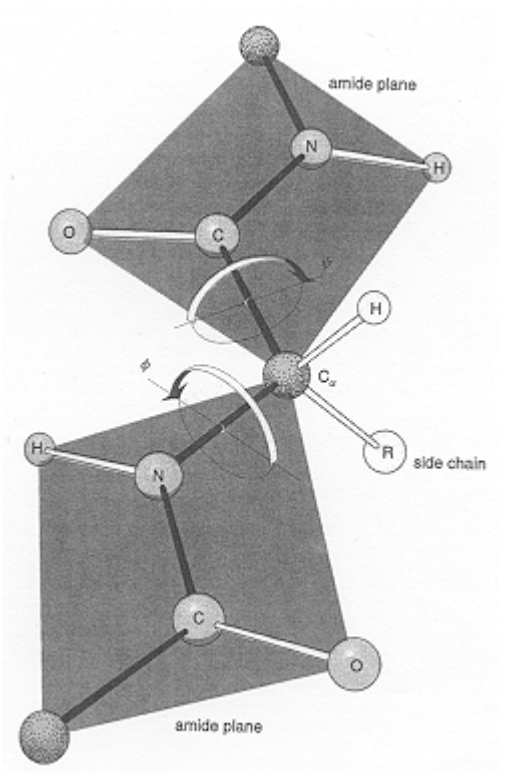


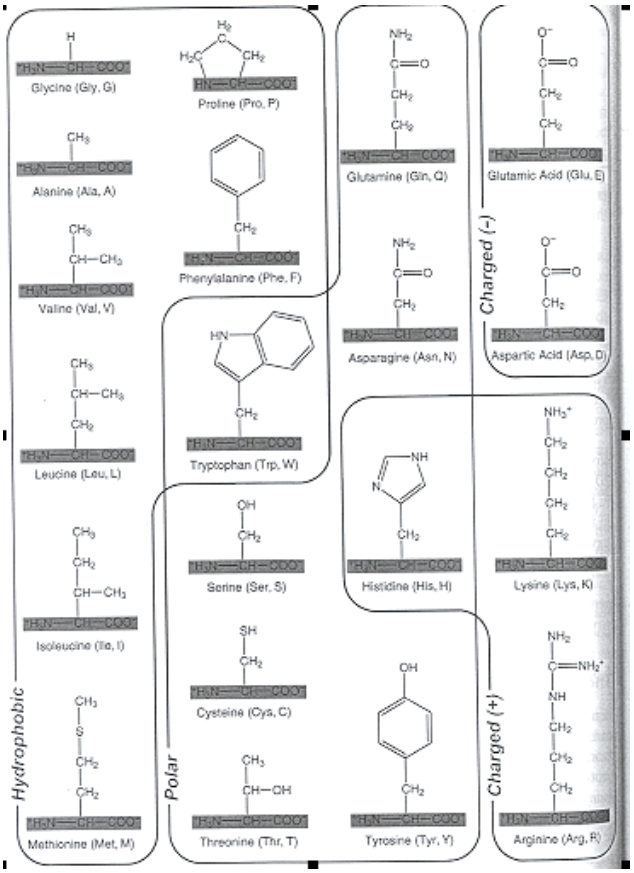
Un lien hydrogène (ou pont hydrogène) est plus faible qu'un lien covalent et est formé par l'attraction entre une zone chargée positivement et une zone chargée négativement. Ceci survient quand un hydrogène impliqué dans un lien covalent rencontre un atome d'une molécule qui a un nuage électrique important qui n'est pas impliqué dans un lien covalent (N ou O). Donc, c'est une attraction entre un proton (H, +) et un nuage électronégatif (N ou O).



2. Introduction à la structure des protéines.

Les protéines sont composées d'une séquence d'acides aminés connectées entre eux par un lien covalent entre l'atome de carbone (C) d'un acide aminé et l'atome d'azote (N) de l'acide aminé suivant.



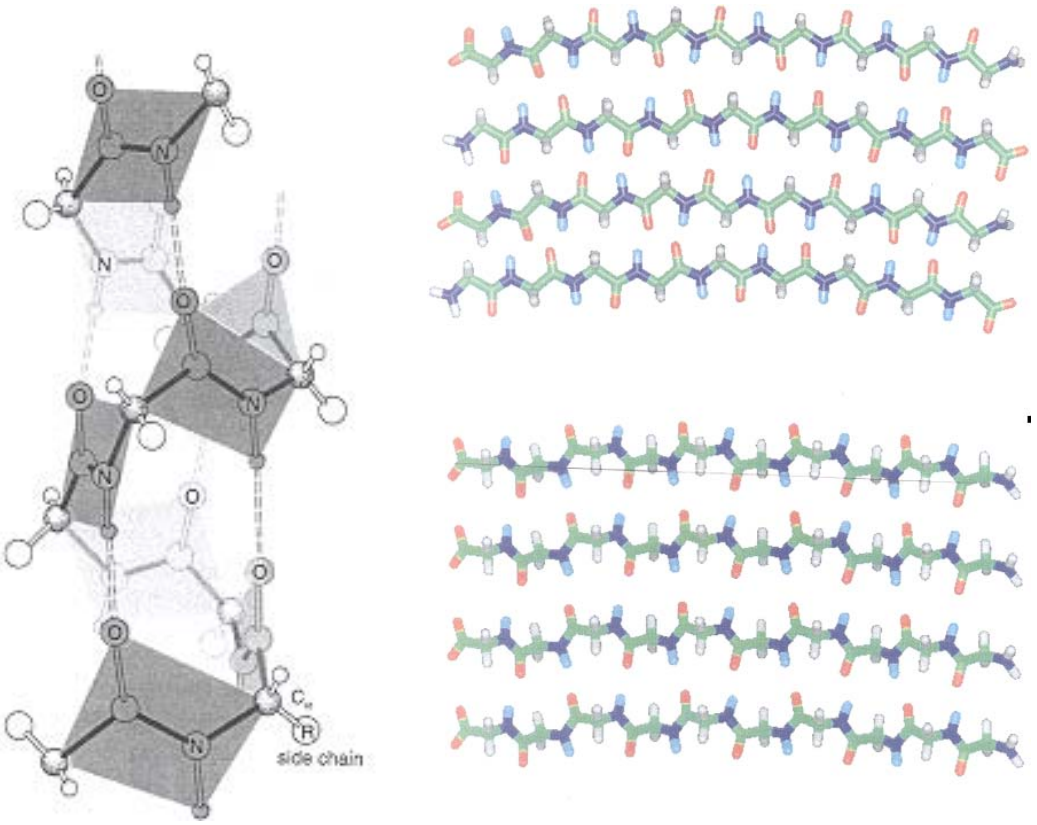


L'élément commun entre les acides aminés forme le « backbone » de la protéine. Les acides aminés sont différenciés par les particularités chimiques de leur chaînes latérales (side chain) telle que montré dans la figure précédente.

Les protéines arrivent à former des structures plus complexes en permettant des liens hydrogènes entre les éléments du backbone. Ceci stabilise grandement les molécules et plusieurs motifs structurels peuvent être observés. Un de ces motifs est l'hélice alpha. Notez (dans la figure qui suit) comment pour caque acide aminé, l'atome d'azote forme un pont hydrogène avec l'acide aminé à la position i+3. Ces interactions permettent à la protéine d'assurer une structure en forme d'hélice.

Un autre type de forme simple est aussi très fréquent dans les protéines, soit les feuillets bêta. Dans ce type de structures, les ponts hydrogènes sont formés entre brins parallèles ou anti-parallèles. Encore une fois, les ponts hydrogènes sont formés entre les éléments du

backbone de la structure. Dans l'image qui suit, les chaînes latérales sont perpendiculaires à la feuille (sortent du plan de la feuille a l'avant et à l'arrière de façon alternée).



Un élément qui motive la formation de structures complexe est la tendance des protéines à vouloir cacher leurs éléments hydrophobes (certaines chaînes latérales) du solvant, soit l'eau :

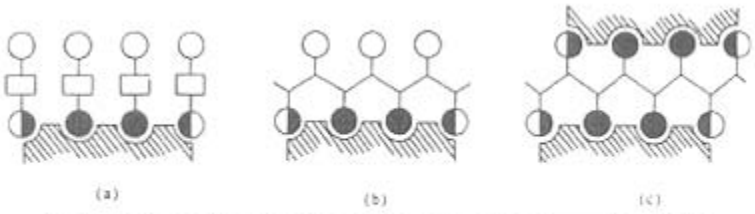


FIG. 25. Solvent accessibility of side chains of β-strands packed within protein molecules. Hatched areas show hydrophobic cores. (a) An end view of a β-sheet; (b) a side view of a β-strand packed on a hydrophobic surface; (c) a side view of a β-strand located between two hydrophobic cores. Solid circles, side chains completely buried in a hydrophobic core; semi-solid circles, partially buried side chains; open circles, side chains completely accessible to solvent molecules.

3. Stratégies de prédiction de la structure des protéines.

- Quand il est possible de trouver une structure connue qui a une bonne similarité de séquence avec la séquence que l'on désire modéliser, alors il est possible d'inférer certaines régions puisque les séquences qui sont exactes devraient avoir des structures similaires. Cette approche est désignée modélisation par homologie.
- Quand l'homologie entre les séquence et les structures connues est moins bonne, il est quand même possible de fixer la structure des petites régions similaires et ensuite d'utiliser l'information sur les forces chimiques pour prédire le reste de la structure.
- Finalement, il est aussi possible de tenter de prédire la structure à partir d'aucune autre information que la simple séquence en acides aminés (ab-initio). Ceci est un problème très difficile à résoudre et est souvent approché à l'aide d'heuristiques. Par exemple, une méthode consiste à explorer l'espace de l'énergie des différentes solutions et de trouver le minimum global. On essaie de trouver le minimum global par « minima hopping » !

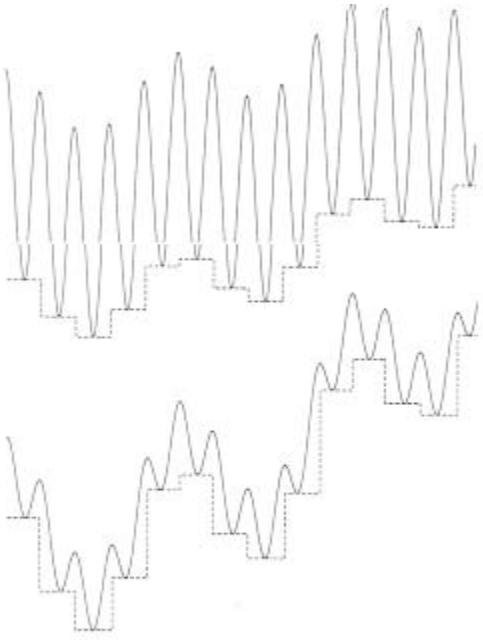


Figure 3: Two different potential energy landscapes: In the upper part the barriers separating basins are very high compared to the energy differences of the various local minima, in the lower part both are comparable.

4. Introduction à la structure des ARNs.

Comme pour les acides aminés, les nucléotides des ARNs ont une partie commune à tous les nucléotides qui forme le backbone de la structure totale. Le backbone est composé d'un phosphate et d'un ribose et les nucléotides sont liés entre eux du carbone 3' au phosphate du nucléotide voisin, ce qui détermine l'orientation 5'-3' de l'ARN.

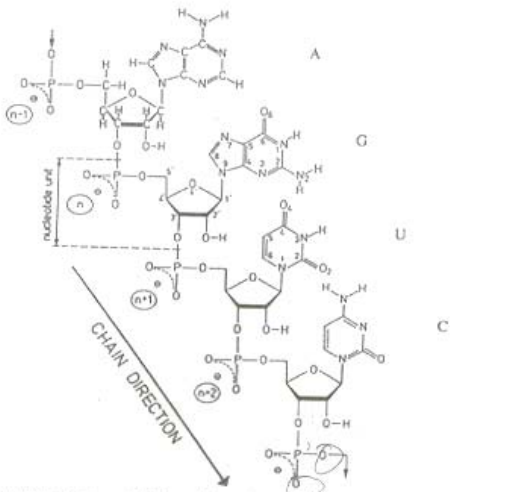
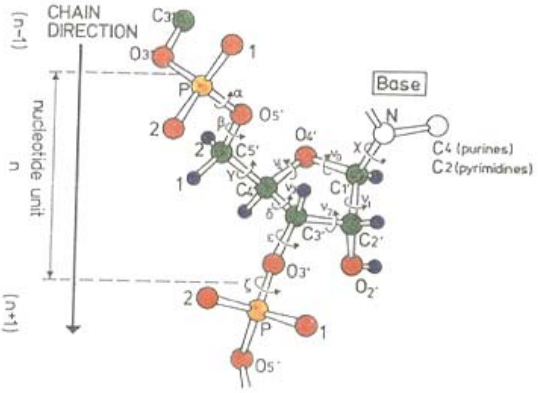
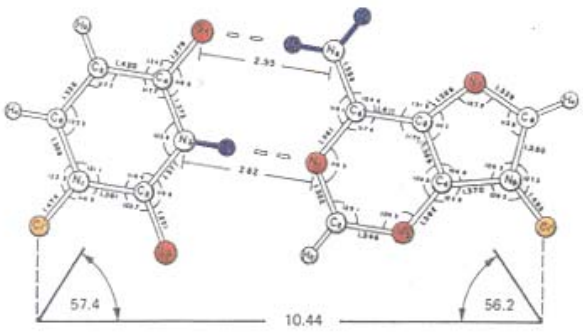
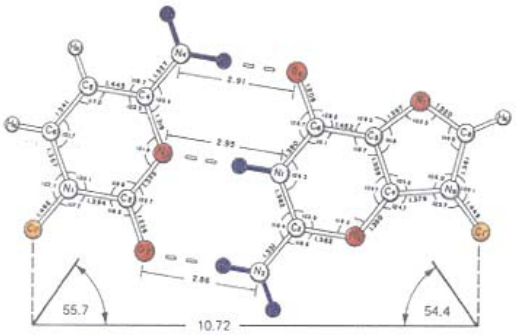


Figure 2-1. Fragment of ribonucleic acid (RNA) with sequence adenosine (A), guanosine (G), uridine (U), cytidine (C) linked by 3',5'-phosphodiester bonds. Chain direction is from 5'- to 3'-end as shown by arrow. Atom numbering scheme is indicated in one framed nucleotide unit, 5'-GMP. All hydrogen atoms drawn in A and only functional hydrogens in other nucleotides. In short notation, this fragment would be pApGpUpCp or pAGUCp. In deoxyribonucleic acid (DNA), the hydroxyl attached to $C_{2'}$ is replaced by hydrogen and uracil, by thymine.

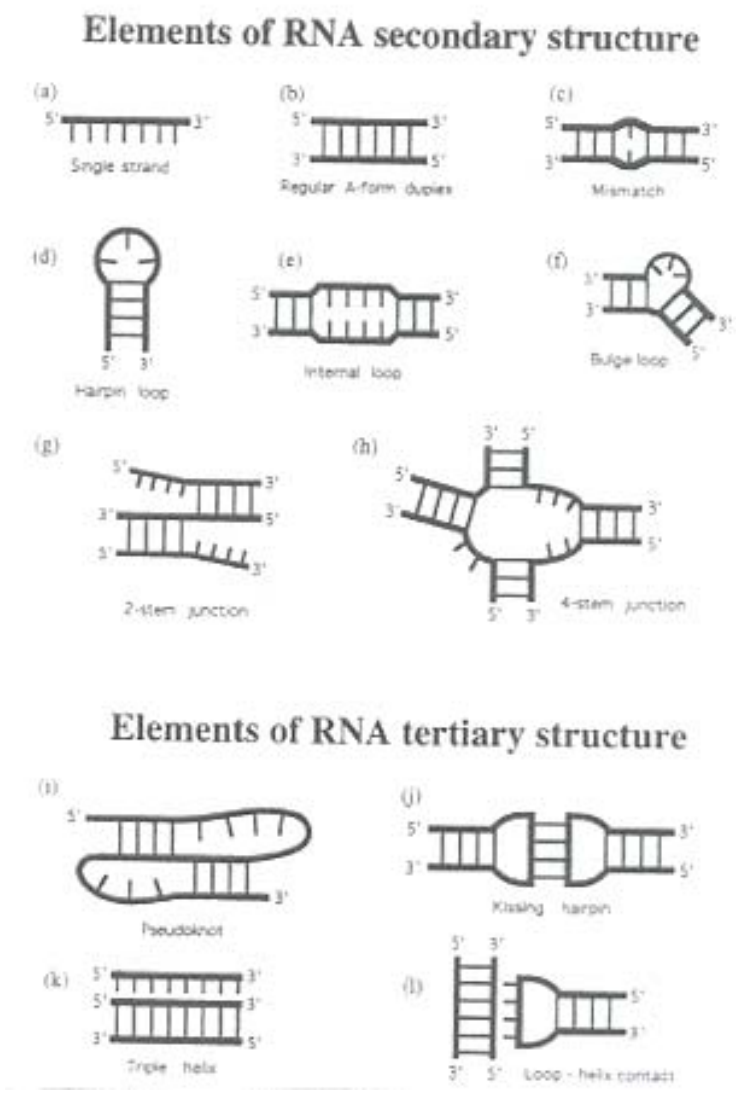


Par contre, contrairement aux protéines, la structure de l'ADN n'est pas assurée par des liens entre les éléments du backbone, mais plutôt par des appariements entre les paires de bases. Entre autre, A fait deux ponts hydrogènes avec U et C fait trois ponts hydrogènes avec G.





Les appariements entre paires de bases créer des motifs facilement reconnaissables. Ces motifs ont été caractérisés en deux dimensions pour simplifier la compréhension des structures de l'ARN, mais ces structures n'existent as réellement, ce sont des représentations créer par l'homme pour simplifier la réalité.



De plus il faut savoir qu'en trois dimensions, l'ARN est capable de faire beaucoup plus d'appariements que les paires Watson-Crick (A-U, C-G). En effet, les nucléotides peuvent faire des interactions stables de plusieurs façons ! Par contre, les problèmes de prédiction de structures et de recherche de molécules ne considèrent souvent que les paires Watson-Crick en plus de la paire Wobble (G-U).

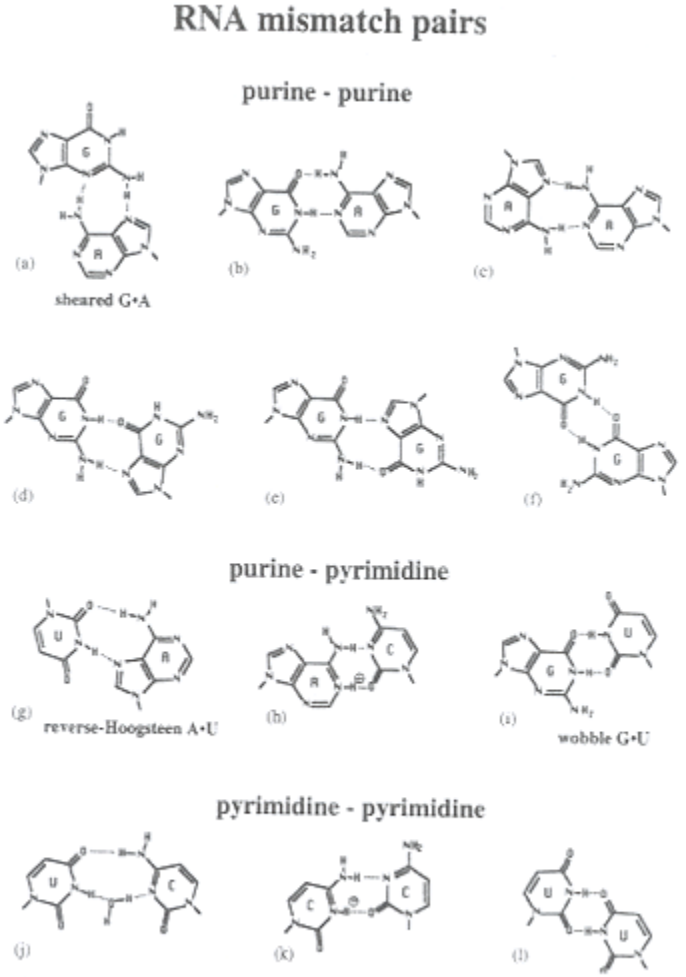


Fig. 18.4. Non-Watson–Crick base pairs observed in solution structures of RNA. (a) Sheared G·A pair (18,58); (b) G·A (59); (c) A·A (60,61); (d)–(f) G·G mismatches (29,30,64). (g) reverse Hoogsteen A·U pair (60,61); (h) protonated A^+·C pair (62,68). (i) wobble G·U pair (36,72); (j) water-mediated U·C pair (72). (k) protonated C^+·C pair (53,75). (l) U·U mismatch (53,75)

5. Stratégies de prédiction de la structure des ARNs.

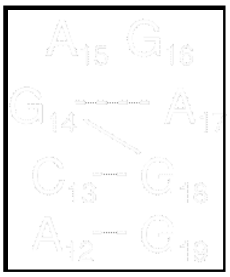
- Méthode combinatoire. (démo 8)
- Minimisation de l'énergie libre. (démo 8)
- Modélisation à partir d'un dictionnaire d'éléments structuraux (Mc-Sym)

Mc-Sym est un logiciel de modélisation de structures tridimensionnelles d'ARN qui utilise des coordonnées et des relations entre résidus extraites de structures cristallographiées, de résultats de RMN et de modèles théoriques. Ces structures proviennent des bases de données de « RCSB Protein Data Bank » et de « Rutgers Nucleic Acid Database » . Le modélisateur permet l'ajout de contraintes à la procédure de construction afin de générer des modèles valides. Les instructions de construction et de contraintes peuvent s'écrire dans un script ou être entrées inter activement dans le modélisateur Mc-Sym.

RNA 3-D Modelling Steps

1. Get structural information (theory, experiment)
2. Convert in *MC-Sym* syntax (human)
3. Generate 3-D structures (*MC-Sym*)
4. Visualize/Analyze generated structures (*MC-Annotate,* clustering, comparison, etc)
5. Goto 1

Structural Data In *MC-Sym* Syntax (the GAGA tetraloop)



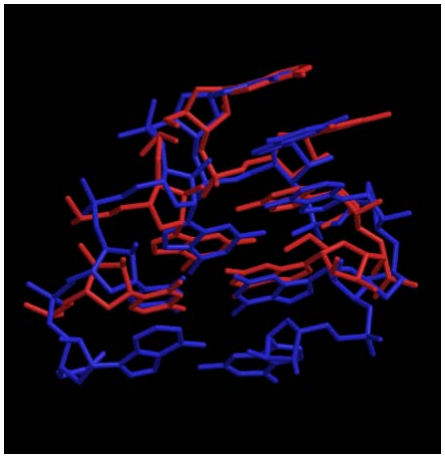

```
sequence( r 13  CGAGAG )
residue(
  13  { type_A }    5
  14  {}            5
  15  {}            5
  16  {}            5
  17  {}            5
  18  { type_A }    5
)
connect(
  13  14  { stack }     5
  14  15  { nostack }  10
  15  16  { stack }     5
  16  17  { stack }     5
  17  18  { helix }     1
)
pair(
  14  17  { XI }    30
  13  18  { wc }     5
)
```

```
gaga = backtrack(
  ( 18 13 )
  ( 13 14 17 )
  ( 17 16 15 )
)
gaga_cache = cache(
  gaga
  rmsd_bound 0.5
  all    no_hydrogen
)
adjacency( gaga 1.0 3.0 )
res_clash(
  gaga
  fixed_distance 1.0
  all  no_hydrogen
)
explore(
  gaga_cache
  fileName_pdb ("gaga-%03d.pdb")
  zipped
)
```

Lemieux, Oldziej, Major "Nucleic Acids : Qualitative Modeling, in" The Encyclopedia of Computational Chemistry, Schleyer et al. (Eds.); John Wiley & Sons: Chichester. 1998 1930-1941.

3-D Structures: The GAGA Tetraloop

(`linux% mcsym gaga.mcc`)



Corell et al. (1998) *PNAS(USA)* **95**, 13436.

- Results: 69 structures.
- Best structure: 1.4Å.
- Variance: 2.9Å.
- Search space: 2 929 687 500.
- Running time: 6.2 sec (PIII 600 MHz).

Best *MC-Sym* structure (red) superimposed with the crystal structure of the Sarcin/Ricin loop from rat rRNA 28S.

6. Introduction à la recherche des ARNs.

Nous avons précédemment vu comment rechercher des séquences avec erreurs (alignement semi-local) et une heuristique pour approcher cette recherche (Blast). La recherche de protéines et de gènes codant pour celles-ci avec Blast fonctionne très bien. La raison est que des protéines ou des gènes ayant la même fonction doivent posséder une forte similarité. Par contre, les molécules d'ARNs ont moins de similarité entre elles au niveau de la séquence, même si elles ont la même fonction. En effet, dans le cas de l'ARN, la composition de la séquence en acides nucléiques est moins importante tant et aussi longtemps que la structure de la molécule reste la même. Ainsi, il est facile de modifier un appariement pour un autre dans la structure secondaire d'un ARN sans changer la structure. Par contre, les molécules résultantes perdront en similarité. Ce phénomène est nommé : covariation. C'est-à-dire qu'un élément d'une molécule peut changer tant et aussi longtemps que la molécule avec laquelle elle interagit pour maintenir la structure change aussi pour préserver la structure globale. La covariation est un phénomène important à considérer dans la recherche de molécules d'ARN puisque la recherche d'une séquence consensus seule arrive rarement à trouver toutes les molécules recherchées. Ainsi, il faut ajouter de l'information sur la structure secondaire de la molécule pour améliorer la recherche.

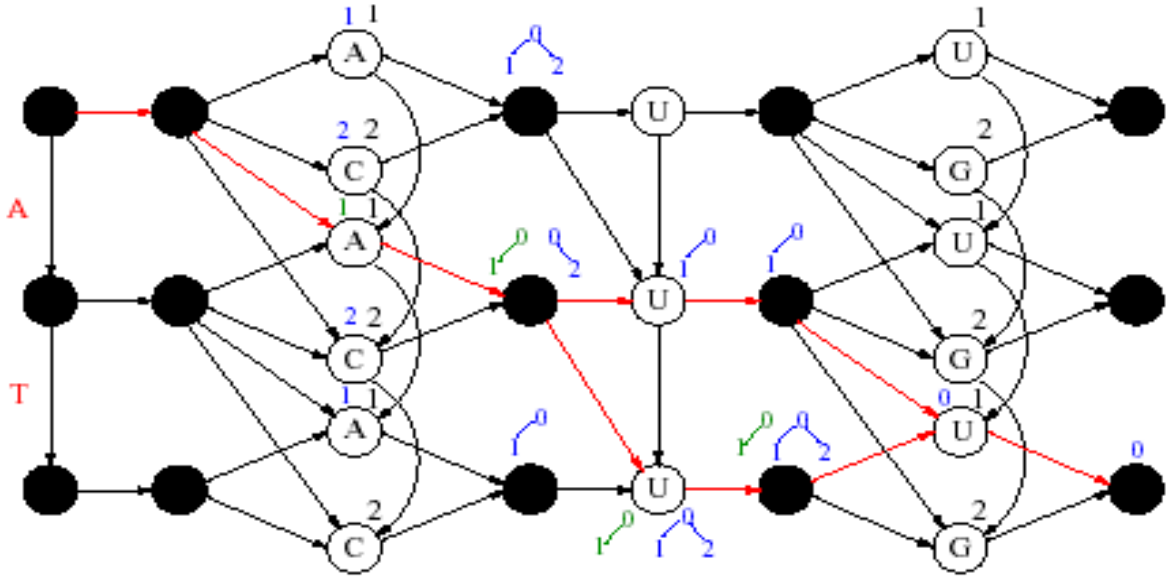
7. Façons de représenter les éléments de structure secondaire.

i. Automate à pile

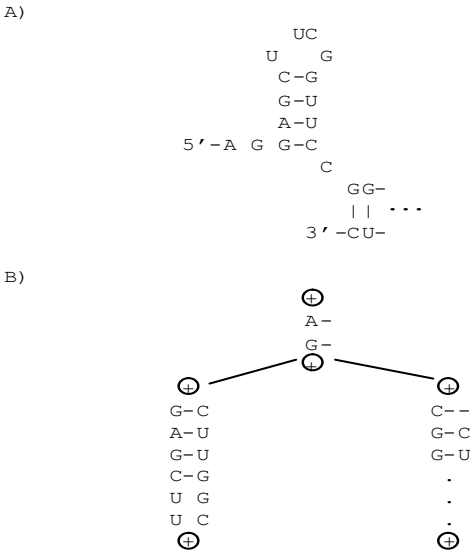
$S = \{(A|C), sl\}\{T, p\}\{(T|G), sr\}$, reading AT.

At most $k = 1$ error.

Blue stacks for $k = 1$; Green stacks for $k = 0$.

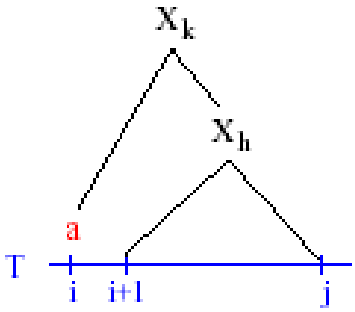


ii. Arbres

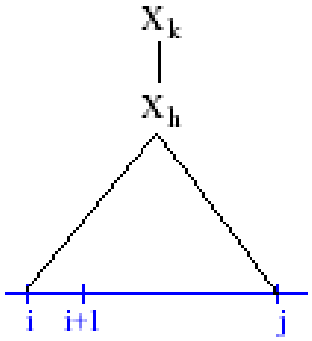


iii. Grammaires

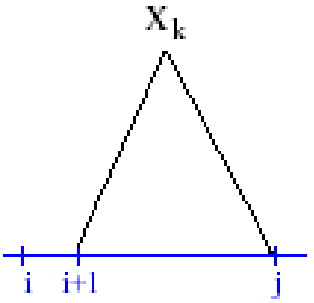
$$X \rightarrow a\,Y\,\overline{a}\,, \quad X \rightarrow a\,Y\,, \quad X \rightarrow Y\,a\,, \quad X \rightarrow a$$



a aligned with T[i]



Deletion of a



Deletion of T[i]