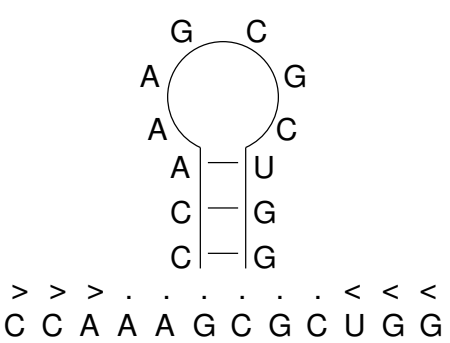
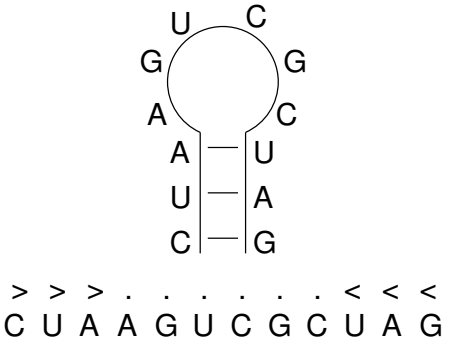
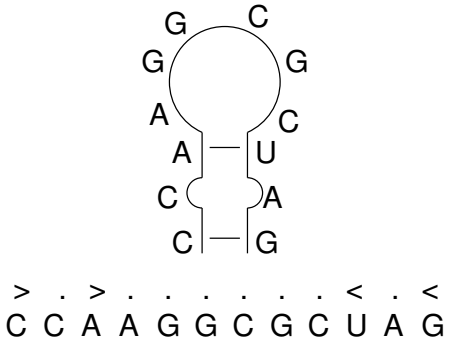
METHODE

Matrices de substitution pour les ARN $\Rightarrow$ structure secondaire







2 sortes de matrices :

– 4×4 pour les boucles

$\mathcal{A} = \{A, C, G, U\}$

– 16×16 pour les tiges

$\mathcal{A} = \{AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, UU\}$

Algorithme de Henikoff-Henikoff

	A	G	G	T	A
	G	G	G	T	G
	A	G	G	T	A
	G	A	G	A	A
A	2	1	0	1	3
C	0	0	0	0	0
G	2	3	4	0	1
T	0	0	0	3	0

$f(a,b)$ nombre d'occurrences d'une substitution de $a \in \mathcal{A}$ en $b \in \mathcal{A}$

A	4			
C	0	0		
G	10	0	10	
T	3	0	0	3
	A	C	G	T

$$P_{obs}(a,b) = \frac{f(a,b)}{\sum_{c,d \in \mathcal{A}} f(c,d)}$$

probabilité observée d'occurrence d'une substitution de $a \in \mathcal{A}$ en $b \in \mathcal{A}$

A	0.13			
C	0	0		
G	0.33	0	0.33	
T	0.1	0	0	0.1
	A	C	G	T

	A	G	G	T	A
	G	G	G	T	G
	A	G	G	T	A
	G	A	G	A	A

A	2	1	0	1	3
C	0	0	0	0	0
G	2	3	4	0	1
T	0	0	0	3	0

$F(a)$ nombre d'occurrences de $a \in \mathcal{A}$

A	7
C	0
G	10
T	3

$$P(a) = \frac{F(a)}{\sum_{b \in \mathcal{A}} F(b)}$$

probabilité observée d'occurrence de $a \in \mathcal{A}$

A	0.35
C	0
G	0.5
T	0.18

$$P_{exp}(a,b) = \begin{cases} P(a)^2 & \text{si } a = b \\ 2 \times P(a) \times P(b) & \text{si } a \neq b \end{cases}$$

probabilité attendue d'occurrence d'une substitution de $a \in \mathcal{A}$ en $b \in \mathcal{A}$

A	0.12			
C	0	0		
G	0.35	0	0.25	
T	0.11	0	0.15	0.02
	A	C	G	T

	A	G	G	T	A
	G	G	G	T	G
	A	G	G	T	A
	G	A	G	A	A

A	2	1	0	1	3
C	0	0	0	0	0
G	2	3	4	0	1
T	0	0	0	3	0

$$S = \left(\log_2 \left(\frac{P_{obs}(a,b)}{P_{exp}(a,b)} \right) \right)_{a,b \in \mathcal{A}}$$

matrice de substitution

A	0.12	$\perp$	-0.07	-0.07
C	$\perp$	$\perp$	$\perp$	$\perp$
G	-0.07	$\perp$	0.42	$-\infty$
T	-0.07	$\perp$	$-\infty$	1.15
	A	C	G	T

$S_{T,G} = -\infty$ car on n'observe aucune substitution de T en G

$S_{C,A}$ est indéterminée, car on n'observe et on n'attend aucune substitution de C en A

Il faut une unique structure secondaire

$\Rightarrow$ structure secondaire "consensus" :

un appariement à chaque endroit où il y en a un dans les structures de départ.

```
structure 1              >>>.>>...<<..>>>...<<<..<<<
structure 2              >>>>.......>>>...<<<.<<<<
structure 3               .>>>>>>...<<..>.>....<.<.<<<.

structure "consensus"    >>>>>>>...<<..>>>...<<<.<<<<
```

RESULTATS

Pour six eucaryotes : *Homo sapiens*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*

matrices sur les groupes de séquences suivants :

- les ARNt correspondant à un même acide aminé
- les ARNt correspondant à un même anticodon et ayant une même structure secondaire

- chaque tige et chaque boucle de tous les ARNt

Même acide aminé

Dans les matrices pour les boucles :

- scores positifs pour les conservations de bases
- scores négatifs pour les substitutions
- scores souvent meilleurs pour les transitions que pour les transversions

Exemple : Homo sapiens, ARNt cystéine.

	A	C	G	U
A	**1.46**	-5.23	-2.81	-4.83
C	-5.23	**2.71**	-3.73	-2.96
G	-2.81	-3.73	**1.82**	-3.10
U	-4.83	-2.96	-3.10	**1.89**

Même acide aminé

Dans les matrices pour les tiges :

- beaucoup de valeurs indéterminées
- les conservations des appariements Watson Crick, GU et UG ont des scores soit positifs, soit $-\infty$, soit indéterminés
- les conservations des autres appariements peuvent avoir de meilleurs scores

Même acide aminé

Dans certaines matrices, les substitutions sont absentes.

organisme	boucles	tiges
D. melanogaster	Asp, **His**, **Phe**, Trp	**His**, **Phe**, Tyr
S. cerevisiae	**Asn**, Asp, **Cys**, His, **SeC**	**Asn**, **Cys**, Tyr, **SeC**
S. pombe	Asn, **Asp**, Cys, **His**, Phe, **Trp**	**Asp**, **His**, **Trp**, Tyr

Même anticodon et même structure secondaire

En ne regardant que les ensembles ayant au moins deux séquences, on remarque que :

- les transversions sont souvent absentes
- les séquences sont souvent identiques dans un même ensemble

organisme	pourcentage d'ensembles où les transversions sont absentes			pourcentage d'ensembles où les séquences sont identiques		
	boucles	tiges	tiges et boucles	boucles	tiges	tiges et boucles
H. sapiens	76%	89%	68%	47%	53%	38%
C. elegans	84%	92%	80%	77%	74%	67%
D. melanogaster	91%	96%	89%	64%	64%	45%
A. thaliana	81%	72%	63%	63%	45%	39%
S. cervisiae	84%	95%	81%	74%	77%	60%
S. pombe	97%	95%	92%	79%	71%	58%

Boucles et tiges

matrices pour les boucles D, $anticodon$, $T\Psi C$ et la boucle $multiple$:

en sommant les valeurs de la diagonale, on ordonne ces matrices

$diag(T\Psi C) > diag(D) > diag(anticodon)$
et $diag(T\Psi C) > diag(D) > diag(multiple)$

Exemple : Homo sapiens

	A	C	G	U
A	**1.28**	-2.33	-1.67	-1.55
C	-2.33	**1.22**	-0.81	0.95
G	-1.67	-0.81	**1.03**	-1.59
U	-1.55	0.95	-1.59	**1.53**

$diag(D) = 5.06$

	A	C	G	U
A	**0.85**	-0.62	0.39	-1.02
C	-0.62	**0.80**	-0.44	-0.33
G	0.39	-0.44	**0.78**	-0.62
U	-1.02	-0.33	-0.62	**0.80**

$diag(anticodon) = 3.23$

	A	C	G	U
A	**1.27**	-2.47	0.44	-1.66
C	-2.47	**2.06**	-2.48	-1.17
G	0.44	-2.48	**1.98**	-1.88
U	-1.66	-1.17	-1.88	**0.92**

$diag(T\Psi C) = 6.23$

	A	C	G	U
A	**0.66**	0.36	0.59	-1.97
C	0.36	**0.32**	0.43	-0.85
G	0.59	0.43	**0.57**	-1.69
U	-1.97	-0.85	-1.69	**0.99**

$diag(multiple) = 2.54$

Travail en cours et perspectives

matrices pour un même organisme sur les groupes de séquences suivants :

- chaque tige et chaque boucle pour les ARNt correspondant à un même acide aminé
- chaque tige et chaque boucle pour tous les ARNt
- tous les ARNt

définir une distance entre matrices

travailler avec d'autres ARN