M2 - STL

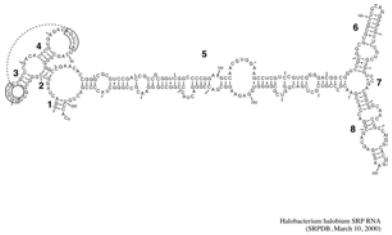
Algorithmes sur les séquences en bioinformatique

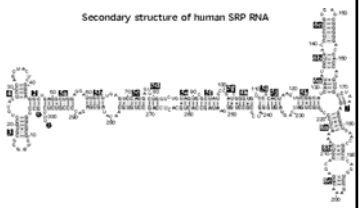
Cours 6 – 2ème partie : Algorithmes de prédiction des structures des ARN

Alessandra Carbone
Université Pierre et Marie Curie

L'ARN est un polymère constitué par 4 unités nucléotidiques A,C, G, U (adénine, cytosine, guanine et uracile)

Le U remplace le T dans l'ADN.





Les rybozymes

L'ARN est la seule macromolécule capable à la fois de transmettre de l'information génétique (il est génome dans de nombreux virus dont le HIV du SIDA), mais également d'effectuer des réactions de catalyse tels que synthèse peptidique et réaction de coupure/ligation des polynucléotides.

Un ribozyme est un ARN possédant une activité catalytique (Ceck et Altman, au début des années 1980) : ils accélèrent les réactions chimiques dans la cellule.

Exemples:

-ARN auto-épissant dans des espèces fungiques (excision des introns, ligation et production d'un ARN maturé)
-ARN auto-coupant dans les viroïdes des plantes
-Hepatitis Delta Virus
-etc.

Le ribosome

Agrégat complexe de trois acides ribonucléiques de longueur différente (~120, 1500 et 1300 nucléotides) et d'une cinquantaine de protéines chez les bactéries, est le siège de la synthèse des protéines.

Cette structure a été déterminée a 5Å de résolution malgré sa taille: 270.000 atomes et 2.MD.

Les structures du ribosome et des ses deux unités ont révélé que la synthèse protéique s'effectuaient dans une région dépourvue de protéines ribosomiques. Cela implique que l'ARN ribosomique est le moteur catalytique de la synthèse protéique:

« Le ribosome est un ribosyme » (Tom Steitz et al., 2000)

Ces découvertes confortent l'hypothèse d'un monde de l'ARN à l'origine de la vie, il y a quelques 3.6-3.8 milliard d'années.

+ le rôle de l'ARN dans l'épissage des introns des gènes eucaryotes
+ les ARN non-codants, qui interviennent dans des étapes successifs à l'épissage (comme l'ARN du contrôle de l'expression du chromosome X)
+ le rôle joue par l'ARN dans la dynamique de l'évolution des génomes ou la duplication d'éléments n'est pas rare : ARN messagers peuvent être re-introduits dans le génome donnant lieu à des gènes actifs ou inactifs (les pseudo-gènes).

L'ADN est un ARN modifié : chimiquement, l'ADN possède un atome d'oxygène de mois que l'ARN.

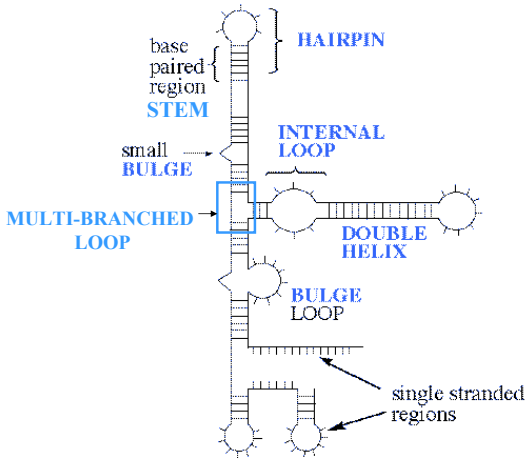
Les structures secondaires et tertiaires

A = U G ≡ C

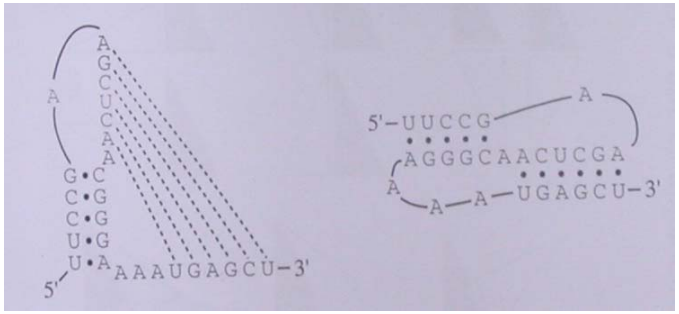
Les paires de bases sont presque co-planaires et elles forment presque tout le temps des piles avec d'autres paires de bases, qui on appelle tiges.

En 3D les tiges forment des doubles hélices.

Les structures constituées par des bases complémentaires sont appelées structures secondaires.

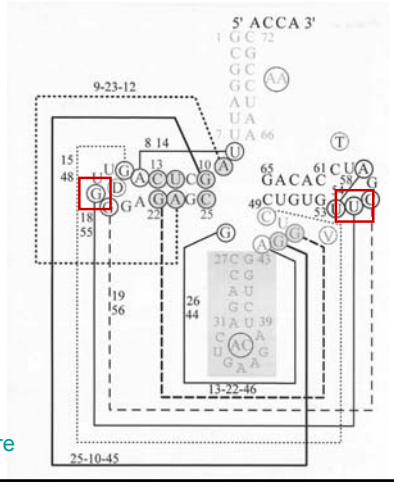


Pseudo-noeuds



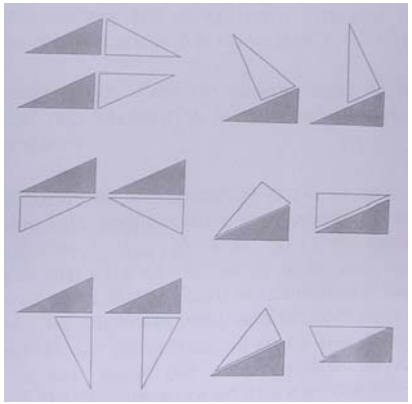
Les appariements non-canoniques

G-U : elles sont thermodynamiquement ainsi favorable que les appariements Crick-Watson.



tARN-phe de la levure

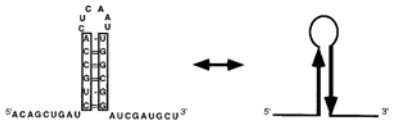
Il y a aussi d'autres appariement….



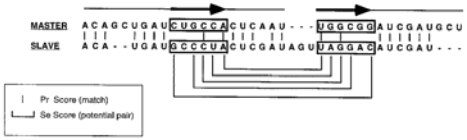
(Nomenclature de Leontis et Westhof, 2000)

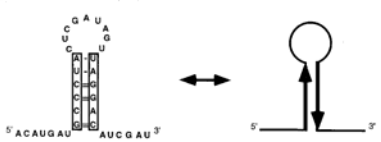
Inférence de la structure par comparaison des séquences

On trouve plusieurs exemples de séquences homologues qui ne sont pas similaires, mais telles que leurs structures secondaires sont identiques.



Mais on trouve aussi plusieurs séquences similaires à mutation corrélées.





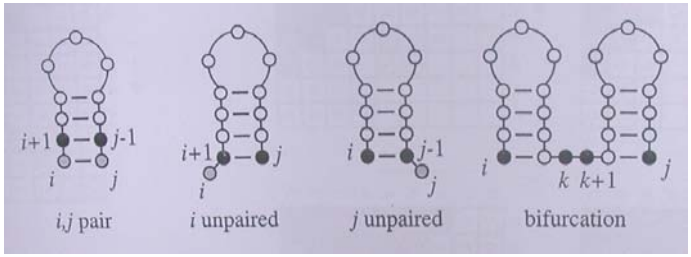
Prédiction de la structure secondaire

Pour un ARN de 200 bases il y a ~10^{50} structures d'appariage possibles. Il faut retrouver la structure qui est biologiquement significative.

L'algorithme de Nussinov (1978)

Algorithme de programmation dynamique. Même si les critères utilises par cet algorithme sont trop simplistes, l'idée est la même pour les algorithmes de repliage qui minimisent l'énergie et pour les algorithmes basés sur les grammaires probabilistes.

Le calcul de Nussinov est récursif : il calcule la structure préférée pour des petites sous-séquences et il étende le résultat à des séquences plus larges.



Idée: il y a seulement 4 façon pour détecter la meilleure structure pour i,,j à partir de la meilleure structure de sous-séquences plus courtes:

1. Ajouter le couple i,,j à la meilleure structure pour la sous-sequence i+1,j-1
2. Ajouter la position non-appariée i à la meilleure structure pour la sous-séquence i+1,j
3. Ajouter la position non-appariée j à la meilleure structure pour la sous-séquence i,j+1
4. Combiner deux sous-structures optimales i,k et k+1,j.

Soit x une séquence de longueur L avec symboles $x_1 \ldots x_L$. Soit $\delta(i,,j)=1$ si x_i et x_j sont des paires de bases complémentaires; sinon $\delta(i,,j)=0$.

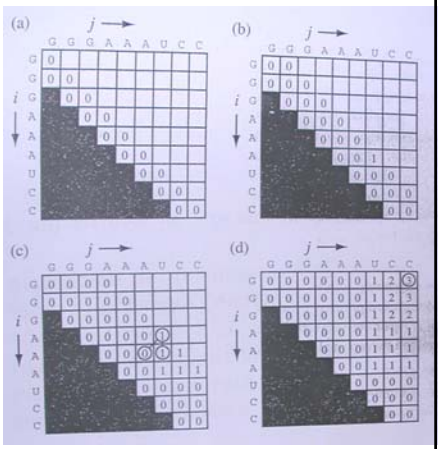
L'algorithme calcule récursivement des scores $\gamma(i,j)$ représentant le nombre maximale de paires de bases qui peuvent être formées pour la sous-séquence $x_i \ldots x_j$.

Algorithme de remplissage de la table de scores:

Base: $\gamma(i,i-1)=0$ pour i=2 a L
$\gamma(i,i) =0$ pour i=1 a L

Recursion: on commence avec toutes sous-séquences de longueur 2, et on les étend jusqu'à la longueur L

$$\gamma(i,j)= \max \begin{cases} \gamma(i+1,j) \\ \gamma(i,j-1) \\ \gamma(i+1,j-1)+(i,j) \\ \max_{1<k<j} [\gamma(i,k)+ \gamma(k+1,j)] \end{cases}$$

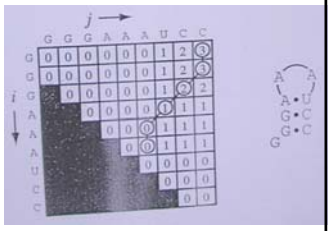


Algorithme de back-tracking de la table de scores:

Initialisation: insérer (1,L) dans la pile (avec l'instruction push(1,L))

Recursion : répéter jusqu'à quand la pile est vide:

-pop(i,j)
-Si i>=j continue;
sinon si $\gamma(i+1,j)= \gamma(i,j)$ alors push(i+1,j);
sinon si $\gamma(i+1,j-1)= \gamma(i,j)$ alors push(i,j-1)
sinon si $\gamma(i+1,j-1)+ \delta_{i,j} = \gamma(i,,j)$:
- mémorise la paire de bases i,,j
- push (i+1,,j-1)
sinon pour k=i+1 à j-1: si $\gamma(i,k)+ \gamma(k+1,j)= \gamma(i,j)$:
- push(k+1,j)
- push(i,k)
- termine



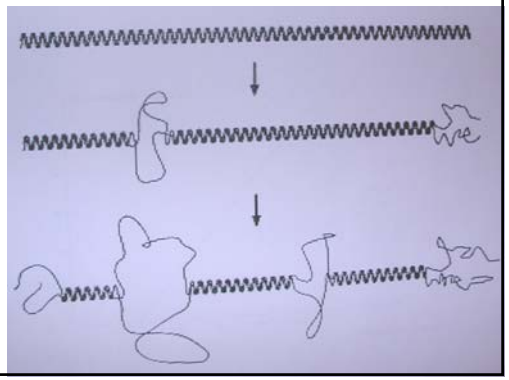
Algorithme de Zucker (1981-1989)

http://www.bioinfo.rpi.edu/applications/mfold/

Algorithme de minimisation de l'énergie associée au polymère (ARN)

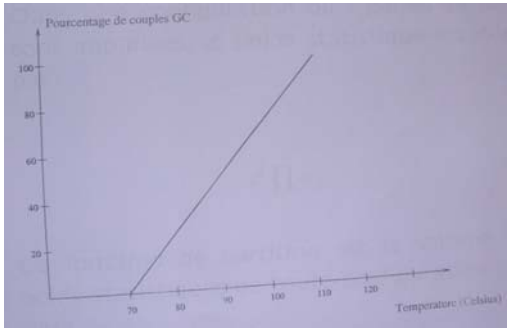
Hypothèse de travail : la structure biologiquement correcte est celle ayant la plus petite énergie libre d'équilibre.

Modèle de mécanique statistique:



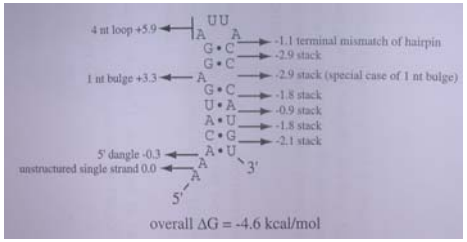
La transition de la double hélice au couple de brins dénaturés demande un grand nombre d'états intermédiaires: ces états dépendent de la séquence en jeu.

A = U G ≡ C



L'énergie libre d'équilibre d'une structure secondaire d'ARN est approximée par la somme des contributions individuelles des boucles, des paires de bases, et des éléments de la structure secondaire (hairpin loops length, bulge loop length, multi-branch loop length, nucléotides solitaires, nucléotides non-appariés terminaux)

Une différence importante avec l'algorithme de Nussinov est que les énergies des tiges sont calculées en ajoutant des facteurs correspondant à des paires de bases empilées (stacks)



Cette modification corresponde à une approximation des données expérimentales plus appropriée mais elle complique l'algorithme de programmation dynamique.

Algorithme de Isambert et Siggia pour étudier la **dynamique** du repliement: applications au ribozyme Hepatitis Delta Virus

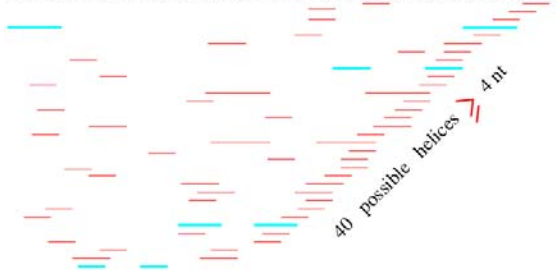
Nouvelles hypothèses :

-le repliage est séquentiel et suit le cours de la transcription.
- les pseudo-nœuds sont considérés

The RNA Folding Problem

Example : tRNA 76 nucleotides

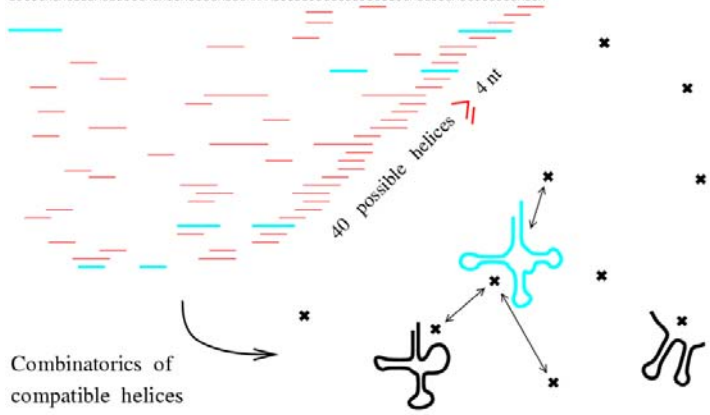
GCGCGAUAGCUCAGUCGGUAGAGCAGGGGAUUGAAAAUCCCCGUGUCCUUGGUUCGAUUCCGAGUCCGCGCACCA



The RNA Folding Problem

Example : tRNA 76 nucleotides

GCGCGAUAGCUCAGUCGGUAGAGCAGGGGAUUGAAAAUCCCCGUGUCCUUGGUUCGAUUCCGAGUCCGCGCACCA



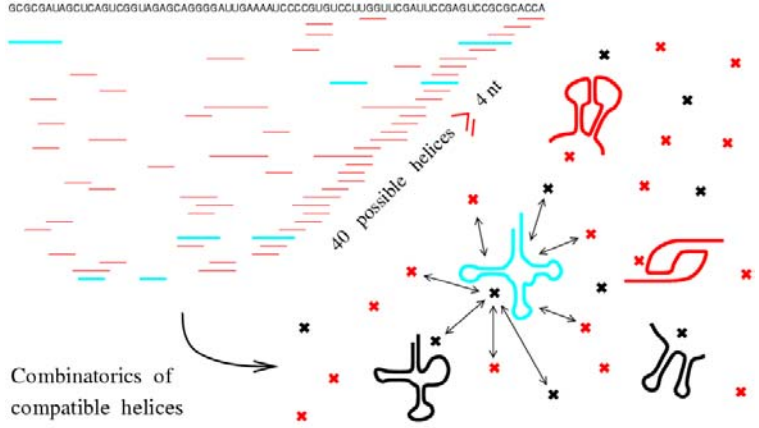
Combinatorics of compatible helices

LARGE Structural Space...

The RNA Folding Problem

Example : tRNA 76 nucleotides

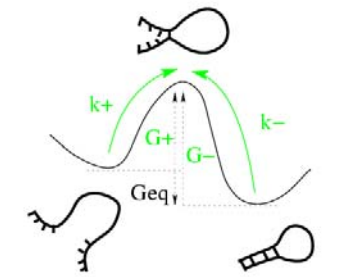
GCGCGAUAGCUCAGUCGGUAGAGCAGGGGAUUGAAAAUCCCCGUGUCCUUGGUUCGAUUCCGAGUCCGCGCACCA



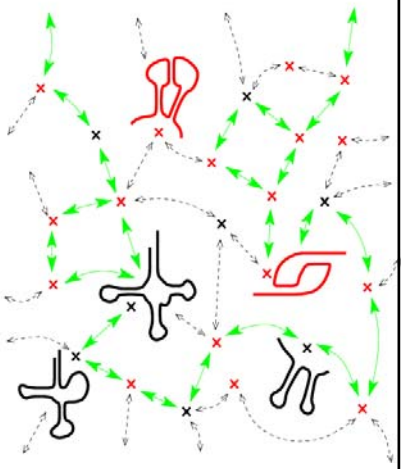
Combinatorics of compatible helices

HUGE Structural Space with Pseudoknots !!

RNA Folding Kinetics

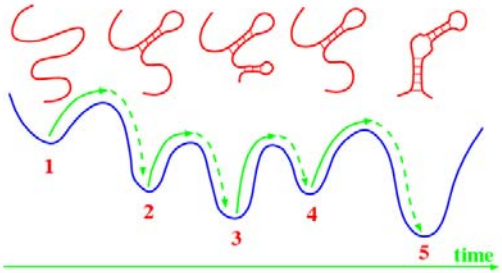


$$k+ = k_0 e^{-\frac{G+}{kT}} \qquad k- = k_0 e^{-\frac{G-}{kT}}$$



Modeling RNA Folding Kinetics

by following the *stochastic* formation and dissociation of *entire* helices



At each configuration along the folding path :

- **calculate *all* rates :** $k_i = k^o \exp(-\Delta G_i / kT)$
- **choose *one transition* stochastically :**

k_1 | k_2 | k_3 ...

$1/\tau = \sum_i k_i$

Le chemin de repliement est prédit par un algorithme de Monte-Carlo cinétique, qui a pour but de créer ou détruire une tige à chaque étape, en changeant la topologie de la structure.

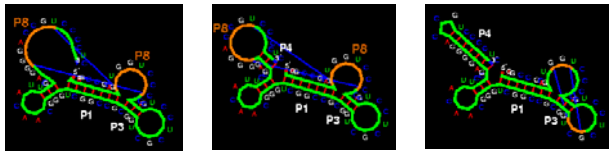
Pour chaque structure obtenue pendant le chemin de repliement :

- Optimisation des frontières entre tiges en compétition
- Calcul de tous les taux de transition
- sélection de la transition suivante avec une probabilité proportionnelle au taux

Cette procédure est re-itérée jusqu'a la distribution moyenne de temps des configurations est stationnaire.

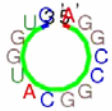
Algorithme pour la détection de structures secondaires avec pseudo-nœuds (Isambert)

KINEFOLD http://kinefold.curie.fr



Decoding RNA Folding Paths: Proposed stability exchange between two competing helices forming sequentially during transcription of hepatitis delta virus ribozyme. The strong, yet transient helix P8 guides the nucleation of P4.

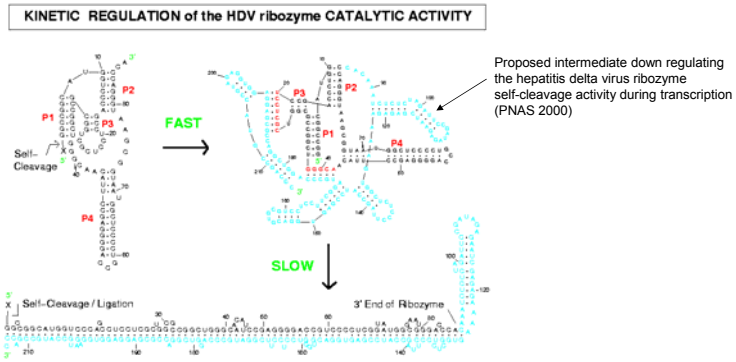
ghdv_123 Generated by KineFold



0.0 kcal/mol reached after 0.000 ms, 3'-end : 14 over 88 bases

Application au ribozyme Hepatitis Delta Virus (HDV)

Il s'agit d'un virus auto-coupant. La structure catalytique est bien connue, mais on ne connaît rien sur le chemin de repliement et sur le repliement du ribozyme une fois qu'il se colle parfaitement à une région de l'HDV.



Quand on suppose que la molécule est complètement dénaturisée et qu'on lui permet de se replier à partir de n'importe quelle région de la séquence, il y a a peu près 1/3 des molécules qui trouvent très facilement la structure native, mais la grande partie des molécules entrent dans des configurations intermédiaires pour quelques minutes avant de se libérer et rentrer dans la configuration native.

Quand on suppose que le repliement suit le processus séquentiel de formation de la molécule d'ARN pendant la transcription, on trouve que presque toutes les molécules entrent dans la configuration native au premier coup.

Le fait que le repliement soit plus efficace pendant la transcription, par rapport au processus qui démarre a partir d'une molécule déjà formée, implique que la séquence code

- la structure, mais aussi
- le chemin de repliement le plus efficace (pendant la transcription)

Extension de l'algorithme de Zucker aux structures avec pseudo-nœuds (Rivas et Eddy, 1999)

Des nouveaux paramètres qui décrivent la stabilité thermodynamique des pseudo-nœuds sont ajoutés.

L'ajoute de contraintes géométriques entre double hélices et brins simples ne permet pas une division en bloques des poids statistiques indépendants.

Complexité de $O(n^6)$, ou n est la taille de la séquence.
(La complexité de l'algorithme de Zucker est de $O(n^3)$.)

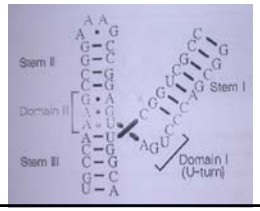
Jamais de prédictions ont été proposées avec ce modèle.

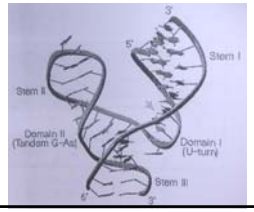
Nouvelles hypothèses?

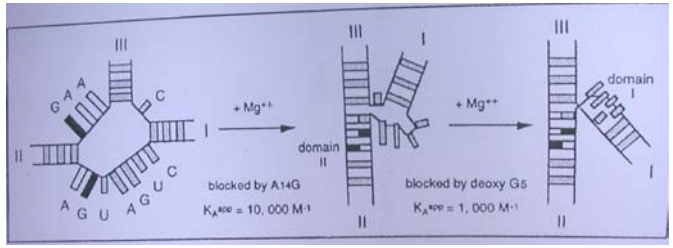
L'ARN a une forte composante électrostatique dans laquelle les ions métalliques jouent un rôle important.

Exemple (D.Lilley 1999): le ribozyme « tête de marteau » (hammerhead ribozyme). La forme géométrique de ce ribozyme est fortement dépendante de la présence de ions métalliques dans la solution.

Il y a deux états de transition pendant le repliage en présence de ions.







Et la structure tertiaire?

On est encore loin….

Quelques références bibliographiques

R.Durbin, S.Eddy, A.Krogh, G.Mitchison, *Biological Sequence Analysis, probabilistic models of proteins and nucleic acids*, Cambridge University Press, 2000.
(vous trouverez ici les références aux autres articles cités)

Isambert H, Siggia E., Modeling RNA folding paths with pseudo-knots: application to hepatitis delta virus ribosyme, *PNAS* 97, 6515-6520, 2000.

O. Perriquet , H. Touzet and M. Dauchet, Finding the common structure shared by two homologous RNAs, *Bioinformatics*, 19:1, 108–116,2003.

Christian Haslinger, Peter F. Stadler, RNA Structures with Pseudo-knots: Graph-theoretical, Combinatorial, and Statistical Properties, *Bulletin of Mathematical Biology* (1999) 61, 437–467.

Cette liste n'est pas représentative de la littérature existante. Elle consiste de qq suggestion de lecture pour démarrer.