# Cours M2 BIM - Séance 3

## Comparaison et interactions

Yann Ponty

Bioinformatics Team
École Polytechnique/CNRS/INRIA AMIB – France

28 Janvier 2011

# Résumé

# Pourquoi aligner structurellement des ARN

Une pression évolutive commune permet d'identifier une fonction commune. Chez certains organismes (et pour certaines familles d'ARN), très faible conservation de la séquence. Cependant, la structure peut être bien plus conservée, et connue (Expérimentalement) ou déterminée par repliement.

Problèmes :

- Édition : Trouver la *distance* entre deux structures $A$ et $B$.
  Quelle est la séquence d'opérations de coût minimal permettant de passer de $A$ à $B$ ? Déjà NP-complet pour deux structures secondaires [BFRS07].
- Alignement : Trouver une super-structure de coût minimal.
  Généralise la notion d'alignement de séquence. Polynomial pour des structures secondaires [BDD+08], NP-complet en 3D [SZS+08].
  Variantes : Alignement local ou global, Recherche de motifs.
- Superposition : Trouver une transformation géométrique (Rotation, translation, zoom) pour superposer *au mieux* les coordonnées de deux ARN de matching connu. Polynomial en 3D [McL82].

$\Rightarrow$ La difficulté algorithme provient de la recherche du matching initial.

# FR3D : Un exemple d'approche géométrique

Quand les structures tertiaires (3D) des ARN sont connues, le problème de l'alignement peut être abordé de façon purement géométrique.

**Problème**

Donnée : Motif $m$ et structure cible $b$ (Ensembles de bases 3D).
Résultat : Matching de $m$ et d'un sous-ensemble de $b$ minimisant une divergence géométrique.

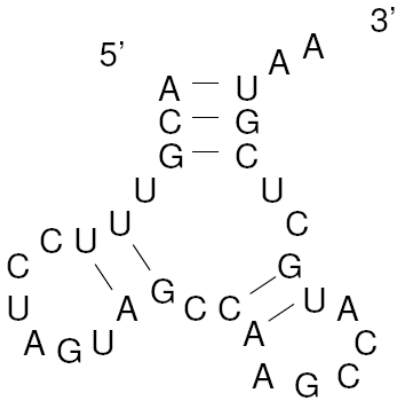
Divergence géométrique : Dans FR3D [SZS+08], une fonction $D$ basée sur deux fonctions $L$ et $A$ d'erreur tenant compte respectivement de la superposabilité ($L$) et de l'orientation des bases ($A$) de $m$ et $b$.

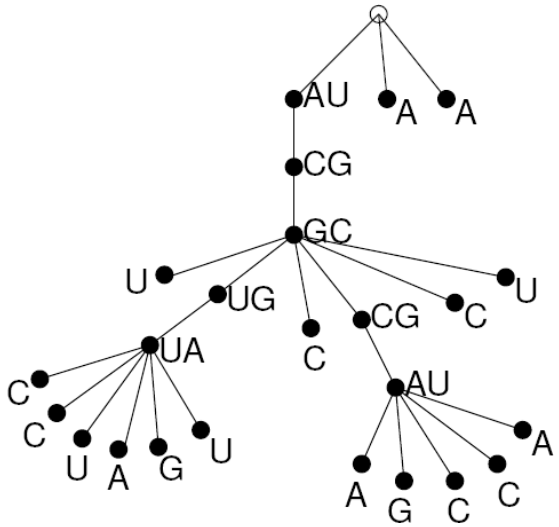
$$L = \sqrt{\min_{R,T} \sum_{i=1}^{m} \|b_i - R(T(m_i))\|^2} \quad A = \sqrt{\sum_{i=1}^{m} \alpha_i^2} \quad D = \frac{1}{m}\sqrt{L^2 + A^2}$$

$R, T$ : Rotation et translation. $c_i$ : Barycentre pour la base $m_i$. $\alpha_i$ : Écart entre les axes barycentre/bases dans $m_i$ et $b_i$.

Exploration (Backtrack) + Élagage incrémental (Bornes sur $D$) $\Rightarrow$ Explosion. Mais recherche exacte pour des petits motifs.

L'alignement de deux structures secondaires est basé sur une représentation arborescente de la structure secondaire [1].

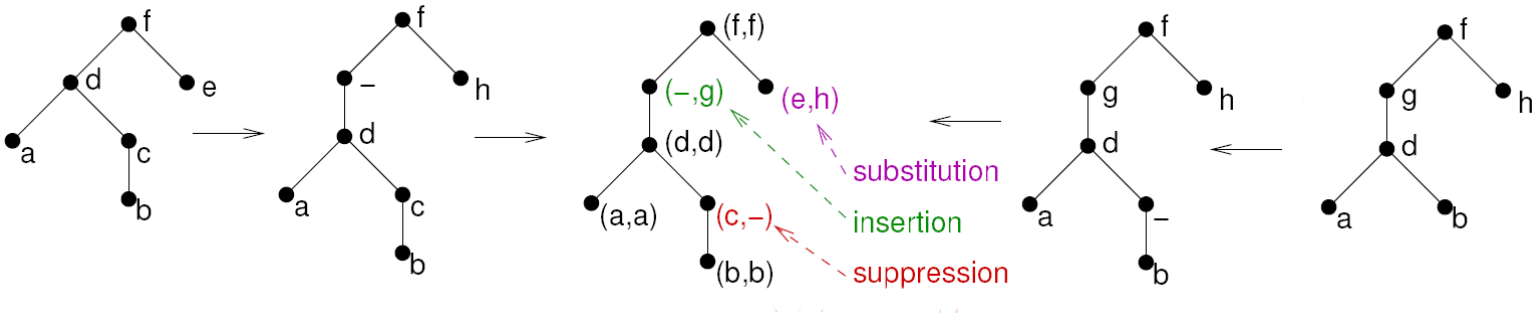



Paires de bases ⇒ noeuds internes

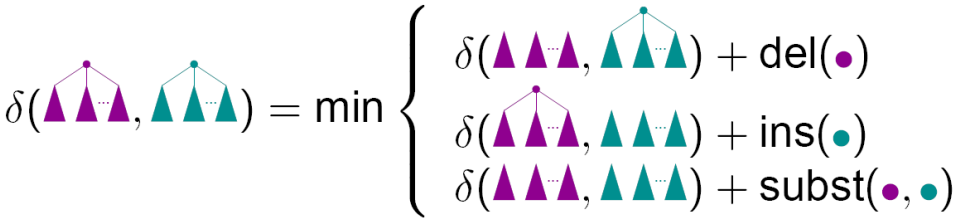
Bases non-appariées ⇒ Feuilles

Alignement = Construction d'un matching complet de coût minimal.

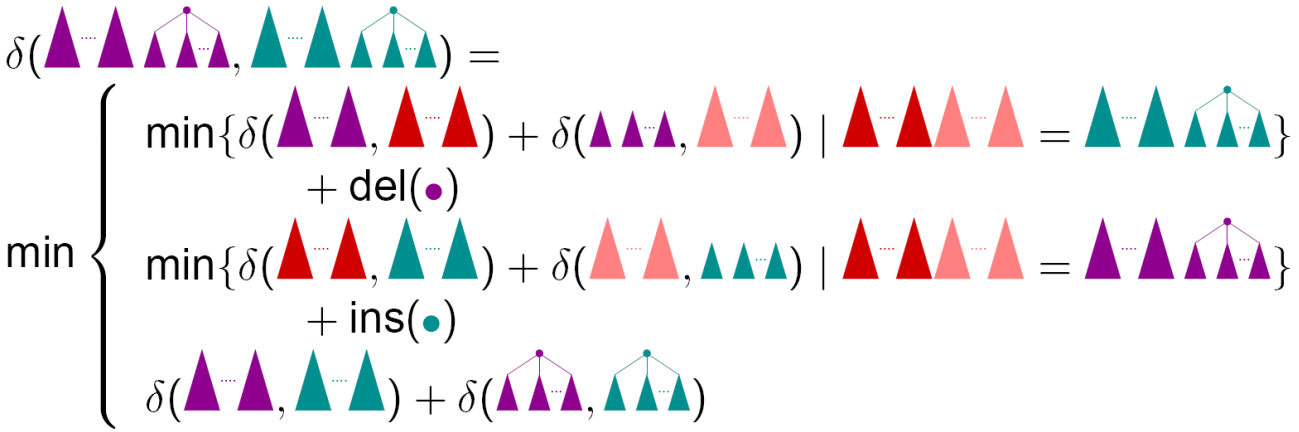



1. Illustrations empruntées à C. Herrbach

Alignement d'arbre [2]



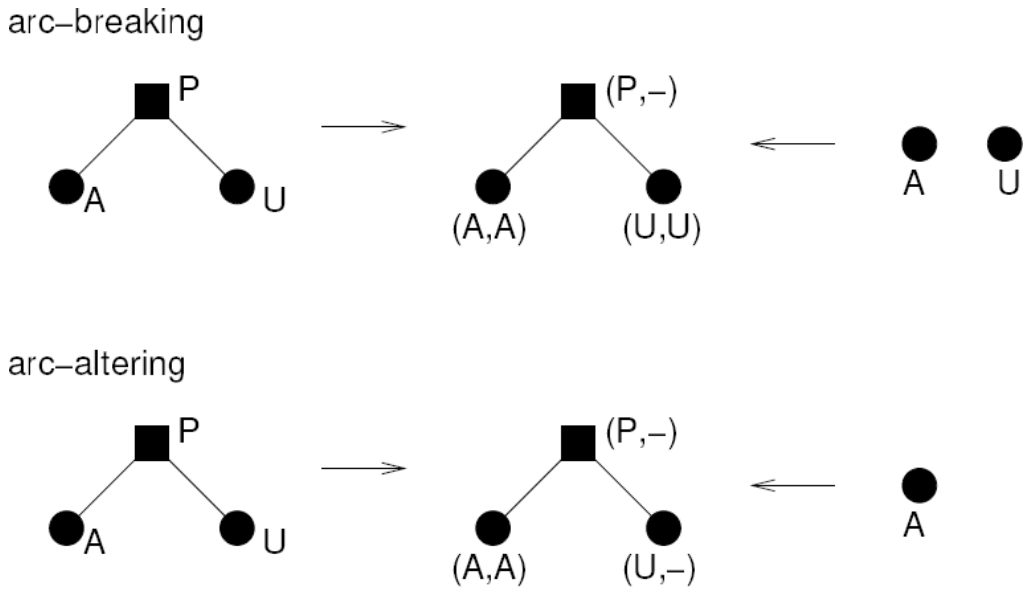

Alignement de forêt




Complexité au pire en $\mathcal{O}(n^4)$ [JWZ94], en moyenne en $\mathcal{O}(n^2)$ [HDD07].
Mais opérations spécifiques à l'ARN manquantes.

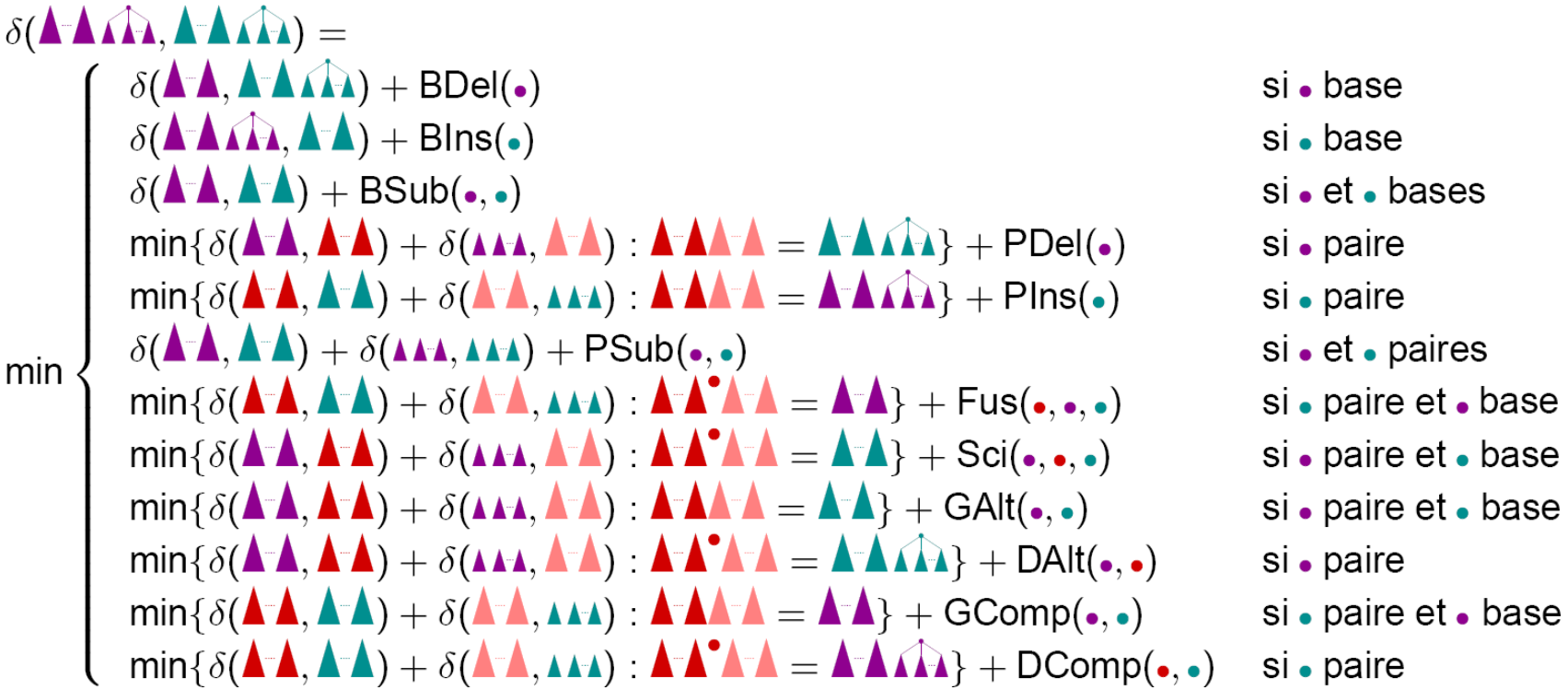
2. Idem

Basé sur l'algorithme de Jiang, Wang & Zhang
+ Intégrations d'opérations spécifiques à l'ARN [3].



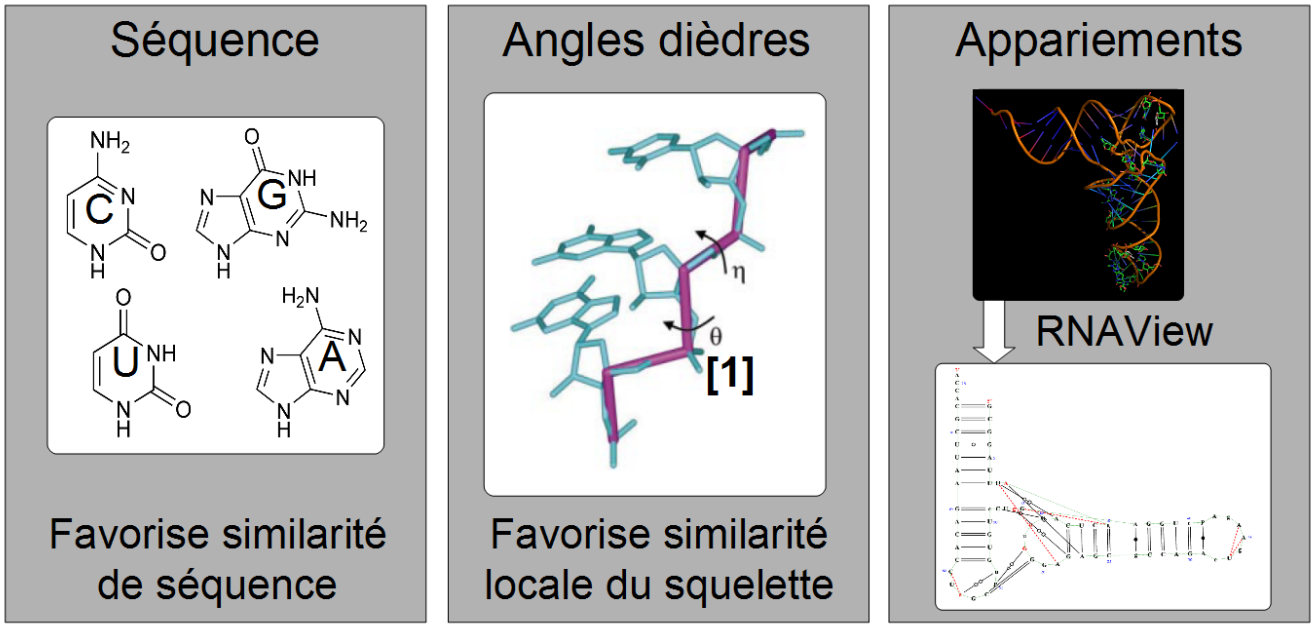

Possibilité de paramètrer les coûts des opérations, mais certaines opérations atomiques dans un modèle réalistes doivent être recomposées à partir des opérations disponibles. Par exemple, la substitution d'un sommet par une feuille est interdite directement.

3. Idem


δ(, ) =
min
δ(, ) + BDel(•) si • base
δ(, ) + BIns(•) si • base
δ(, ) + BSub(•, •) si • et • bases
min{δ(, ) + δ(, ) : = } + PDel(•) si • paire
min{δ(, ) + δ(, ) : = } + PIns(•) si • paire
δ(, ) + δ(, ) + PSub(•, •) si • et • paires
min{δ(, ) + δ(, ) : = } + Fus(•, •, •) si • paire et • base
min{δ(, ) + δ(, ) : = } + Sci(•, •, •) si • paire et • base
min{δ(, ) + δ(, ) : = } + GAlt(•, •) si • paire et • base
min{δ(, ) + δ(, ) : = } + DAlt(•, •) si • paire
min{δ(, ) + δ(, ) : = } + GComp(•, •) si • paire et • base
min{δ(, ) + δ(, ) : = } + DComp(•, •) si • paire

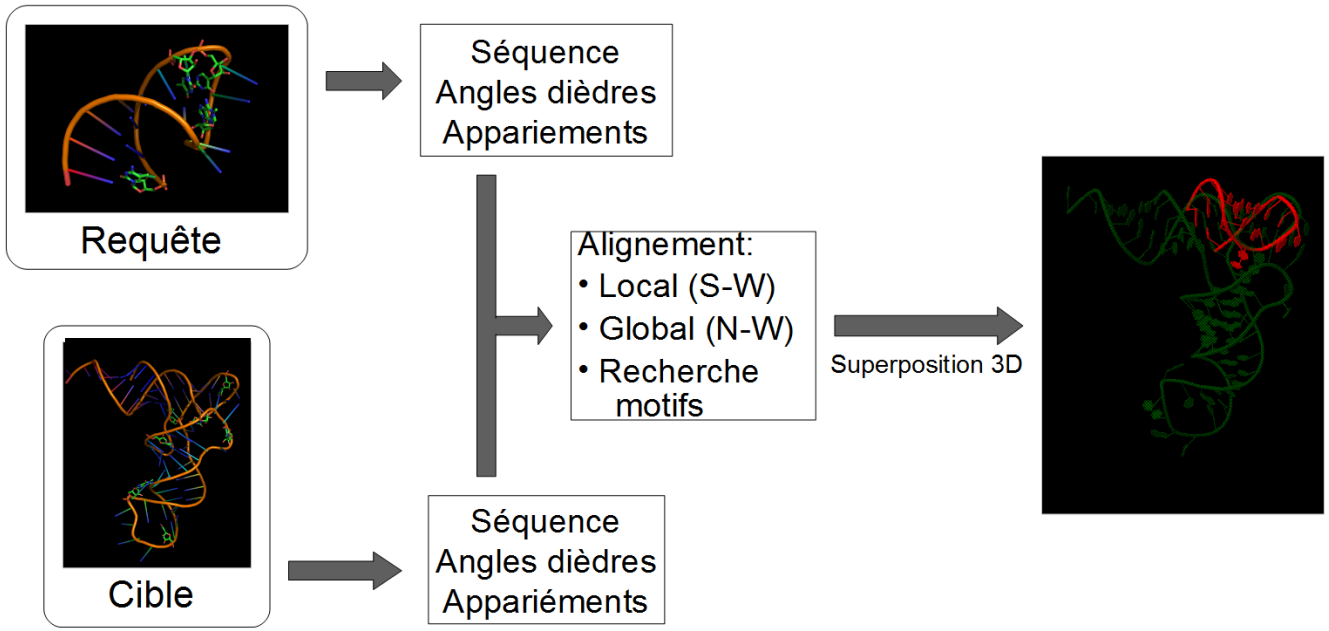
DIAL [FPLC07] est une méthode hybride qui se concentre sur les comportements locaux.
Idée : L'ARN est flexible, petite variation locale peuvent entraîner des grandes déviations géométriques.

DIAL capture les similarités locales à trois niveau :



DIAL [FPLC07] est une méthode hybride qui se concentre sur les comportements locaux.
Idée : L'ARN est flexible, petite variation locale peuvent entraîner des grandes déviations géométriques.

Un algorithme d'alignement de séquence est alors utilisé



Tout dépend de ce que l'on a et veut :

- Modèle 3D :
  - Recherche d'un motif peu conservé en séquence : FR3D
  - Recherche d'un motif conservé : FR3D, DIAL ou DARTS
  - Recherche d'une structure entière : DIAL ou DARTS

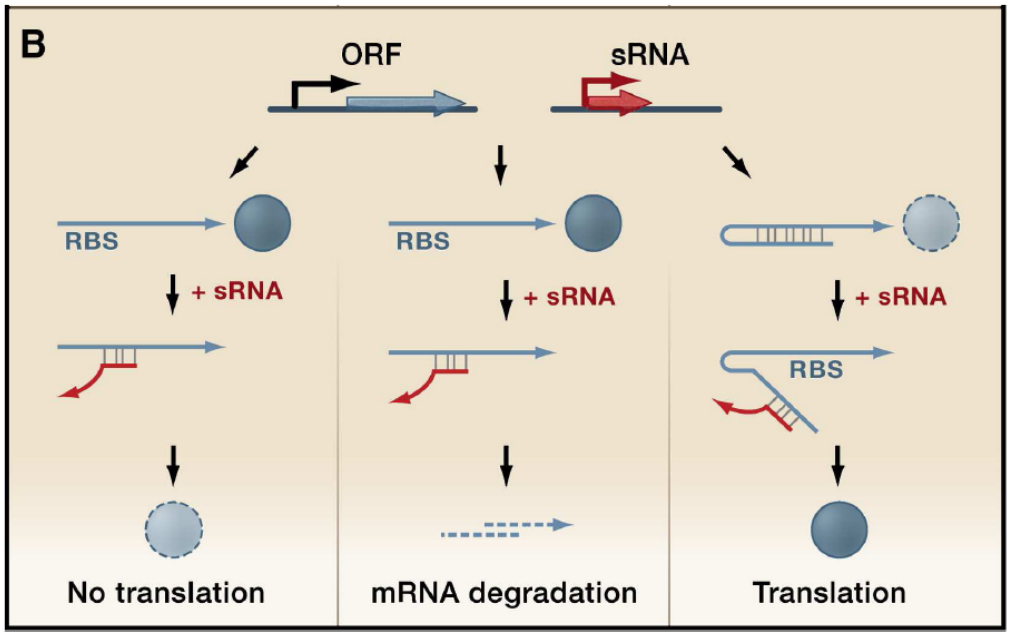
- Structure secondaire :
  - Recherche d'un motif : NestedAlign
  - Alignement structure : RNAForester, NestedAlign

De nombreux autres programmes disponibles : Migal, Magnolia, ...
+ Explosion des approches *par fragments* : FASTR3D, RNA FRABASE, ...

# Prédire les interactions : Motivation

- La prédiction du repliement des macromolécules vise *in fine* à la prédiction de fonction.
- La fonction s'exprime dans le contexte des réseaux d'interactions.
- Comprendre le repliement permet d'identifier les acteurs du réseau.
- Prédire les interactions permet d'établir les liens du réseau.

Exemple : Petits ARNs régulateurs chez les bactéries

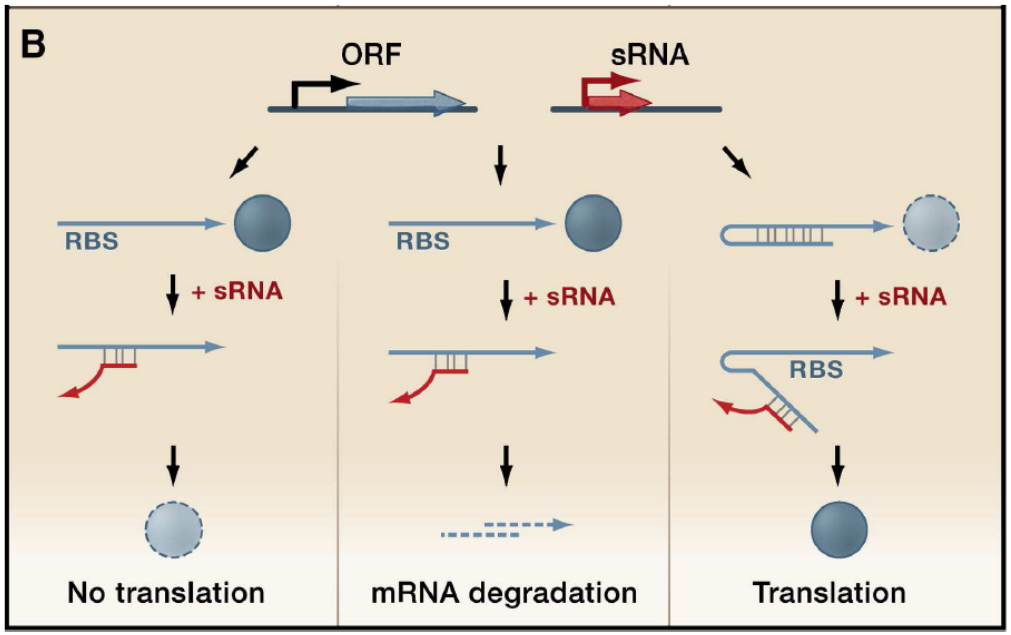



(Waters and Storz, Cell 2009)

# Prédire les interactions : Motivation

- La prédiction du repliement des macromolécules vise *in fine* à la prédiction de fonction.
- La fonction s'exprime dans le contexte des réseaux d'interactions.
- Comprendre le repliement permet d'identifier les acteurs du réseau.
- Prédire les interactions permet d'établir les liens du réseau.

Exemple : Petits ARNs régulateurs chez les bactéries




(Waters and Storz, Cell 2009)

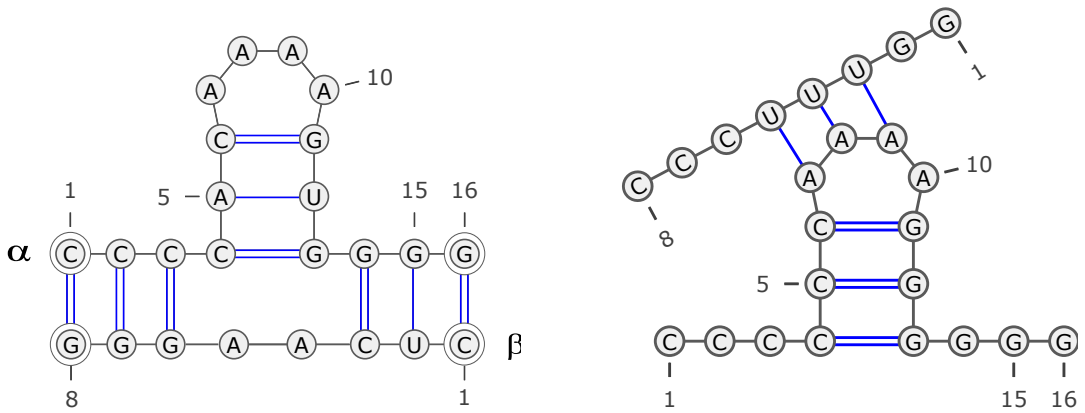
⇒ Comment prédire les interactions ?

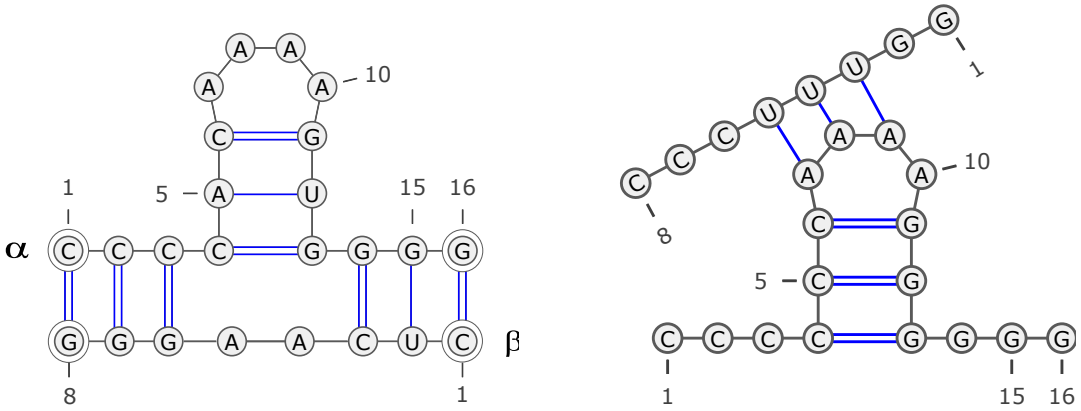
Données : Deux ARN de séquence $\alpha$ et $\beta$.

Idéalement, on souhaiterait parcourir simultanément et efficacement :

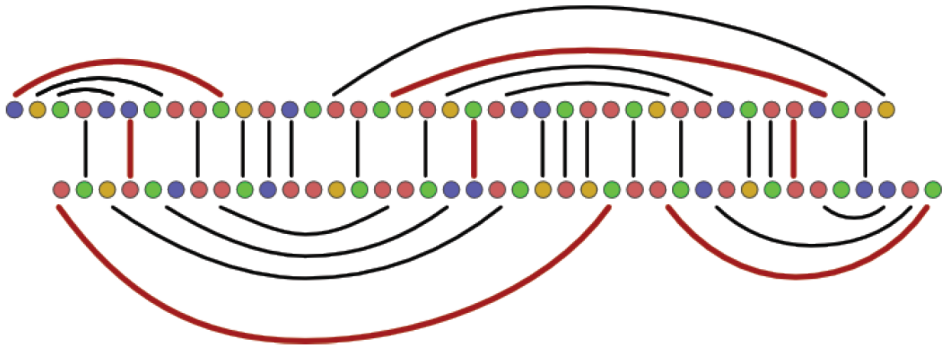
- toutes les structures secondaires valables pour $\alpha$,
- toutes les structures secondaires valables pour $\beta$,
- toutes appariements de base libres entre $\alpha$ et $\beta$,

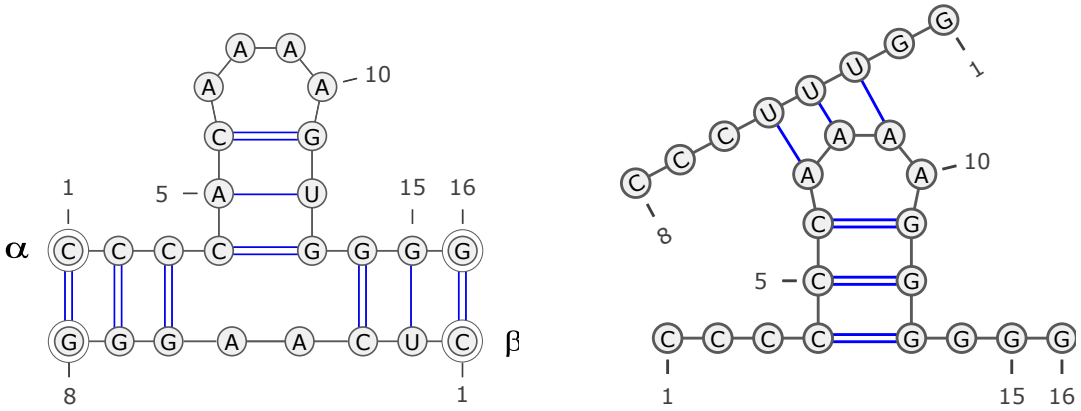
en associant à chaque combinaison une énergie (par exemple additive), et renvoyer la configuration d'énergie minimale.



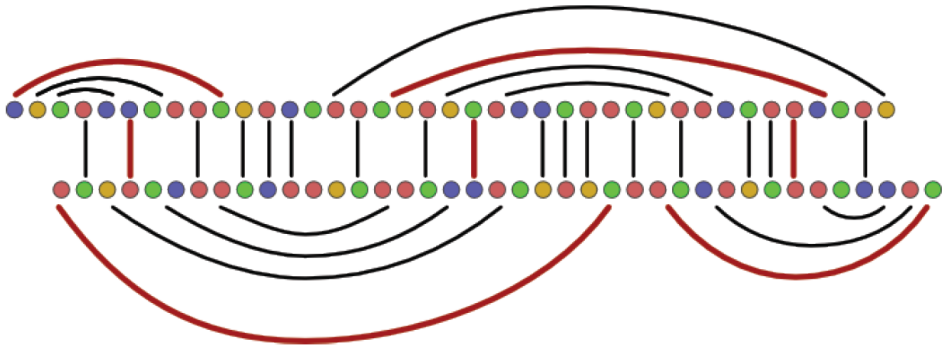



Malheureusement, il est fort probable qu'un tel algorithme n'existe pas, car le problème devient NP-COMPLET dès que des motifs en zig-zags de longueur arbitraire sont autorisés [AKN$^+$06] :

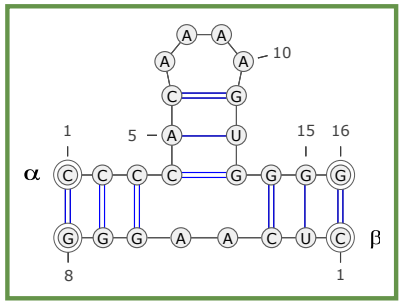


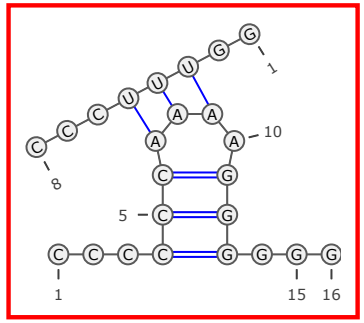



Malheureusement, il est fort probable qu'un tel algorithme n'existe pas, car le problème devient NP-COMPLET dès que des motifs en zig-zags de longueur arbitraire sont autorisés [AKN+06] :

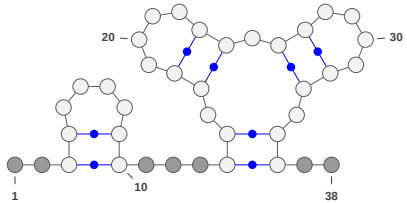


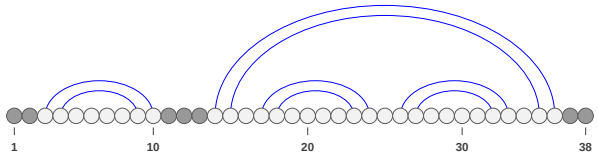
⇒ Considérer des sous-ensembles d'interactions



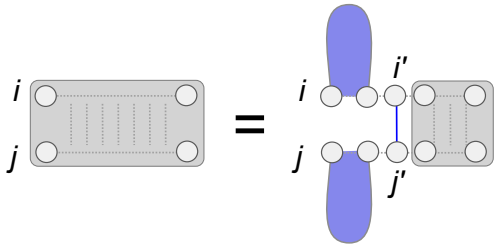



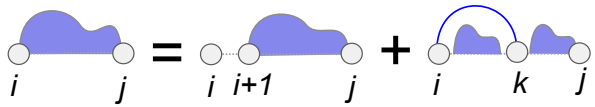

- Interaction = Alignement



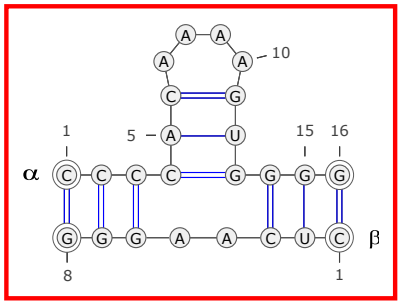



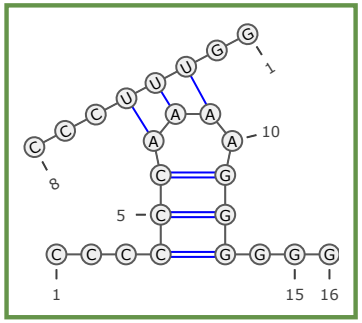

Seules les positions extérieures sont autorisées à interagir (RNACofold).



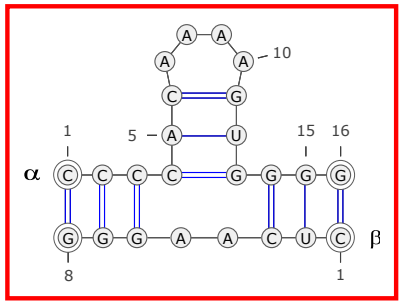



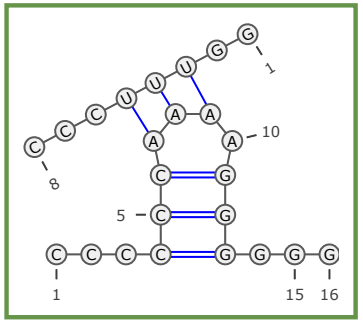
# Approches principales



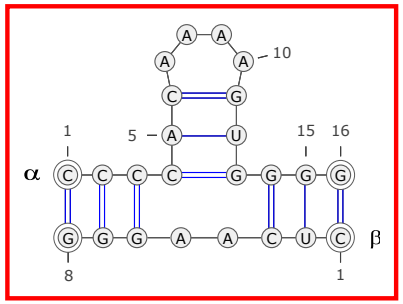



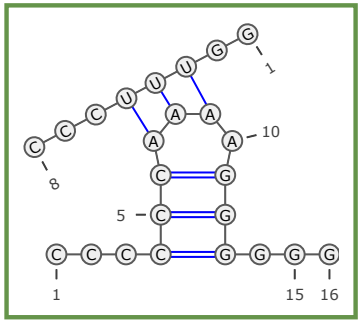

- Interaction = Alignement
- Interaction = Appariement de deux zones privilégiées
  Stratégie RNAUp [MTH+06] :



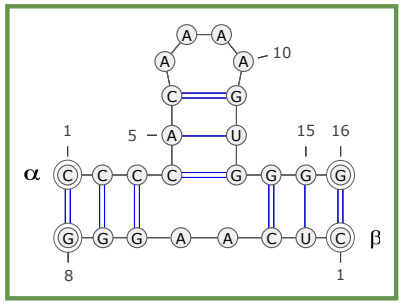



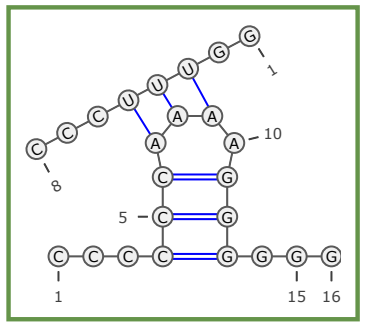

- Interaction = Alignement
- Interaction = Appariement de deux zones privilégiées
  Stratégie RNAUp [MTH+06] :
  - Repérer une zone dans $\alpha$ susceptible d'être impliquée (Probabilité de Boltzmann pour le non-appariement)



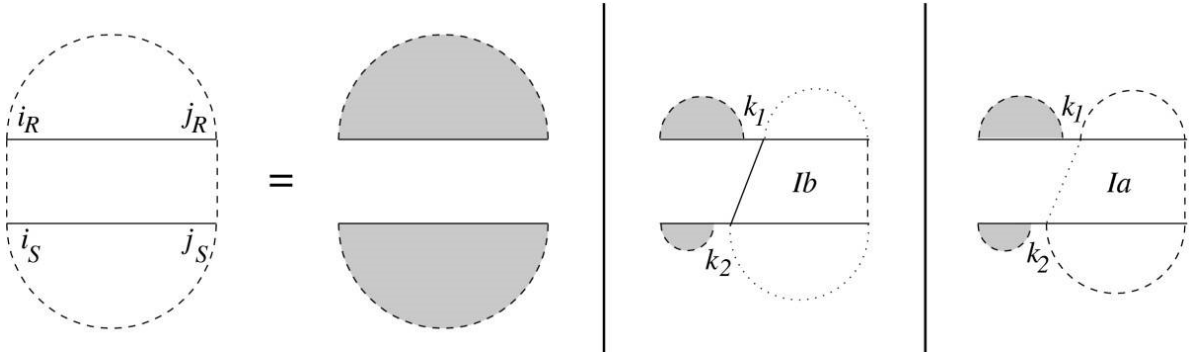




- Interaction = Alignement
- Interaction = Appariement de deux zones privilégiées
  Stratégie RNAUp [MTH+06] :
  - Repérer une zone dans $\alpha$ susceptible d'être impliquée (Probabilité de Boltzmann pour le non-appariement)
  - Simuler un repliement joint localement



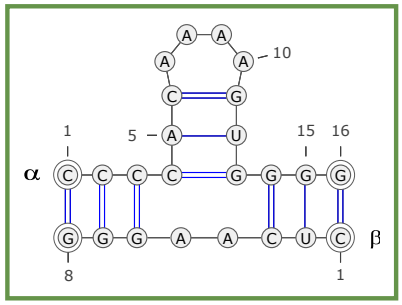



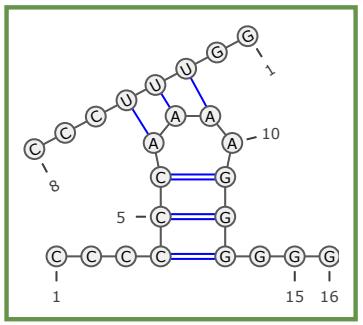

- Interaction = Alignement
- Interaction = Appariement de deux zones privilégiées
  Stratégie RNAUp [MTH+06] :
  - Repérer une zone dans $\alpha$ susceptible d'être impliquée (Probabilité de Boltzmann pour le non-appariement)
  - Simuler un repliement joint localement
- Plus complexe/complet...
  Approche mixte Repliement/alignement [SBS10]




Mais complexité élevée $\mathcal{O}(n^3 \cdot m^3)$.

# Approches principales




- Interaction = Alignement
- Interaction = Appariement de deux zones privilégiées
  Stratégie RNAUp [MTH+06] :
  - Repérer une zone dans $\alpha$ susceptible d'être impliquée (Probabilité de Boltzmann pour le non-appariement)
  - Simuler un repliement joint localement
- Plus complexe/complet...
  Approche mixte Repliement/alignement [SBS10]
  Mais complexité élevée $\mathcal{O}(n^3 \cdot m^3)$.

Champs de recherche extrêmement actif/compétitif !

Quel serait votre algorithme (TP) ?

# References I


Can Alkan, Emre Karakoç, Joseph H. Nadeau, S. Cenk Sahinalp, and Kaizhong Zhang.

**Rna-rna interaction prediction and antisense rna target search.**
*Journal of Computational Biology*, 13(2) :267–282, 2006.

G. Blin, A. Denise, S. Dulucq, C. Herrbach, and H. Touzet.

**Alignment of rna structures.**
*Transactions on Computational Biology and Bioinformatics*, ... :..., 2008.
A paraître.

Guillaume Blin, Guillaume Fertin, Irena Rusu, and Christine Sinoquet.

**Extending the Hardness of RNA Secondary Structure Comparison.**
In Bo Chen, Mike Paterson, and Guochuan Zhang, editors, *ESCAPE'07*, volume 4614 of *LNCS*, pages 140–151, Hangzhou, China, Apr 2007.

F. Ferrè, Y. Ponty, W. A. Lorenz, and Peter Clote.

**Dial : A web server for the pairwise alignment of two RNA 3-dimensional structures using nucleotide, dihedral angle and base pairing similarities.**
*Nucleic Acids Research*, 35(Web server issue) :W659–668, July 2007.

Claire Herrbach, Alain Denise, and Serge Dulucq.

**Average complexity of the jiang-wang-zhang pairwise tree alignment algorithm and of a rna secondary structure alignment algorithm.**
In *Proceedings of MACIS 2007, Second International Conference on Mathematical Aspects of Computer and Information Sciences*, 2007.

M. Hochsmann, B. Voss, and R. Giegerich.

**Pure multiple RNA secondary structure alignments : A progressive profile approach.**
01(1) :53–62, 2004.

Tao Jiang, Lusheng Wang, and Kaizhong Zhang.

**Alignment of trees - an alternative to tree edit.**
In *CPM '94 : Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, pages 75–86, London, UK, 1994. Springer-Verlag.

# References II


D. McLachlan.

**Rapid comparison of protein structures.**
*Acta cristallographica A*, 38(6) :871–873, 1982.

Ulrike Muckstein, Hakim Tafer, Jorg Hackermuller, Stephan H. Bernhart, Peter F. Stadler, and Ivo L. Hofacker.

**Thermodynamics of RNA-RNA binding.**
*Bioinformatics*, 22(10) :1177–1182, 2006.

Raheleh Salari, Rolf Backofen, and S. Cenk Sahinalp.

**Fast prediction of rna-rna interaction.**
*Algorithms Mol Biol*, 5 :5, 2010.

M. Sarver, C. Zirbel, J. Stombaugh, A. Mokdad, and N. B. Leontis.

**FR3D : Finding local and composite recurrent structural motifs in RNA 3D.**
*Journal of Mathematical Biology*, 56(1–2) :215–252, January 2008.