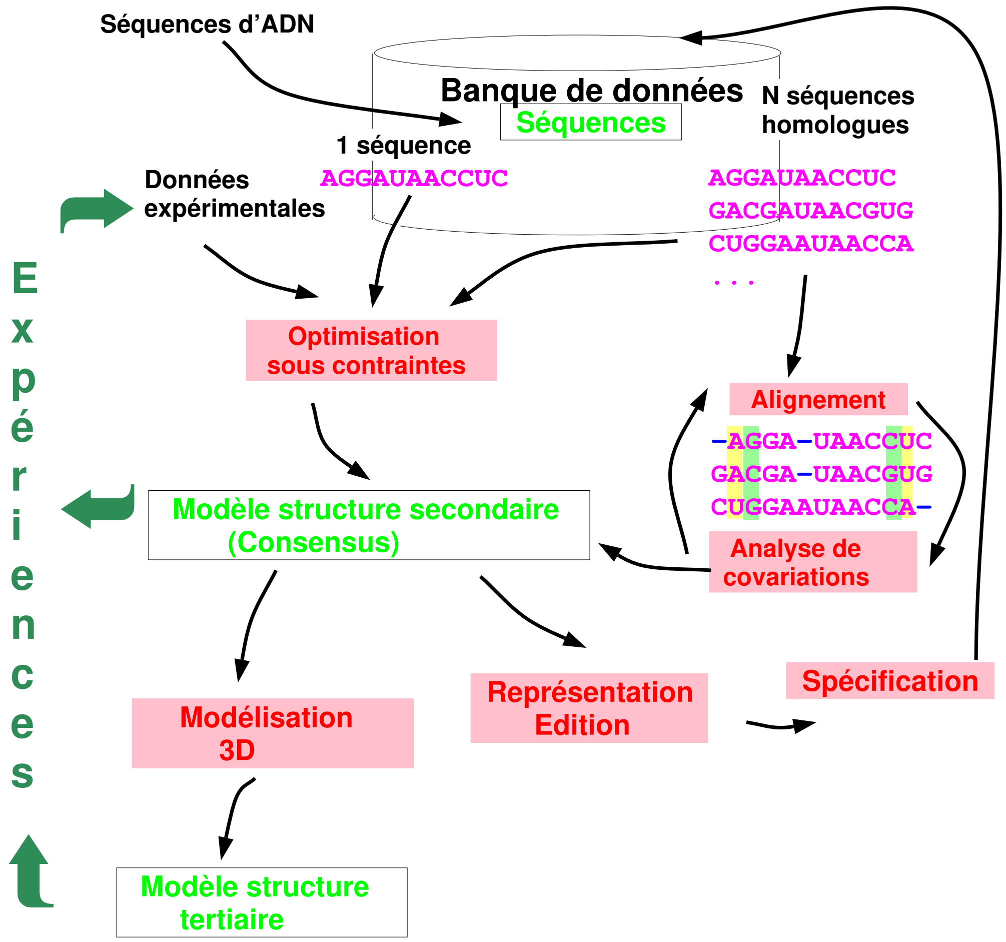
Autour des molécules d'ARN

Introduction aux problèmes et solutions algorithmiques

Christine Gaspin

INRA,Toulouse
email : gaspin@toulouse.inra.fr

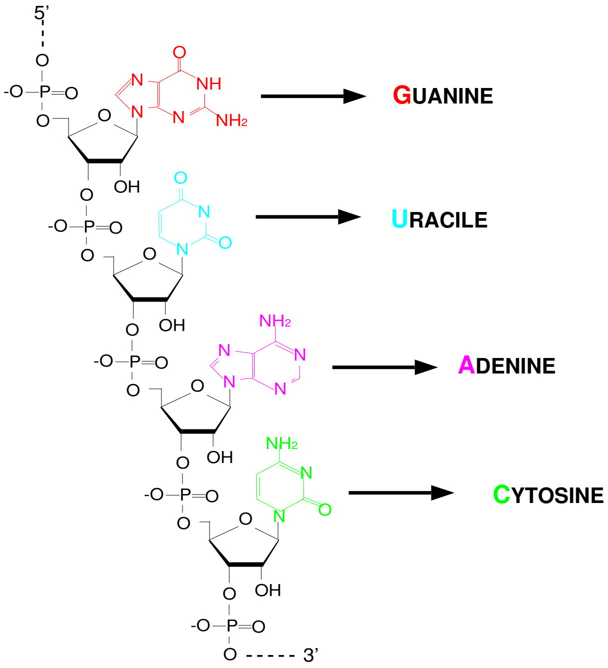
Autour des séquences d'ARN



Qu’est-ce qu’un ARN ?

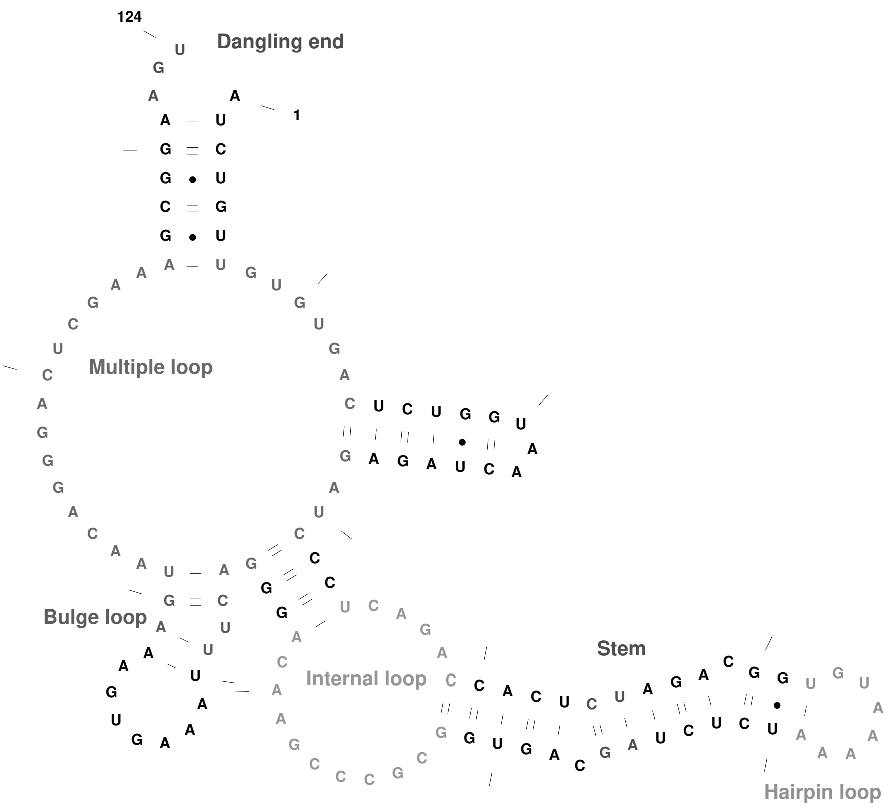
Acide RiboNucléique

Molécule constituée á partir de 4 briques élémentaires: les nucléotides



Représentation et stockage par une suite de lettres caractéristique de la molécule: la séquence ou structure primaire

...G U A C ...



Structures d'ARN-4

Une seule séquence disponible : optimisation sous contraintes

- **Objectif**
  Calculer LA (un sous-ensemble de) structure(s) (sous-)optimale(s) parmi 1.8^N structures possibles

- **Pourquoi**
  Le shéma d'appariements de la structure secondaire optimale est "proche" de celui de la séquence repliée *in vivo*

- **Comment**
  - Recherche arborescente : $O(2^L)$ avec L hélices
  - Heuristiques : recherche "gloutonne" polynomiale en $O(N^2)$
  - Programmation dynamique : $O(N^3)$ à $O(2^N)$ selon la prise en compte des boucles

Programmation dynamique

Hypothèses de base

- Absence de pseudonœuds
- Au plus un appariement par base
- Additivité du modèle énergétique : l'énergie d'une structure est la somme des énergies de ses structures élémentaires

  - Structures élémentaires = paires de bases

  $$E(S) = \sum_{i.j\ in\ S} e(i,j)$$

  - Structures élémentaires = boucles fermées par (i,j)

  $$E(S) = \sum_{i.j\ in\ S} e(L_{i.j})\ +\ e(L_{ext})$$

- Importance du modèle thermodynamique sous-jacent

Structures élémentaires = paires de bases

- Structure optimale : maximise le nombre d'appariements

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

	G	G	A	C	C	U	U	U	G	G	A	C	U	C	
G	**	0	0	1	1	0	0	0	0	0	6	0	0	1	G
G	0	**	0	2	0	0	0	0	5	0	0	2	0	2	G
A	0	0	**	0	0	3	0	0	0	4	0	0	3	0	A
C	1	1	0	**	0	0	0	0	4	4	0	0	11	0	C
C	2	1	0	0	**	0	0	0	5	0	0	10	0	10	C
U	2	1	1	0	0	**	0	0	0	0	6	0	0	9	U
U	2	1	1	0	0	0	**	0	0	0	7	0	0	9	U
U	2	1	1	0	0	0	0	**	0	0	8	0	0	9	U
G	2	2	1	1	1	0	0	0	**	0	0	9	0	9	G
G	2	2	2	2	1	0	0	0	0	**	0	10	0	10	G
A	3	2	2	2	1	1	1	1	0	0	**	0	11	0	A
C	3	3	2	2	2	1	1	1	1	1	0	**	0	0	C
U	3	3	3	3	2	1	1	1	1	1	1	0	**	0	U
C	4	4	3	3	3	2	2	2	2	2	1	0	0	**	C
	G	G	A	C	C	U	U	U	G	G	A	C	U	C	

- Structure optimale : maximise l'énergie libre

$$W[i,j] = min \begin{cases} W(i,j-1) \\ W(i+1,j-1) \ + \ e(i,j) \\ min_{i<k<j}(W(i,k-1) \ + W(k,j) \end{cases}$$

Structures élémentaires = boucles fermées par (i,j)

$$E(S) = \sum_{i.j\ in\ S} e(L_{i.j})\ +\ e(L_{ext})$$

Objectif : Prise en compte d'un modèle thermodynamique plus élaboré (nature et nombre de bases dans les différentes boucles)

Deux quantités

- W(i,j) : énergie minimum d'une structure contenant au moins une paire de bases dans la séquence (i..j) et W(i,i)=∞

  $$W[i,j] = min_{i \leq i_1 < j_1 < i_2 < j_2 < \ldots < i_k < j_k \leq j\ ,\ k>0} \{e(L_{ext}) + \sum_{p=1}^{k} V(i_p, j_p)\}$$

- V(i,j) : énergie minimum d'une structure contenant la paire de bases (i,j) et V(i,i)=∞

  $$V[i,j] = min_{i \leq i_1 < j_1 < i_2 < j_2 < \ldots < i_k < j_k \leq j\ ,\ k \leq 0} \{e(L_{i.j}) + \sum_{p=1}^{k} V(i_p, j_p)\}$$

Si on tient compte des boucles externes :

$O(2^{j-i})$ paires à examiner $\Longrightarrow O(2^N)$

Si on ne tient pas compte des boucles externes :

$$W[i,j] = min \begin{cases} W(i+1,j) \\ W(i,j-1) \\ V(i,j) \\ min_{i<k<j}(W(i,k) \ + W(k+1,j) \end{cases}$$

On réécrit $V[i,j]$

$$V[i,j] = min \begin{cases} VH(i,j) \\ es(i,j) + V(i+1,j-1) \\ VBI(i,j) \\ VM(i,j) \end{cases}$$

Selon les choix réalisés pour la prise en compte des boucles, on aura

Paires empilées	$O(N^2)$
Épingles à cheveux	$O(N^3)$
Renflements/Boucles internes	$O(N^5)$ à $O(N^2)$
Boucles multiples	$O(2^N)$ à $O(N^3)$

- **Problèmes**

  - Hypothèses : additivité du modèle thermodynamique, absence de pseudonœud)

  - Nécessité de plus de souplesse dans l'exploitation de l'espace des structures sous-optimales (Interactivité, jeux de contraintes disponible, visualisation plus claire des choix...)

  - Nécessité d'intégrer des données expérimentales

- **Quelques références**

- *M. Huynen et al.*, **Base pairing probabilities in a complete HIV-1 RNA**, Journal of computational biology, Vol. 3, No. 2, pp. 253-274, 1996.

- *C. Gaspin & E. Westhof*, **An interactive framework for RNA secondary structure prediction with a dynamical treatment of constraints**, J. Mol. Biol., Vol. 254, pp. 163-174, 1995.

- *J. S. Mc Caskill*, **The equilibrium partition function and base pair binding probabilities for RNA secondary structure**, Biopolymers, Vol. 29, 1105-1119, 1990.

- *M. Zuker*, **The use of dynamic programming algorithms in RNA secondary structure prediction**, in Mathematical Methods for DNA sequences, chapter 7, 1989.

- *M. Gouy*, **Secondary structure prediction of RNA**, in Nucleic acids and protein sequence analysis, a practical approach, chapter 11, 1987.

Alignement de séquences d'ARN

- **Objectif**

  Prédiction de structure consensus, phylogénie.

- **Pourquoi**

  - Séquences homologues : même fonction
  - Hypothèse : fonction et structure liées

- **Comment**

  Complexité en $O(L^N)$

  L: longueur de la séquence ; N: nombre de séquences

  - Alignement multiple global en séquence

  $$\begin{cases} -Alignement\ par\ paires\ de\ séquences \\ et\ classification\ hiérarchique \end{cases}$$

  - Alignement multiple global en séquence et structure

  $$\begin{cases} -Approche\ probabiliste \\ -Théorie\ des\ graphes \end{cases}$$

  - Alignement multiple local en séquence et structure

  $$\begin{cases} -Programmation\ dynamique \end{cases}$$

  - Alignement interactif

- **Problèmes**

  - Alignement multiple global en séquence pas assez informatif

  - Approche probabiliste : pas de pseudonœud, limitation par la taille et le nombre des séquences

  - Approche globale ne voit pas les similitudes locales

- **Quelques références**

- *J. Gorodkin et al.*, **Finding the most significant common sequence and structure motifs in a set of RNA sequences**, NAR, Vol. 25, No. 18, pp. 3724-3732, 1997.

- *J. E. Tabaska & G. D. Stormo*, **Automated alignment of RNA sequences to pseudoknotted structures**, ISMB97, pp. 311-318, 1997.

- *S. Eddy & R. Durbin*, **RNA sequence analysis using covariance models**, NAR, Vol. 22, No. 11, pp. 2079-2088, 1994.

-*Y. Sakakibara et al.*, **Stochastic context-free grammars for modeling RNA**, Hawaii international conference on system sciences, pp. 284-293, 1994.

- *J.D. Thompson, D.G. Higgins*, **CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**, CABIOS, Vol. 22, pp. 4673-4680, 1994.

- *M. Waterman*, **Consensus methods for folding single-stranded nucleic acids**, in Mathematical Methods for DNA sequences, chapter 8, 1989.

Analyse de covariations

- **Objectif**

  Identifier des couples de positions dont les variations sont concertées

- **Pourquoi**

  Interprétation des mutations compensées en terme de contraintes structurales

- **Comment**

  - À partir d'un alignement : Comptabiliser les effectifs

  - Identifier les couples qui covarient : test statistique sur une mesure de l'information mutuelle

  - Interpréter les résultats

$$\left\{\begin{array}{l}
-\textit{Test phylogénétique} \\
(\textit{voir E. Westhof et F. Michel } 1994) \\
-\textit{Validation structurale des couples identifiés} \\
(\textit{voir E. Westhof et F. Michel } 1994 \textit{ et Gautheret et al. } 1995) \\
-\textit{Validation expérimentale}
\end{array}\right.$$

- **Problèmes**

  - Contacts tertiaires plus diversifiés donc plus difficiles à mettre en évidence

  - Forte dépendance à l'alignement

  - Alignement bon $\Longrightarrow$ structure secondaire retrouvée

- **Quelques références**

D. Gautheret & R. Gutell, **Inferring the conformation of RNA base pairs and triples from patterns of sequence variation**, NAR, Vol. 25, No. 8, pp. 1559-1564, 1997.

R.R. Gutell et al., **Identifying constraints on the higher-order structure of RNA: continued development and application of comparative analysis methods**, NAR, Vol. 20, No. 21, pp. 5785-5795, 1992.

E. Westhof & F. Michel, **Prediction and experimental investigation of RNA secondary and tertiary foldings**, in RNA-Protein interactions: frontiers in molecular biology, pp. 25-51, 1994.

D. K. Y. Chiu & T. Kolodziejczak, **Inferring consensus structure from nucleic acid sequences**, CABIOS, Vol. 7, No. 3, pp. 347-352, 1991.

Représentation et édition de structures secondaires d'ARN

- **Objectif**

  Représentation claire et *utile* d'une structure secondaire.

- **Pourquoi**

  Exploitation des éléments de structure secondaire, mise en évidence des éléments importants, publication.

- **Comment**

  - Circles,
  - Domes,
  - Mountains,
  - Dotplots,
  - Représentation conventionnelle.

Voir ces différentes représentations dans GCG et les différentes références données

- **Problèmes non complètement résolus**

  - Calcul automatique d'une représentation sans recouvrement
  - Calcul automatique d'une représentation avec pseudonœuds

- **Quelques références récentes**

- *P. De Rijk & R. De Wachter*, **RnaViz, a program for the visualisation of RNA secondary structure**, Nucleic Acids Research, Vol. 25, No. 22, pp. 4679-4684, 1997.

- *Chetouani et al.*, **ESSA: an integrated and interactive computer tool for analysing RNA scondary structures**, Nucleic Acids Research, Vol. 25, No. 17, pp. 3514-3522, 1997.

- *O. Matzura & A. Wennborg*, **RNAdraw: an integrated program for RNA secondary structure calculation and analysis under 32-bit microsoft windows**, CABIOS, Vol. 12, No. 3, pp. 247-249, 1996.

Spécification de structure secondaire d'ARN

- **Objectif**

  Identification des séquences génomiques

- **Pourquoi**

  Les motifs conservés en séquence et en structure constituent une empreinte de la molécule.

- **Comment**

  1. Élaboration de l'empreinte
     - A partir d'un alignement correct

       $$\begin{cases} -Codée\ dans\ un\ algorithme \\ -Codée\ dans\ un\ langage\ dédié \\ -Structure\ consensus \end{cases}$$

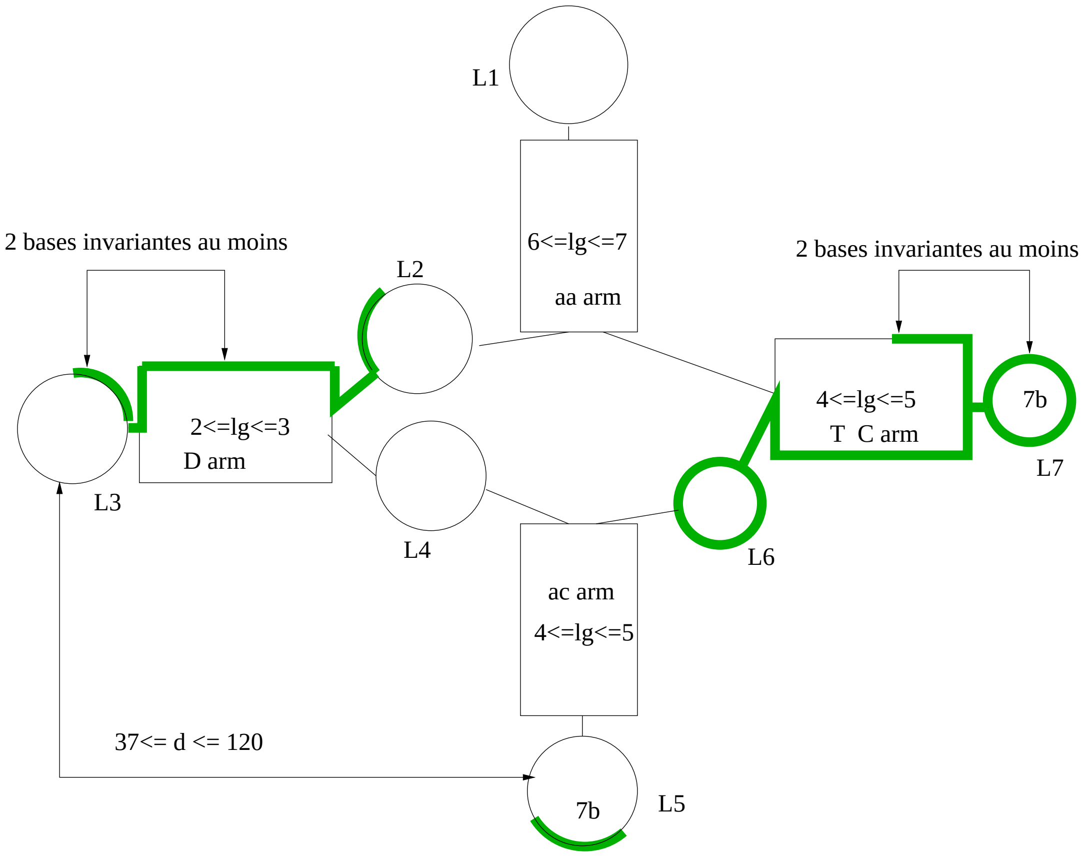
     - Par apprentissage

       $$\begin{cases} -Approches\ probabilistes \\ -Théorie\ des\ graphes \end{cases}$$

  2. Test d'une séquence contre cette empreinte
     - Alignement d'une séquence contre une séquence (banque) de structure connue

       $$\begin{cases} -Recherche\ arborescente \\ -Algorithme\ génétique \\ -Programmation\ dynamique \end{cases}$$

Exemple d'empreinte possible



Pour exemple de matrice de score (portions en vert sur la figure), voir les références.

- **Problèmes**

  - Coût des méthodes d'apprentissage lié au nombre et à la taille des séquences

  - Sensibilité dépendante de l'alignement et du nombre de séquences disponibles

  - Manque de sensibilité des méthodes "manuelles"

- **Quelques références**

- *C. Notredame et al.*, **RAGA: RNA sequence alignment by genetic algorithm**, NAR, Vol. 25, No. 22, pp. 4570-4580, 1997.

- *T. Lowe & S. Eddy*, **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**, NAR, Vol. 25, No. 5, pp. 955-964, 1997.

- *Billoud et al.*, **Palingol: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence databases**, NAR, Vol. 24, No. 8, pp. 1395-1403, 1996.

- *N. El Mabrouk & F. Lisacek*, **Very fast identification of RNA motifs in genomic DNA. Application to tRNA search in the *Yeast* genome**, J. Mol. Biol., Vol. 264, pp. 46-55, 1996.

- *T. Dandekar and M. Hentze*, **Finding the hairpin in the haystack: searching for RNA motifs**, TIG, Vol. 11, No. 2, 1995.

- *F. Corpet & B. Michot* **RNAlign program: alignment of RNA sequences using both primary and secondary structures**, CABIOS, Vol. 10, No. 4, pp. 389-399, 1994.

Modélisation de structure tertiaires

- **Objectif**

  Obtenir un modèle atomique

- **Pourquoi**

  Compréhension des relations structure-fonction liée à l'existence de la structure ou d'un modèle.

- **Comment**

  - Approche automatique

$$\begin{cases} -\textit{Géométrie des distances} \\ -\textit{Satisfaction de contraintes} \end{cases}$$

  - Approche manuelle assistée

$$\begin{cases} -\textit{Utilisation des structures obtenues par} \\ \textit{RMN ou cristallographie} \\ -\textit{Mécanique moléculaire} \\ -\textit{Longue expérience de la modélisation} \end{cases}$$

- **Problèmes**

  - Géométrie des distances : chiralité, présence de nœuds difficiles à détecter.
  - Satisfaction de contraintes : combinatoire des solutions
  - Manuelle assistée : biais de l'oeil humain

- **Quelques références**

- E. Westhof et al., **RNA tectonics: towards RNA design**, Folding and design, Vol. 1, pp. 78-88, 1996.

- *J. M. Hubbard & J. E. Hearst*, **Predicting the three-dimensional folding of transfer RNA with a computer modeling method**, Biochemistry, Vol. 30, pp. 5438-5475, 1991.

- *F. Major et al.*, **The combination of symbolic and numerical computation for three-dimensional modeling of RNA**, Science, Vol. 253, pp. 1255-1260, 1991.

Conclusion

- Encore beaucoup de travail à réaliser !!!
- Avec le nombre croissant de séquences

  - Identifier rapidement les séquences génomiques comme étant des ARN

  - Importance des programmes permettant d'obtenir rapidement des alignements de qualité
- Importance de la modélisation

  - Aucun programme ne fait actuellement mieux que l'homme

  - Difficulté de la modélisation des boucles
- Bien identifier le problème qu'on veut résoudre
- Importance du dialogue entre les différentes communautés