Génération aléatoire de séquences génomiques, codantes *et* structurées

Yang Zhang[1], **Yann Ponty**[2], Mathieu Blanchette[3], Éric Lecuyer[2] et Jérôme Waldispühl[3]

1 Univ. de Montréal

2 CNRS Ecole Polytechnique
Inria Saclay

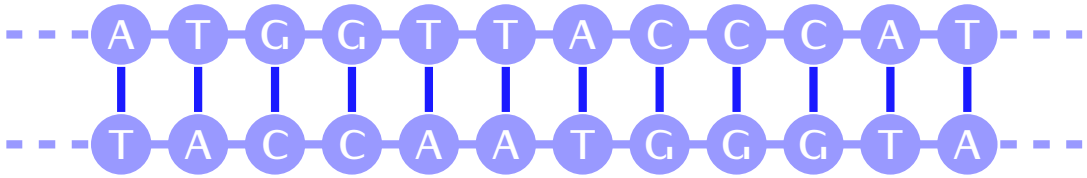
3 Univ. McGill

Juillet 2016

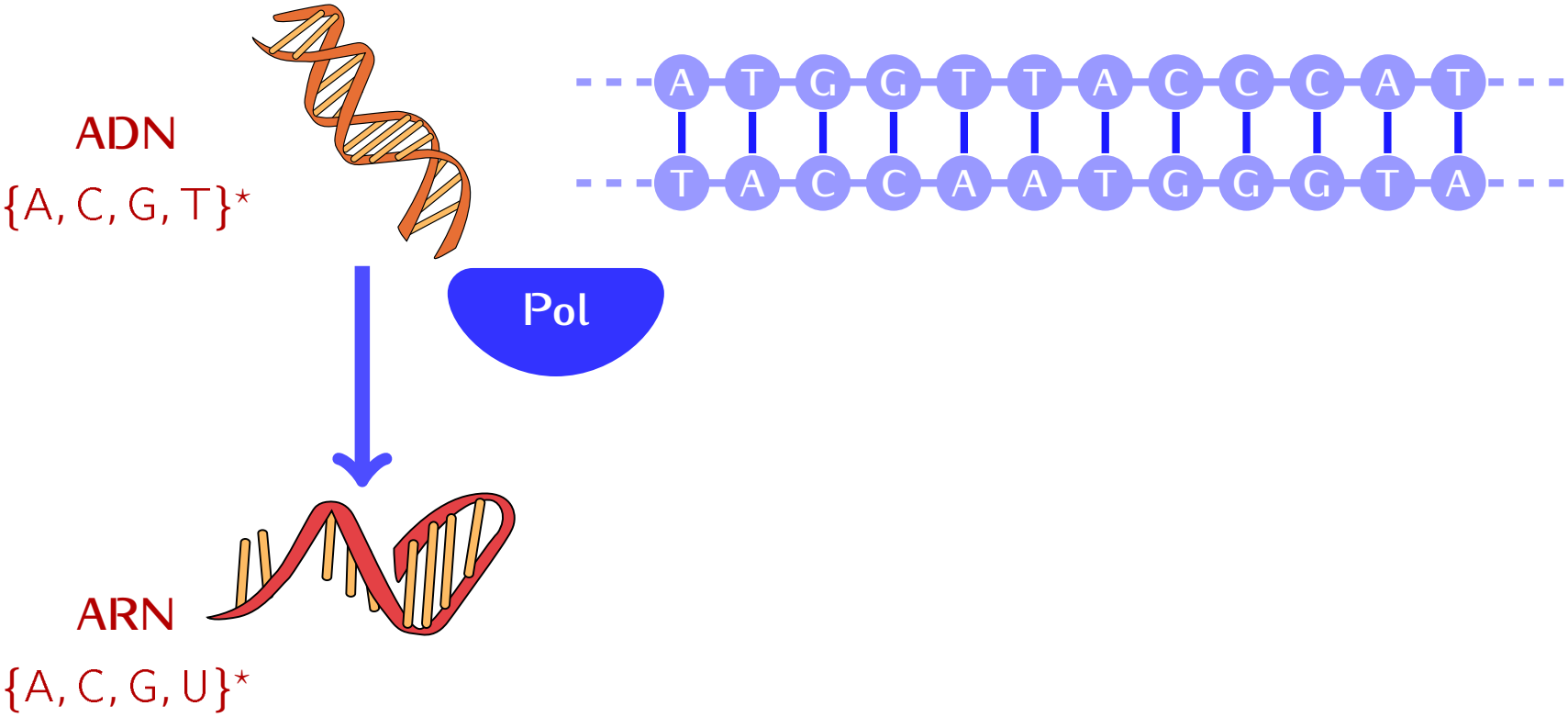
Dogme fondateur de la biologie moléculaire

ADN

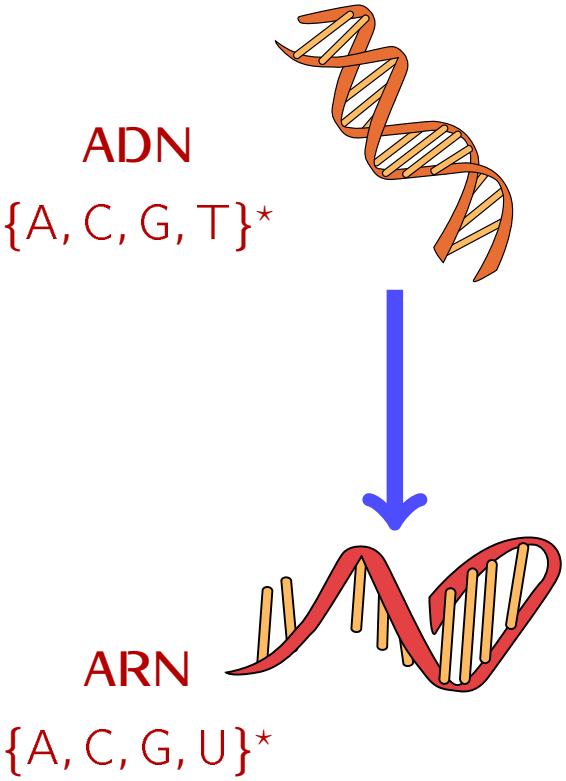
$\{A, C, G, T\}^\star$

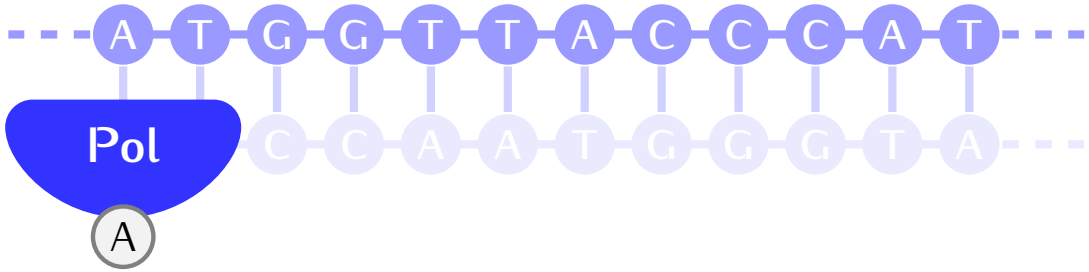


Dogme fondateur de la biologie moléculaire



Dogme fondateur de la biologie moléculaire

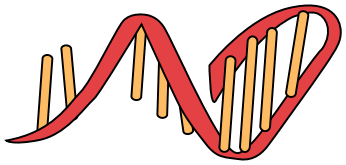




Dogme fondateur de la biologie moléculaire

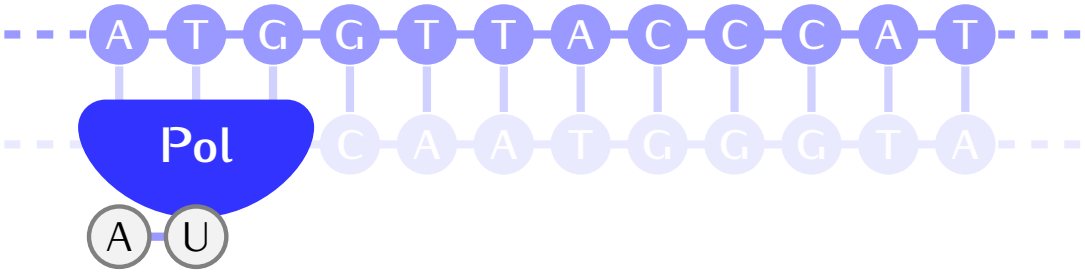
ADN

$\{A, C, G, T\}^*$



ARN

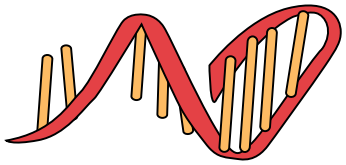
$\{A, C, G, U\}^*$



Dogme fondateur de la biologie moléculaire

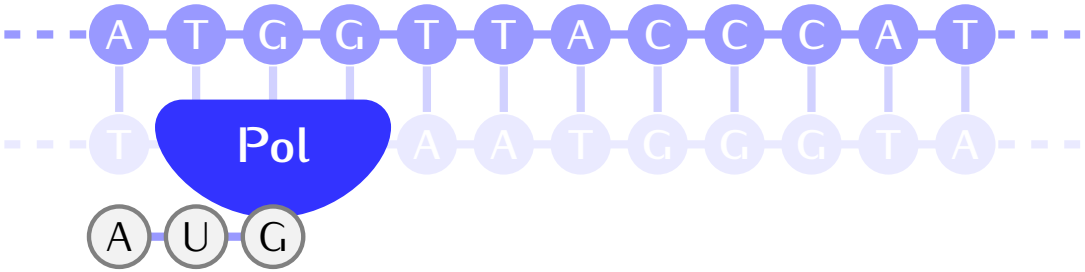
ADN

$\{A, C, G, T\}^*$



ARN

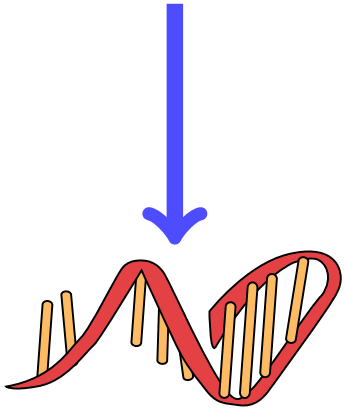
$\{A, C, G, U\}^*$



Dogme fondateur de la biologie moléculaire

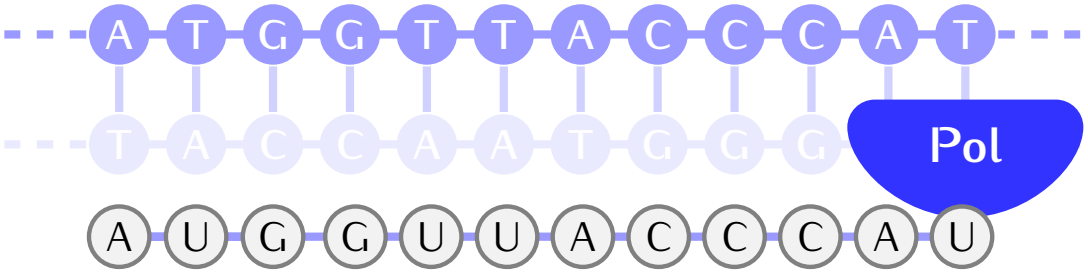
ADN

$\{A, C, G, T\}^*$



ARN

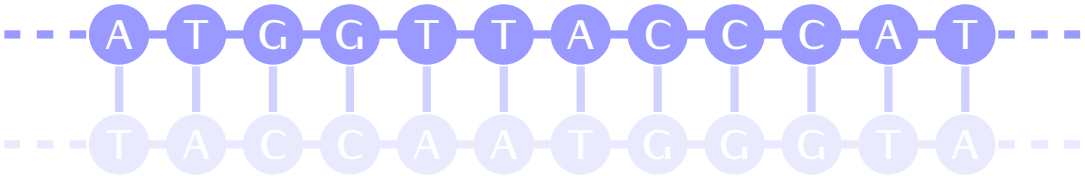
$\{A, C, G, U\}^*$

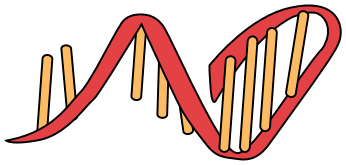


Dogme fondateur de la biologie moléculaire

ADN

$\{A, C, G, T\}^*$





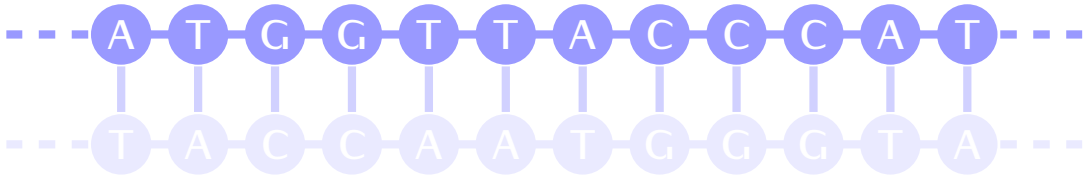
ARN

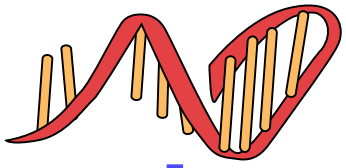
$\{A, C, G, U\}^*$

Dogme fondateur de la biologie moléculaire

ADN

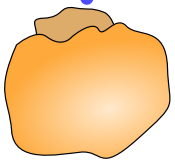
$\{A, C, G, T\}^\star$





ARN

$\{A, C, G, U\}^\star$



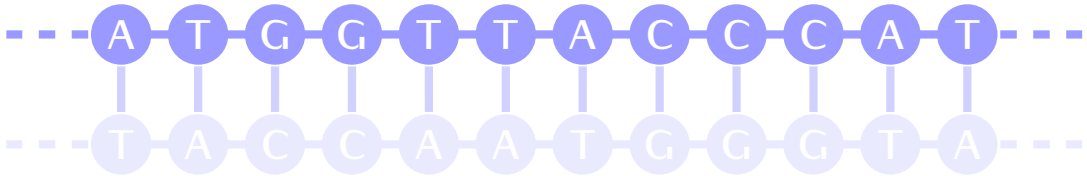
Protéines

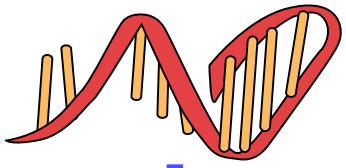
$\underbrace{\{\mathrm{Ala}, \mathrm{Arg}, \ldots, \mathrm{Val}\}^\star}_{20^+ \text{ Acides aminés}}$

Dogme fondateur de la biologie moléculaire

ADN

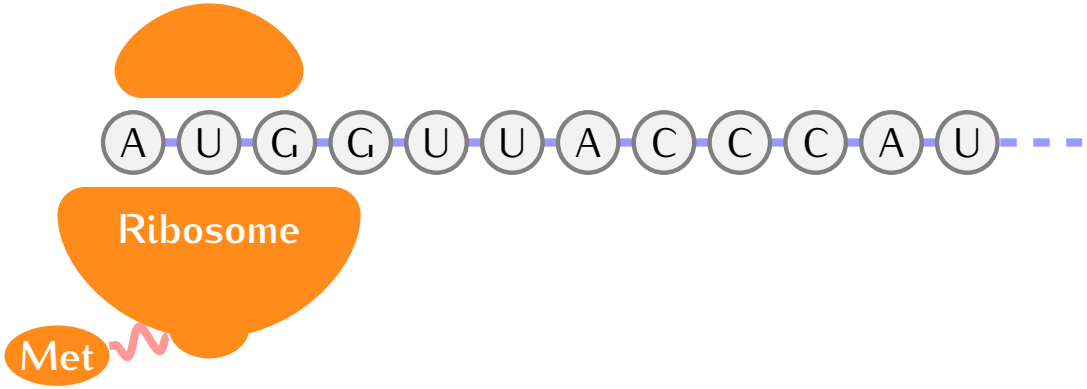
$\{A, C, G, T\}^\star$

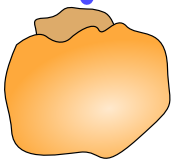




ARN

$\{A, C, G, U\}^\star$





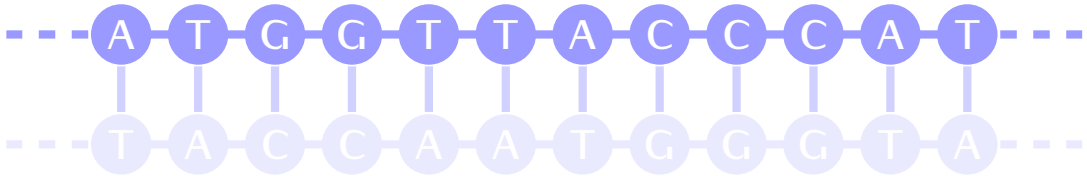
Protéines

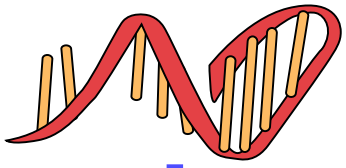
$\underbrace{\{\mathrm{Ala}, \mathrm{Arg}, \ldots, \mathrm{Val}\}^\star}_{20^+ \text{ Acides aminés}}$

Dogme fondateur de la biologie moléculaire

ADN

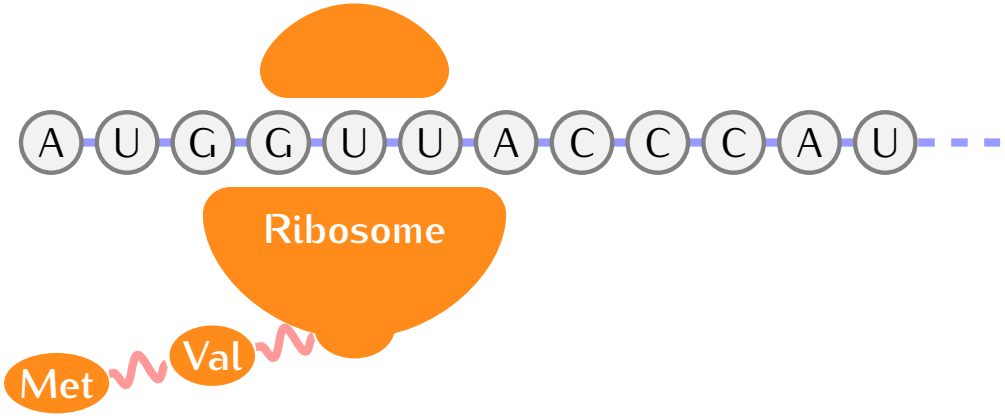
$\{A, C, G, T\}^\star$

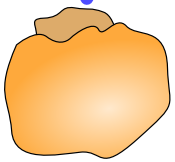




ARN

$\{A, C, G, U\}^\star$





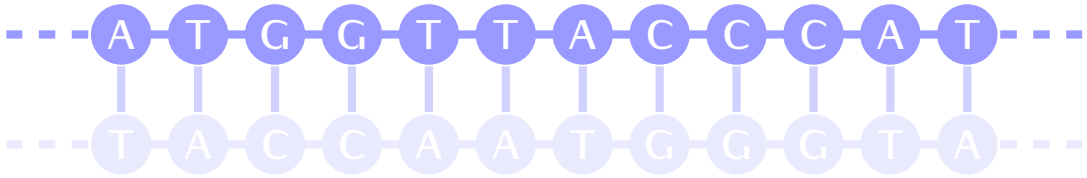
Protéines

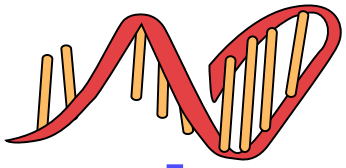
$\underbrace{\{\mathrm{Ala}, \mathrm{Arg}, \ldots, \mathrm{Val}\}^\star}_{20^+ \text{ Acides aminés}}$

Dogme fondateur de la biologie moléculaire

ADN

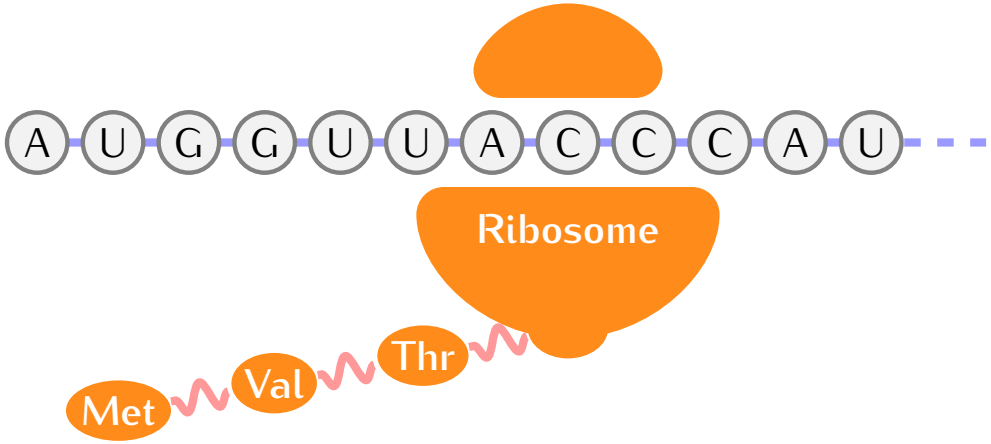
$\{A, C, G, T\}^\star$

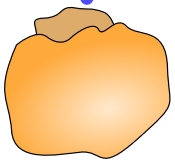




ARN

$\{A, C, G, U\}^\star$





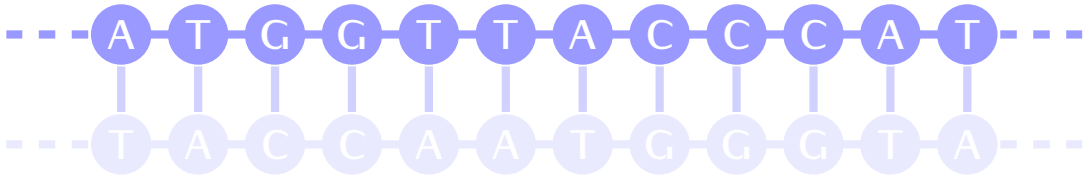
Protéines

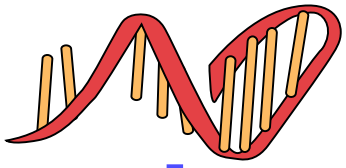
$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Acides aminés}}$

Dogme fondateur de la biologie moléculaire

ADN

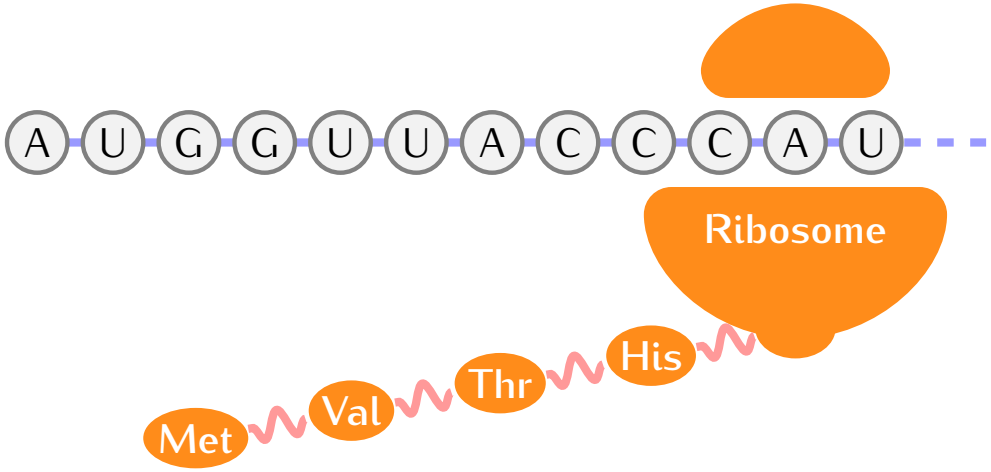
$\{A, C, G, T\}^\star$

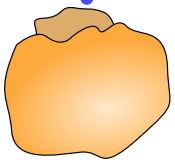




ARN

$\{A, C, G, U\}^\star$





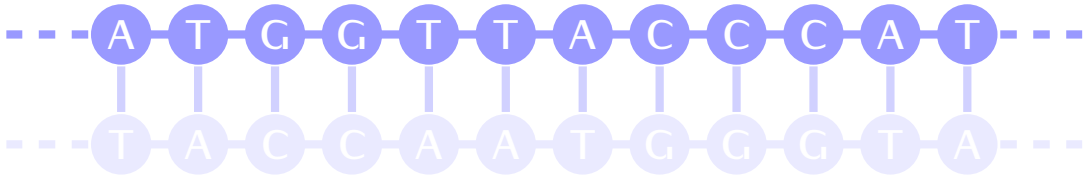
Protéines

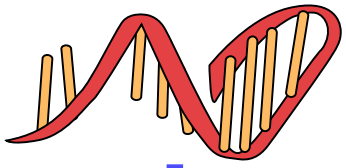
$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Acides aminés}}$

Dogme fondateur de la biologie moléculaire

ADN

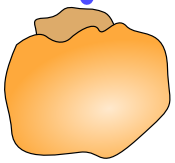
$\{A, C, G, T\}^\star$





ARN

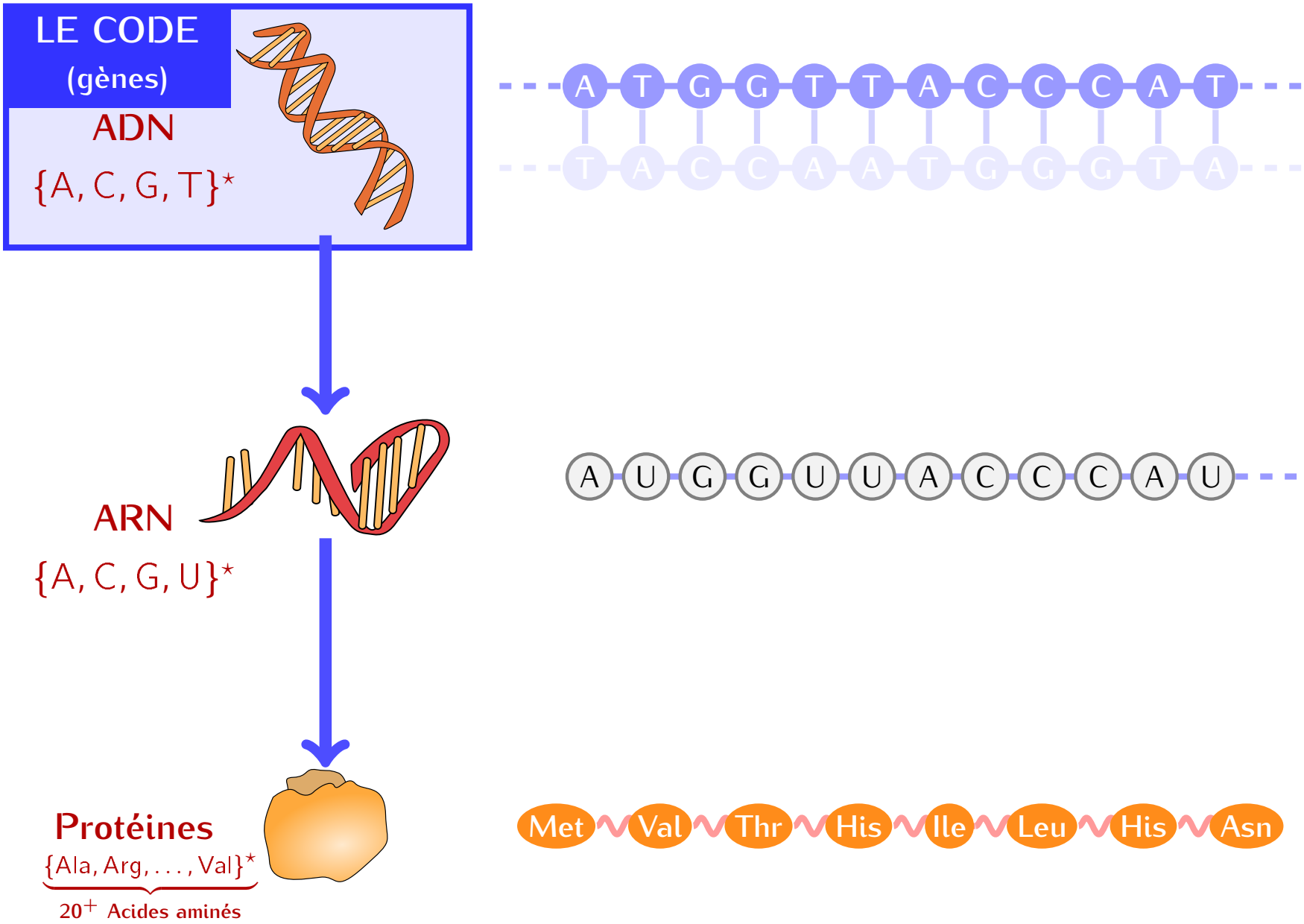
$\{A, C, G, U\}^\star$



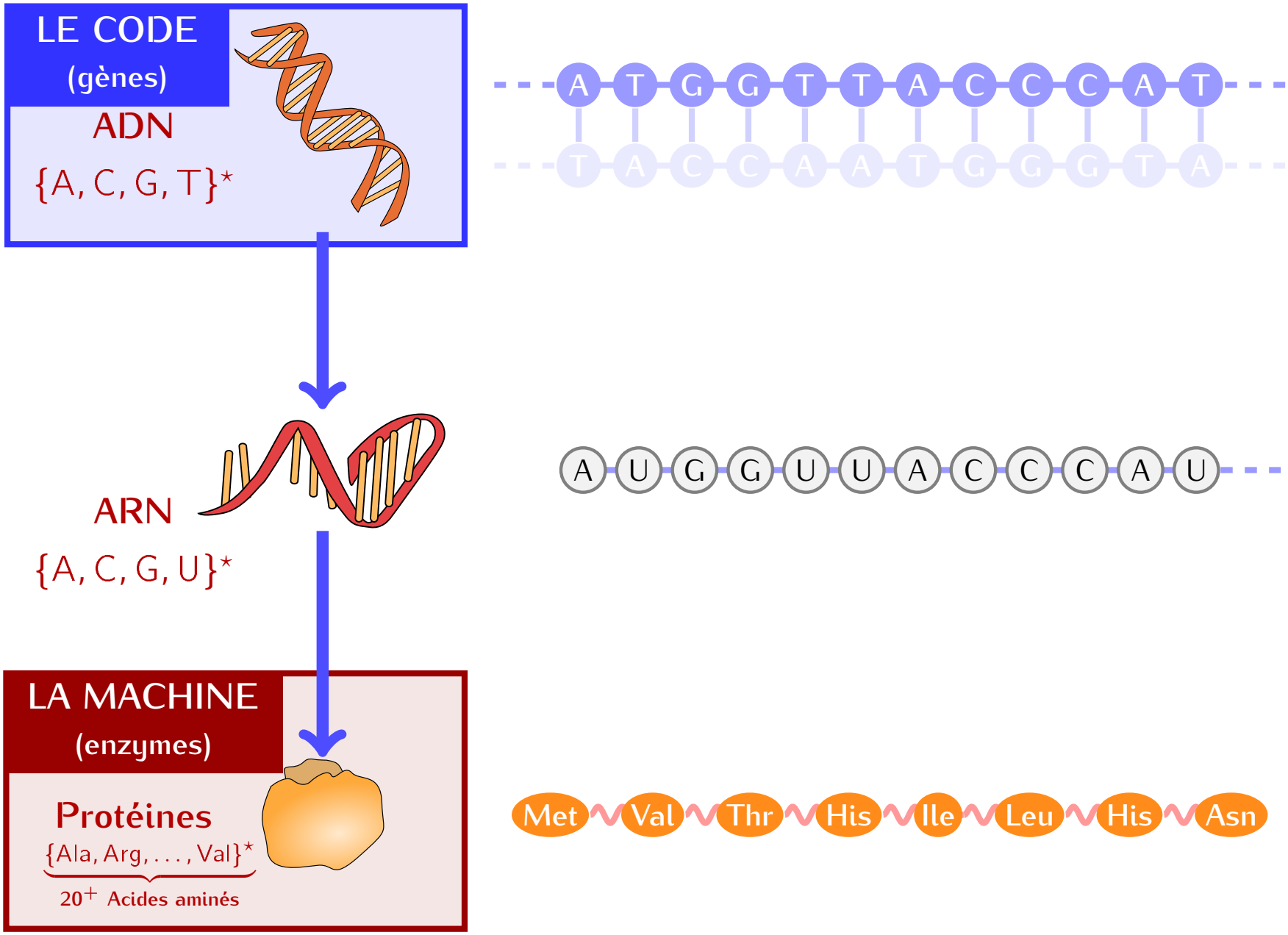
Protéines

$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Acides aminés}}$

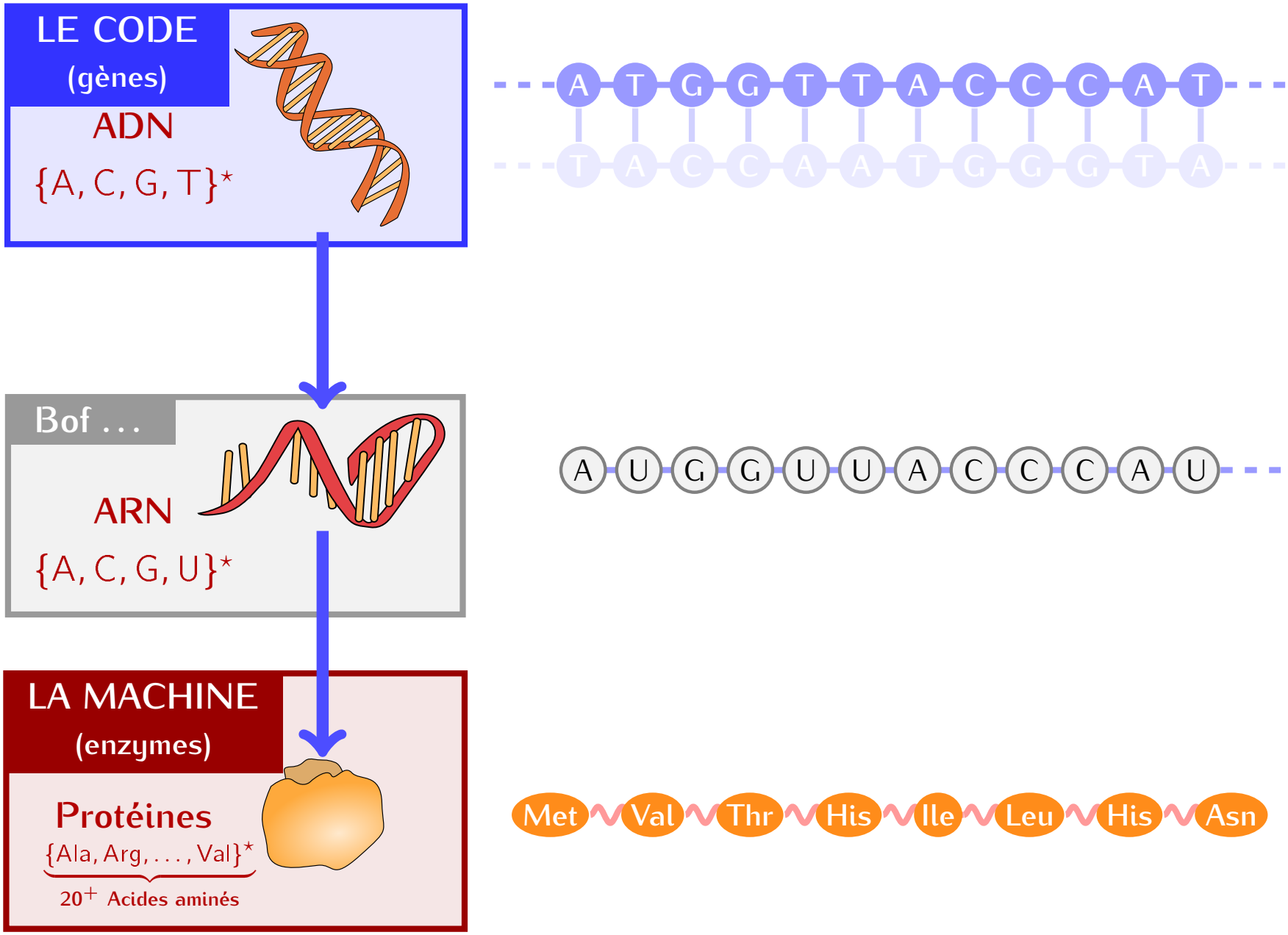
Dogme fondateur de la biologie moléculaire



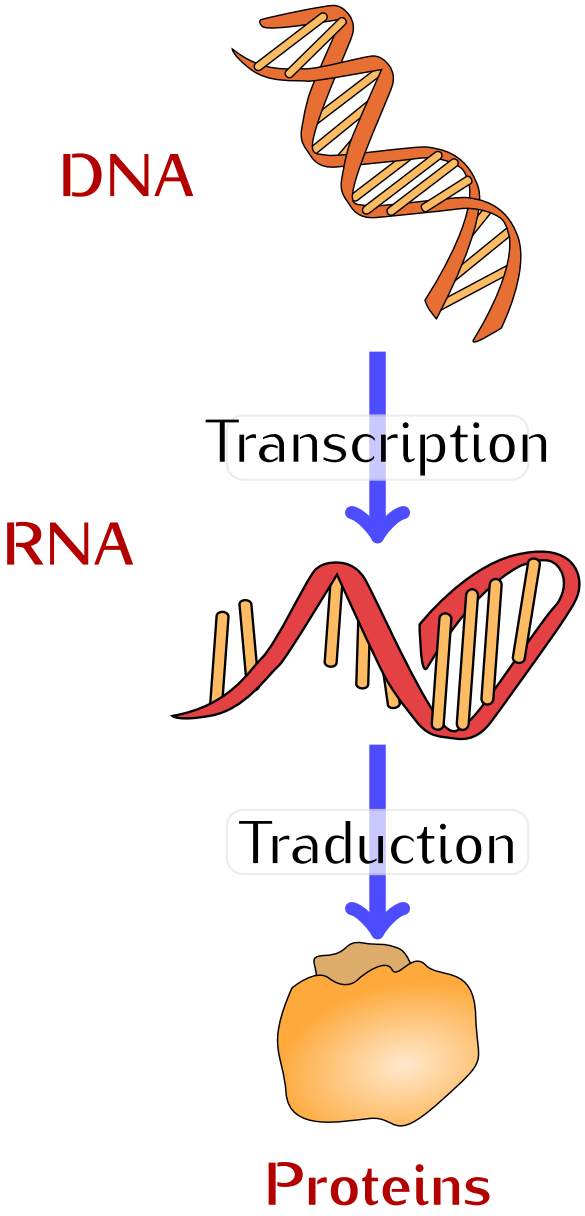
Dogme fondateur de la biologie moléculaire



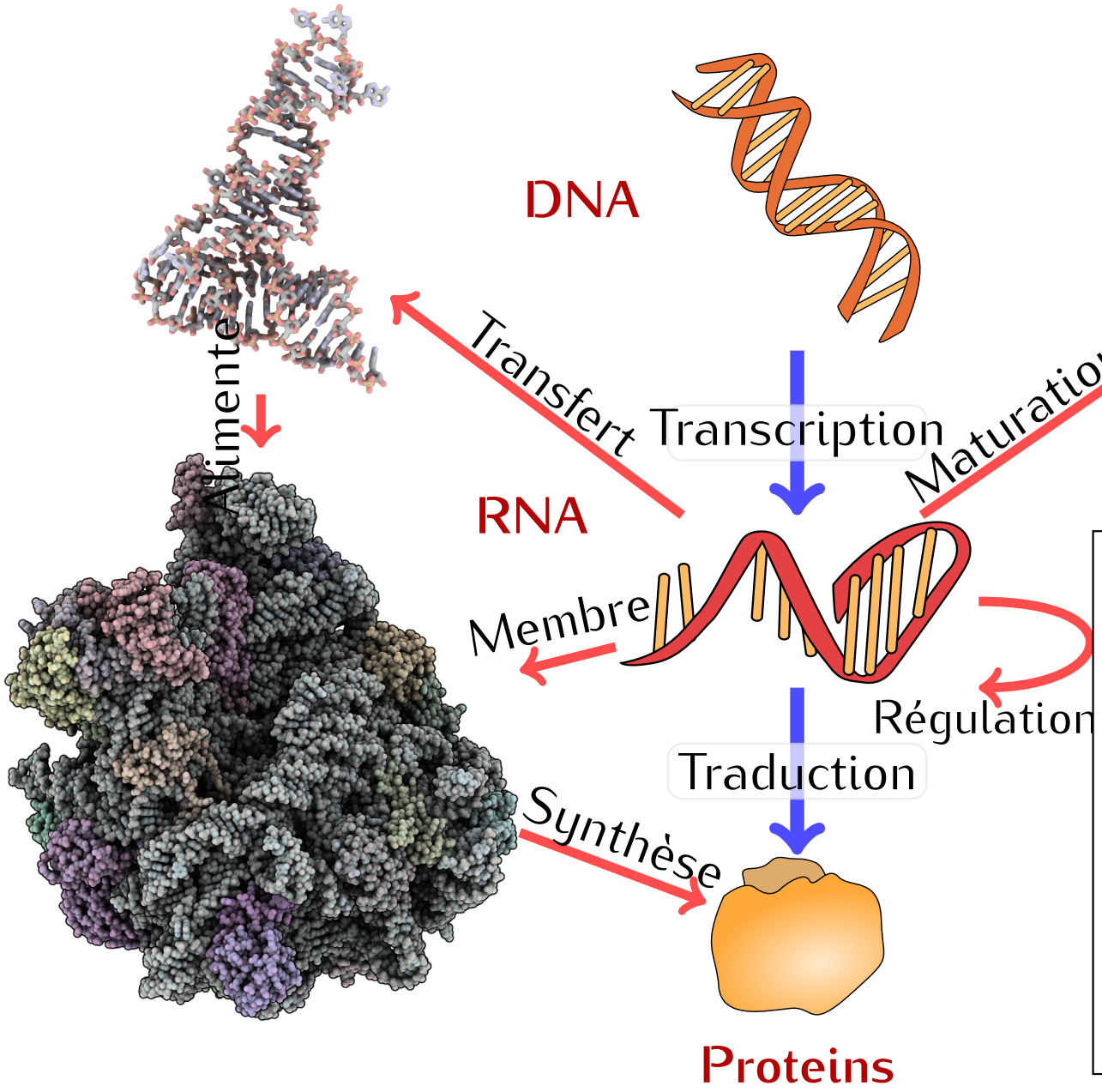
Dogme fondateur de la biologie moléculaire



Dogme fondateur de la biologie moléculaire



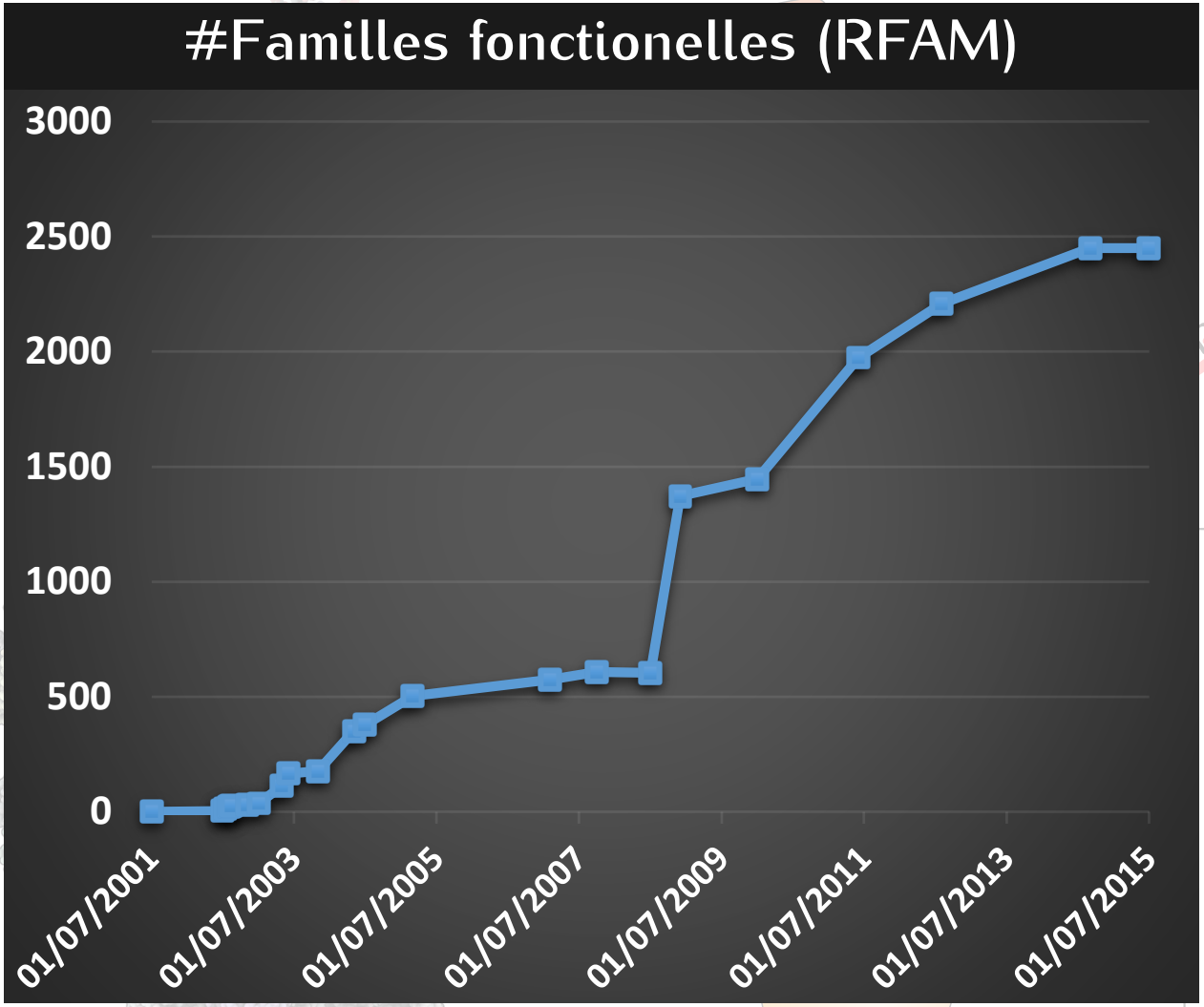
Dogme fondateur de la biologie moléculaire (v2.0)



ARN : fonctions

- Traduction
- Régulation
- Enzyme
- Messager
- Localisation
- ...

Dogme fondateur de la biologie moléculaire



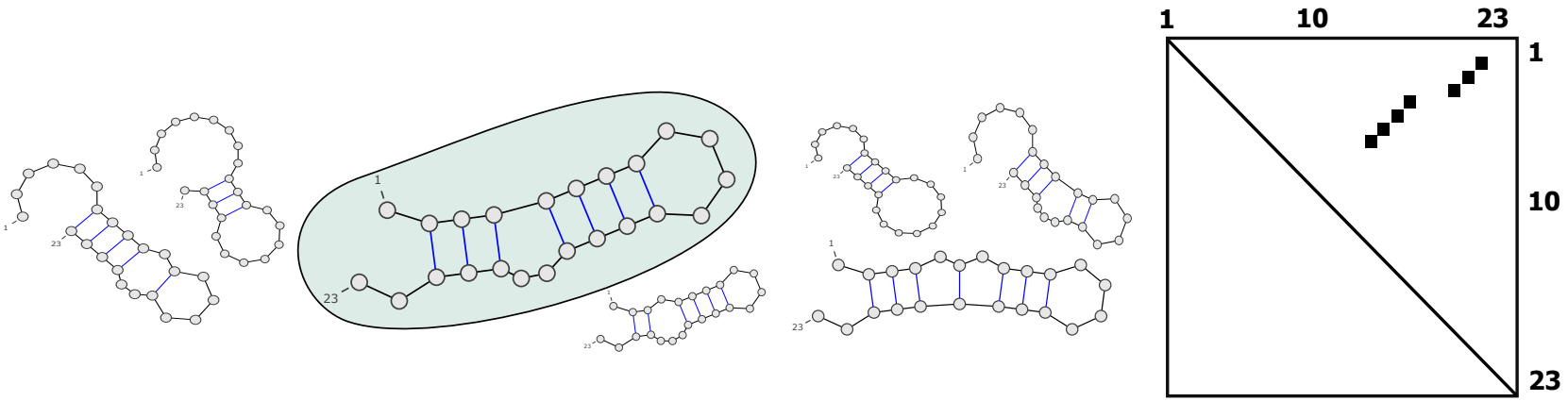
ARN : fonctions

- Traduction
- Régulation
- Enzyme
- Messager
- Localisation
- ...

Proteins

Structure(s) des ARN

- Structure(s) adoptée(s) par l'ARN influence(nt) sa fonction
- ARN messagers (ARNm) : Motifs structurels stables, important pour la localisation, l'épissage alternatif ...
- Détecter des régions (dé)structurées et/ou dynamiques



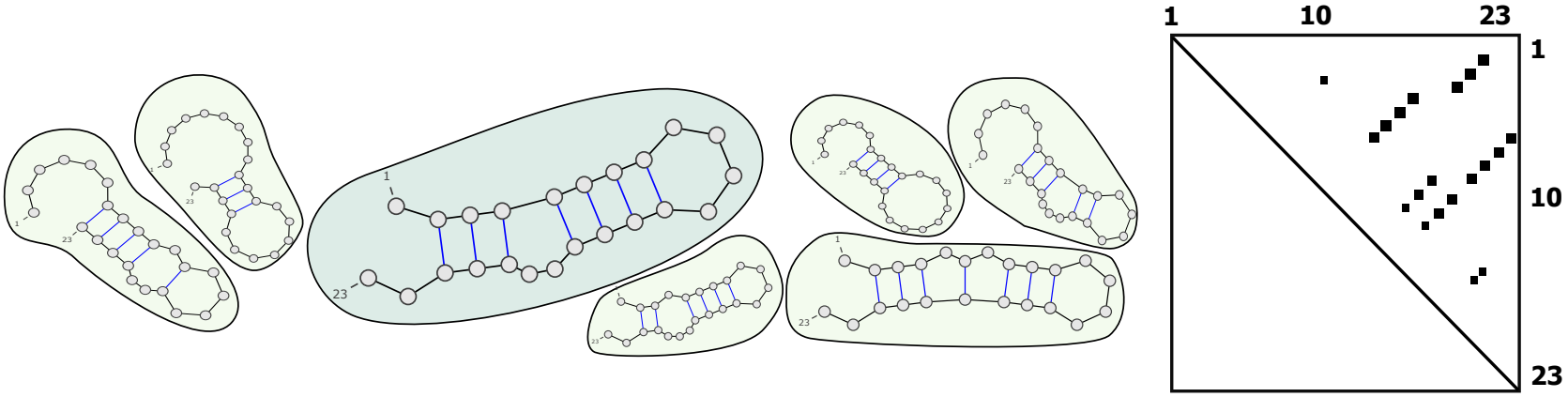
Équilibre de Boltzmann/fonction de partition d'un ARN w :

$$p_{i,j} = \frac{\sum_{\substack{s \in \mathcal{S}_w \\ (i,j) \in s}} e^{-E_{w,s}/RT}}{\mathcal{Z}_w} \qquad \mathcal{Z}_w := \sum_{s \in \mathcal{S}_w} e^{-E_{w,s}/RT}$$

Remarque : $\mathcal{Z}$ et $\{p_{i,j}\}$ calculables en $\Theta(n^3)$ [McCaskill 1990]

Structure(s) des ARN

- Structure(s) adoptée(s) par l'ARN influence(nt) sa fonction
- ARN messagers (ARNm) : Motifs structurels stables, important pour la localisation, l'épissage alternatif ...
- Détecter des régions (dé)structurées et/ou dynamiques



Équilibre de Boltzmann/fonction de partition d'un ARN w :

$$p_{i,j} = \frac{\sum_{\substack{s \in \mathcal{S}_w \\ (i,j) \in s}} e^{-E_{w,s}/RT}}{\mathcal{Z}_w} \qquad \mathcal{Z}_w := \sum_{s \in \mathcal{S}_w} e^{-E_{w,s}/RT}$$

Remarque : $\mathcal{Z}$ et $\{p_{i,j}\}$ calculables en $\Theta(n^3)$ [McCaskill 1990]

Significativité et génération aléatoire

Avant de s'étonner d'un phénomème quantitatif observé, vérifier qu'il n'est pas induit par des mécanismes déjà compris.

Test d'hypothèse : Soit $B \in \mathbb{R}$ un observable, observé à b
Hypothèse H_0 : "B induit par des mécanismes connus"
Si $\mathbb{P}(B = b \mid H_0) = \varepsilon$, alors anguille sous roche ...
$\Rightarrow H_0$ n'explique pas *toute* la réalité observée

Méta-approche *boîte noire* : Valeur observée b

- Génération aléatoire sous contrainte (H_0)
- Estimation moyenne μ et écart type σ pour B sous H_0
- Calcul du Z-score(b) $:= \frac{b-\mu}{\sigma}$
- Interprétation/hypothèse $\rightarrow$ Validation ($\rightarrow$ Ajout à H_0)

Significativité et génération aléatoire

Avant de s'étonner d'un phénomème quantitatif observé, vérifier qu'il n'est pas induit par des mécanismes déjà compris.

Test d'hypothèse : Soit $B \in \mathbb{R}$ un observable, observé à b
Hypothèse H_0 : "B induit par des mécanismes connus"
Si $\mathbb{P}(B = b \mid H_0) = \varepsilon$, alors anguille sous roche ...
$\Rightarrow H_0$ n'explique pas *toute* la réalité observée

Méta-approche *boîte noire* : Valeur observée b

- Génération aléatoire sous contrainte (H_0)
- Estimation moyenne μ et écart type σ pour B sous H_0
- Calcul du Z-score(b) $:= \frac{b-\mu}{\sigma}$
- Interprétation/hypothèse $\rightarrow$ Validation ($\rightarrow$ Ajout à H_0)

Significativité et génération aléatoire

Avant de s'étonner d'un phénomème quantitatif observé, vérifier qu'il n'est pas induit par des mécanismes déjà compris.

Test d'hypothèse : Soit $B \in \mathbb{R}$ un observable, observé à b
Hypothèse H_0 : "B induit par des mécanismes connus"
Si $\mathbb{P}(B = b \mid H_0) = \varepsilon$, alors anguille sous roche ...
$\Rightarrow H_0$ n'explique pas *toute* la réalité observée

Méta-approche *boîte noire* : Valeur observée b

- Génération aléatoire sous contrainte (H_0)
- Estimation moyenne μ et écart type σ pour B sous H_0
- Calcul du Z-score(b) $:= \frac{b-\mu}{\sigma}$
- Interprétation/hypothèse $\rightarrow$ Validation ($\rightarrow$ Ajout à H_0)

Modèle aléatoire contraint pour les ARNm

Observables : Fenêtre de largeur m dans un ARNm w,

- Somme des probabilités d'appariement $\mathcal{B}(w) = \sum_{i,j} p_{i,j}$
- Entropie des appariements $\mathcal{H}(w) = -\sum_{i,j} p_{i,j} \log(p_{i,j})$

Propriétés connues des ARN :

- Mutations synonymes : conservation de la protéine

- Stabilité globale des ARNm préservée par l'évolution
  $\approx$ ARNm préservant les fréquences de dinucléotides

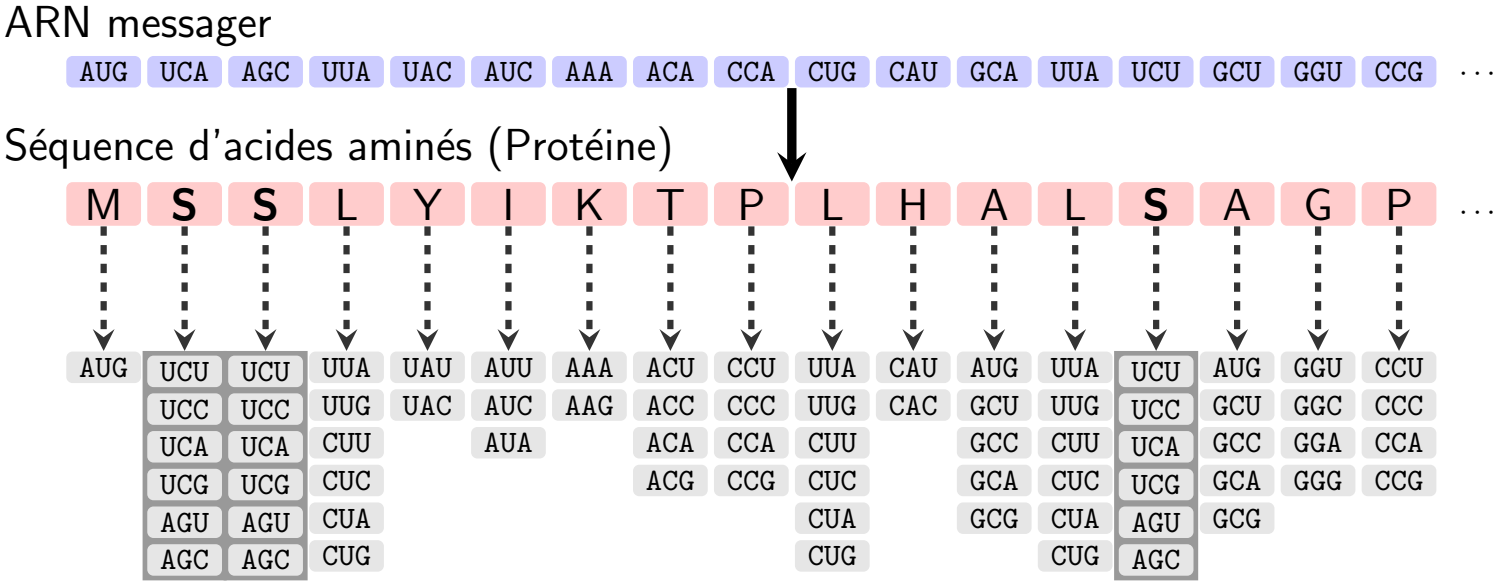
Modèle aléatoire contraint pour les ARNm

Observables : Fenêtre de largeur m dans un ARNm w,

- Somme des probabilités d'appariement $\mathcal{B}(w) = \sum_{i,j} p_{i,j}$
- Entropie des appariements $\mathcal{H}(w) = -\sum_{i,j} p_{i,j} \log(p_{i,j})$

Propriétés connues des ARN :

- Mutations synonymes : conservation de la protéine



- Stabilité globale des ARNm préservée par l'évolution
  $\approx$ ARNm préservant les fréquences de dinucléotides

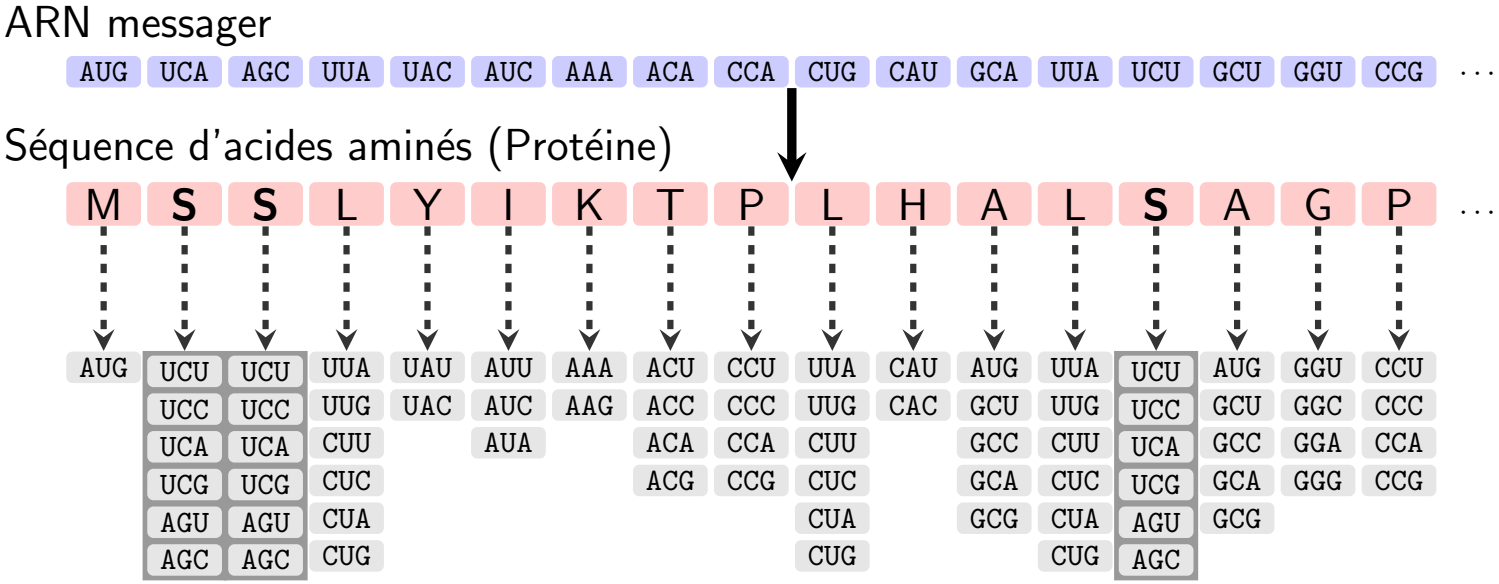
Modèle aléatoire contraint pour les ARNm

Observables : Fenêtre de largeur m dans un ARNm w,

- Somme des probabilités d'appariement $\mathcal{B}(w) = \sum_{i,j} p_{i,j}$
- Entropie des appariements $\mathcal{H}(w) = -\sum_{i,j} p_{i,j} \log(p_{i,j})$

Propriétés connues des ARN :

- Mutations synonymes : conservation de la protéine



- Stabilité globale des ARNm préservée par l'évolution
  $\approx$ ARNm préservant les fréquences de dinucléotides

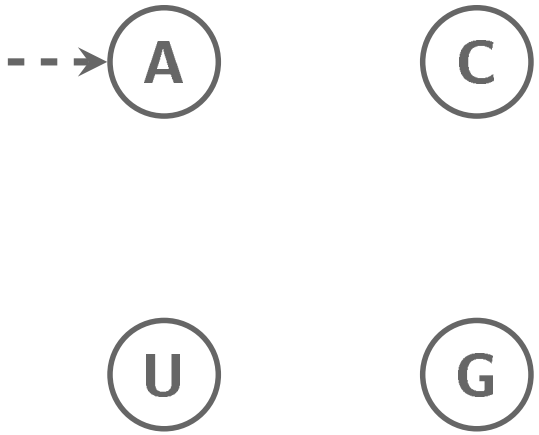
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC



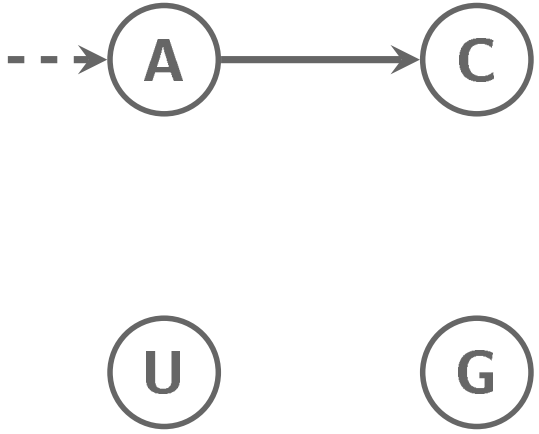
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : $\boxed{\mathrm{AC}}$GUAUAGGC



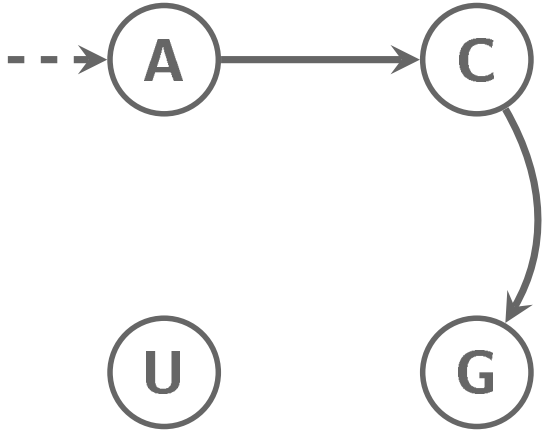
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : A $\boxed{\mathrm{CG}}$ UAUAGGC



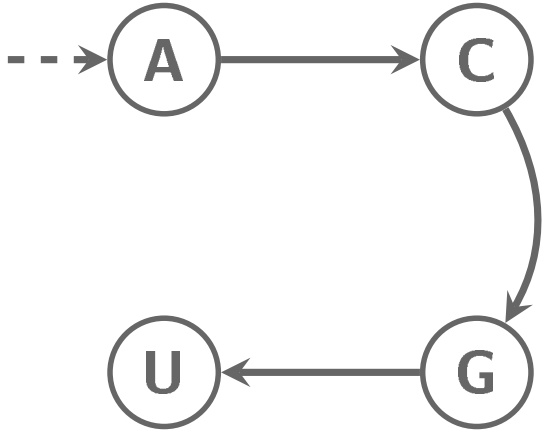
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : AC $\boxed{\text{GU}}$ AUAGGC



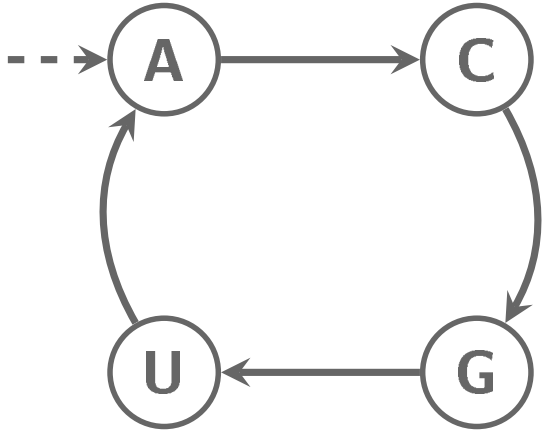
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACG $\boxed{\text{UA}}$ UAGGC



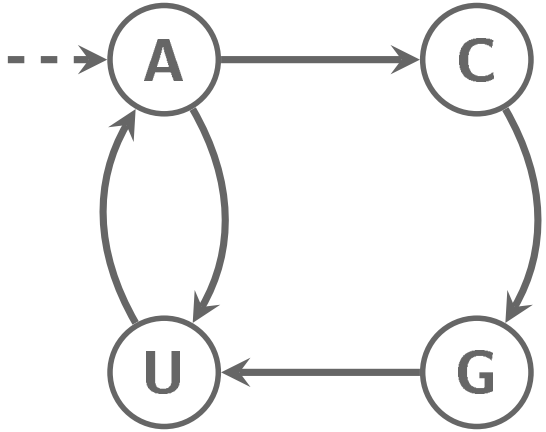
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGU $\boxed{\text{AU}}$ AGGC



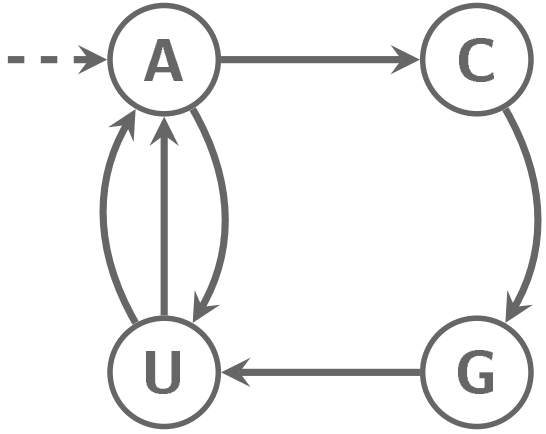
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUA $\boxed{\text{UA}}$ GGC



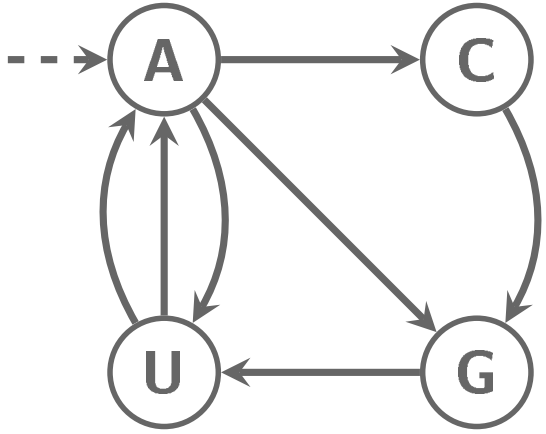
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAU$\boxed{\mathrm{AG}}$GC



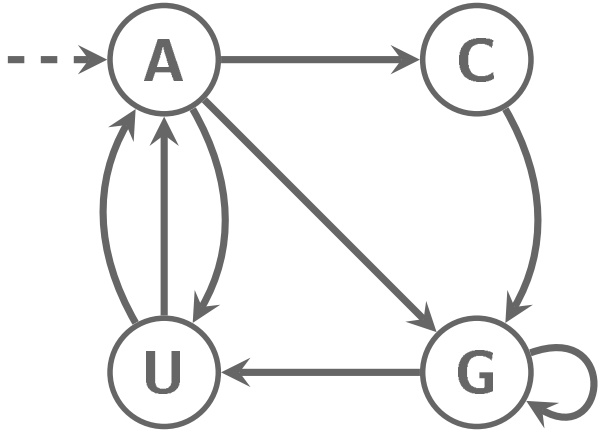
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUA $\boxed{\mathrm{GG}}$ C



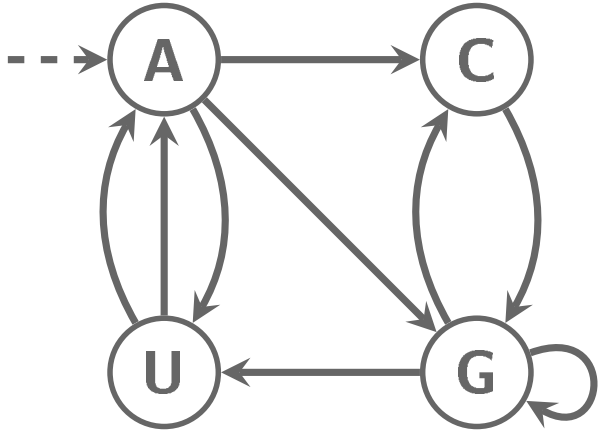
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall(\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAG $\boxed{\text{GC}}$



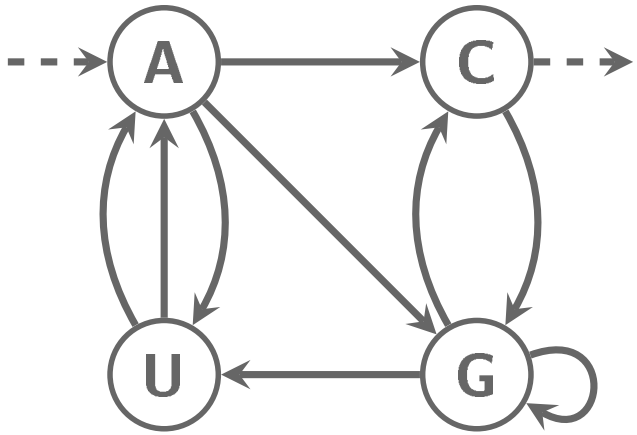
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC



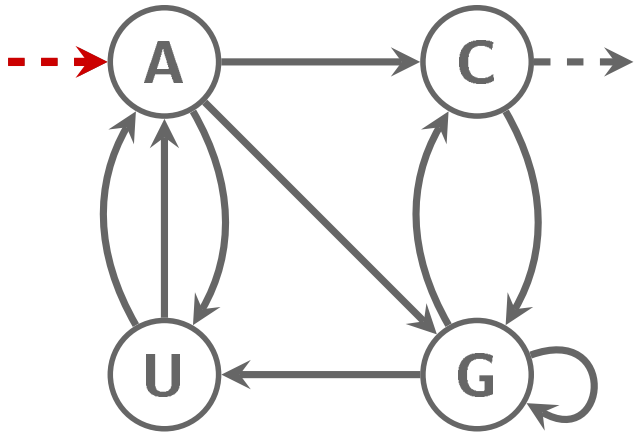
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall(\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC (Alt. : AGCGGUAUAC ...)



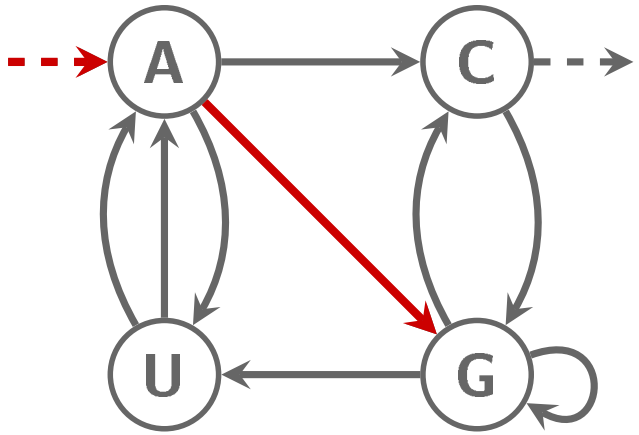
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC (Alt. : AGCGGUAUAC ...)



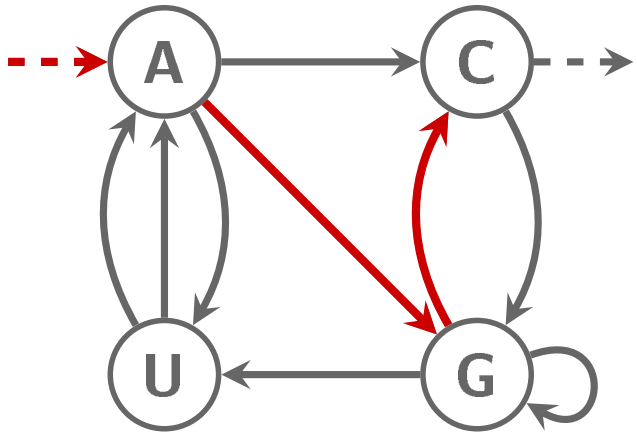
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall(\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC (Alt. : AGCGGUAUAC ...)



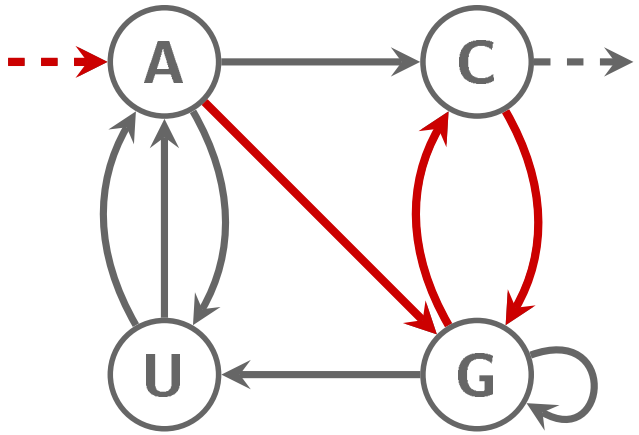
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC (Alt. : AGCGGUAUAC ...)



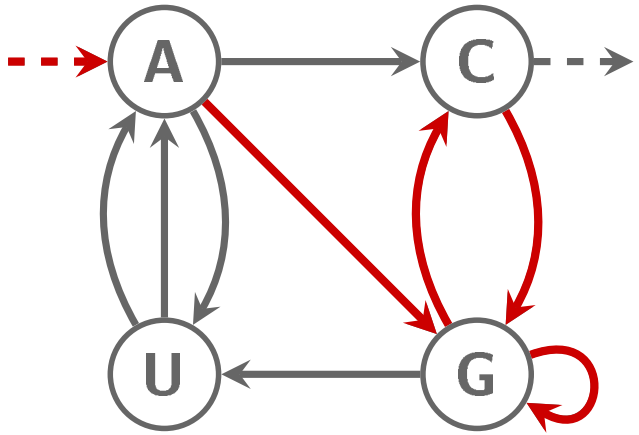
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC (Alt. : AGCGGUAUAC ...)



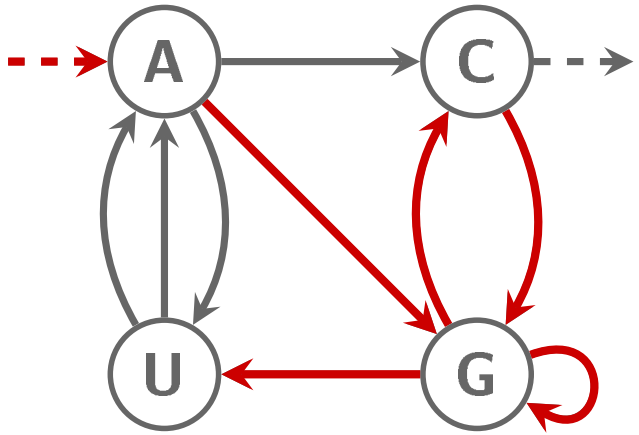
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC (Alt. : AGCGGUAUAC ...)



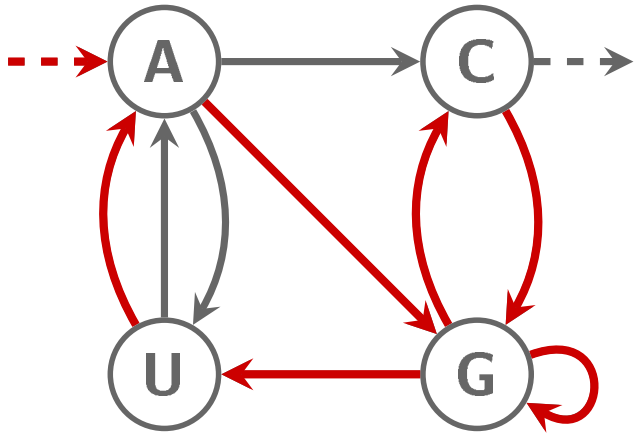
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC (Alt. : AGCGGUAUAC ...)



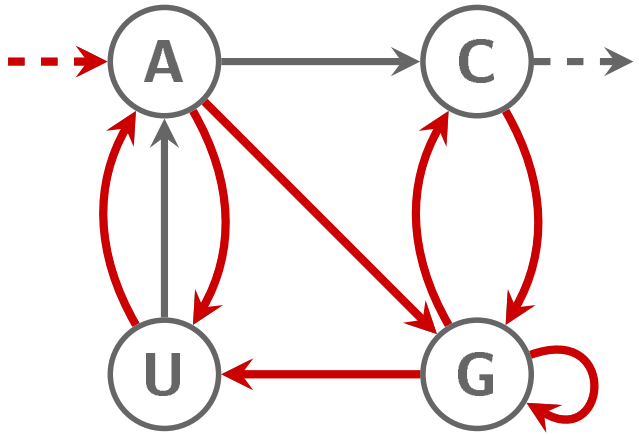
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC (Alt. : AGCGGUAUAC ...)



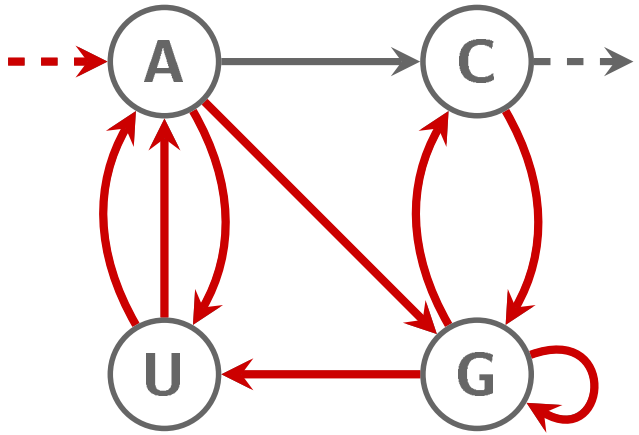
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC (Alt. : AGCGGUAUAC ...)



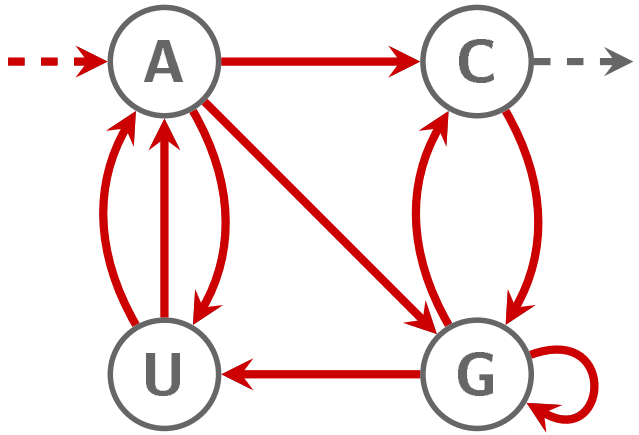
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC (Alt. : AGCGGUAUAC ...)



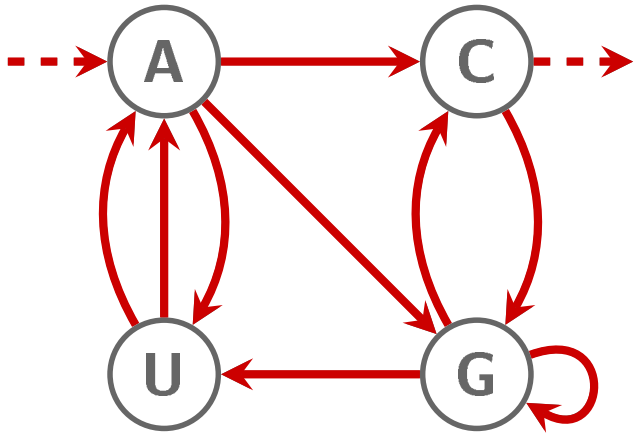
Génération aléatoire sous contrainte de dinucléotides

Entrée : Séquence d'ARN $w \in \mathcal{B}^\star$, $\mathcal{B} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$
Sortie : Séquence w' engendrée uniformément telle que

$$|w'|_{\mathrm{XY}} = |w|_{\mathrm{XY}}, \forall (\mathrm{X}, \mathrm{Y}) \in \mathcal{B} \times \mathcal{B}$$

$\Leftrightarrow$ Engendrer un chemin eulérien dans le multigraphe $G_w = (\mathcal{B}, E_w)$ où $E_w := \{\mathrm{w_i} \to \mathrm{w_{i+1}}\}_{i=0}^{|w|-1}$ (avec multiplicité).

Exemple : ACGUAUAGGC (Alt. : AGCGGUAUAC ...)



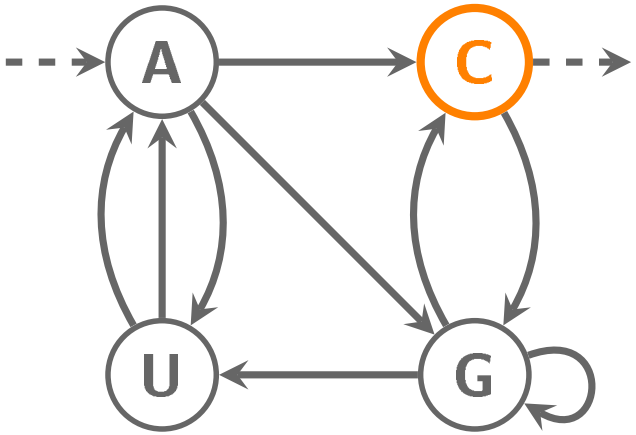
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

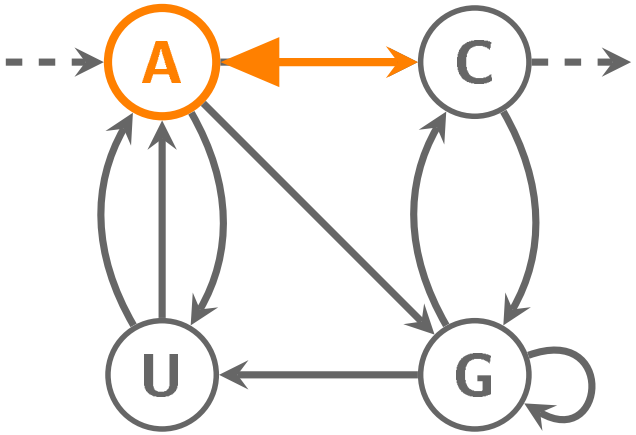
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

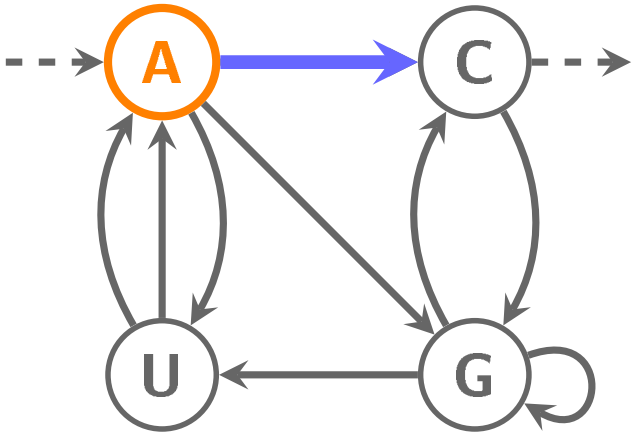
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

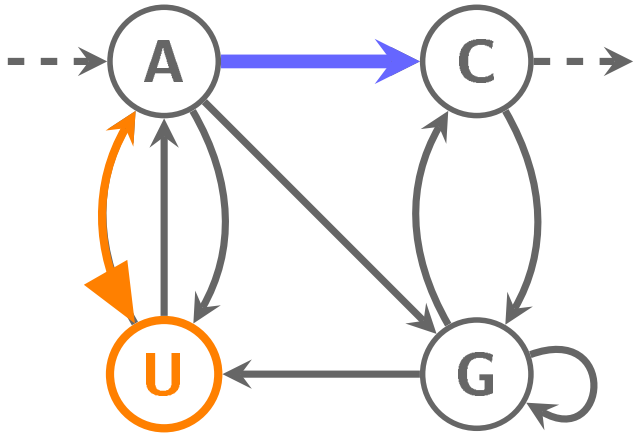
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

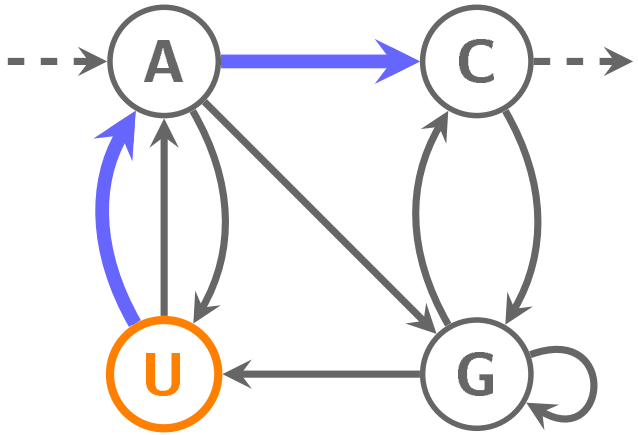
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

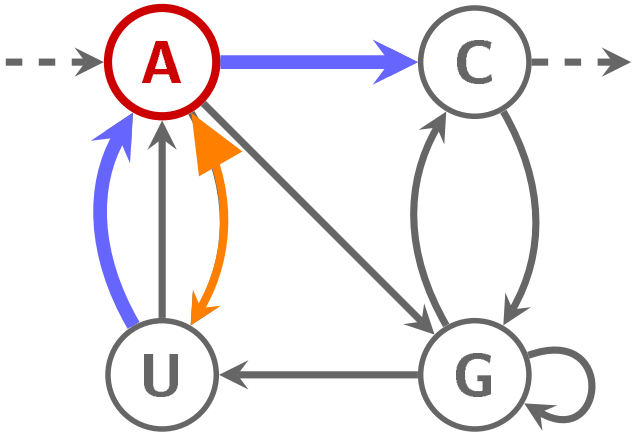
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

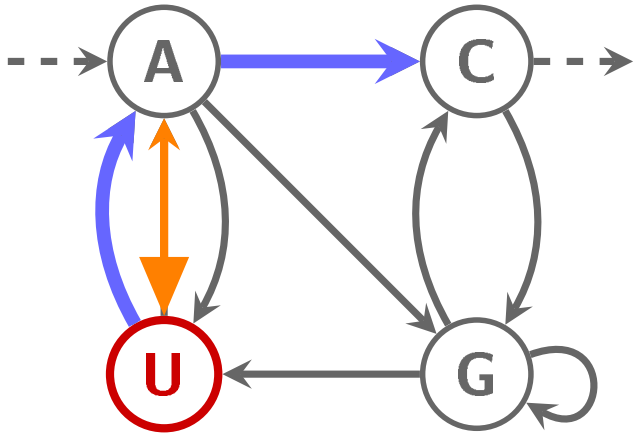
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

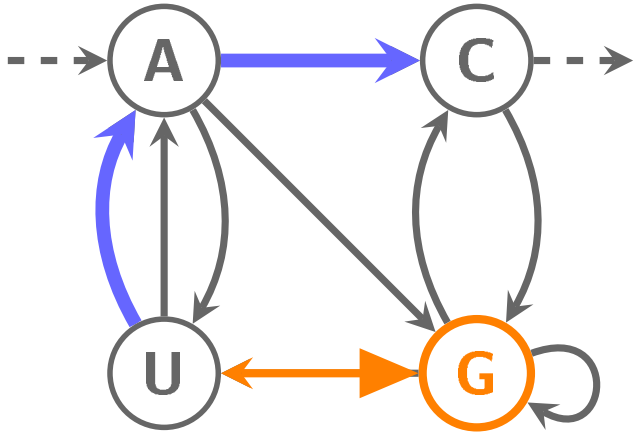
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

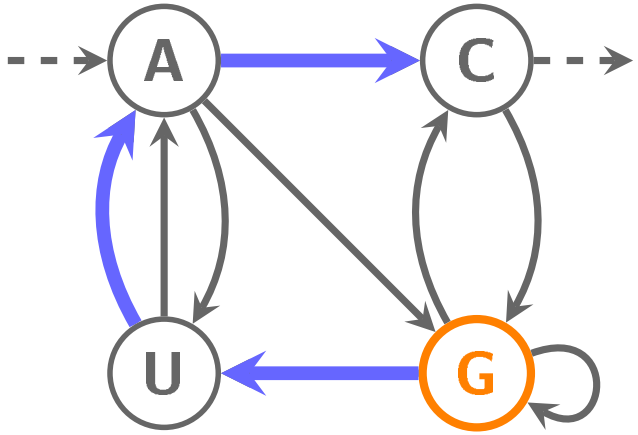
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

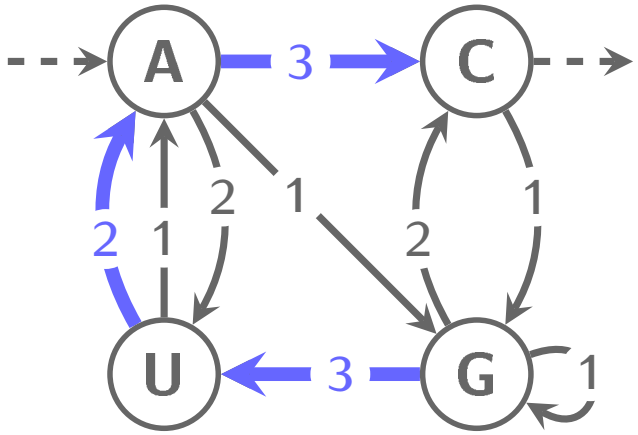
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

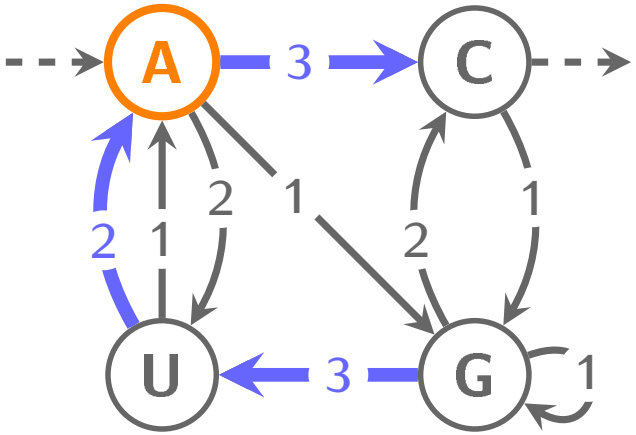
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

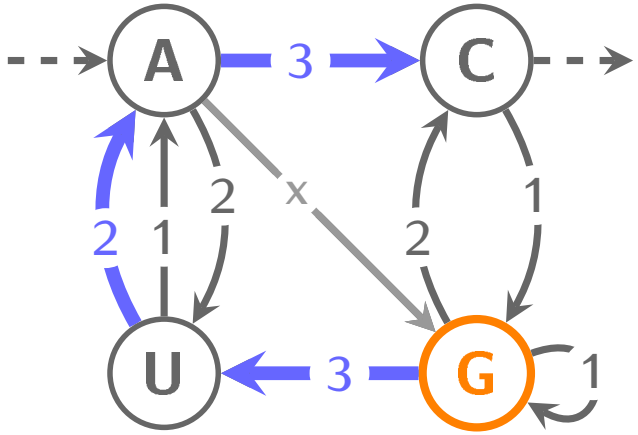
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

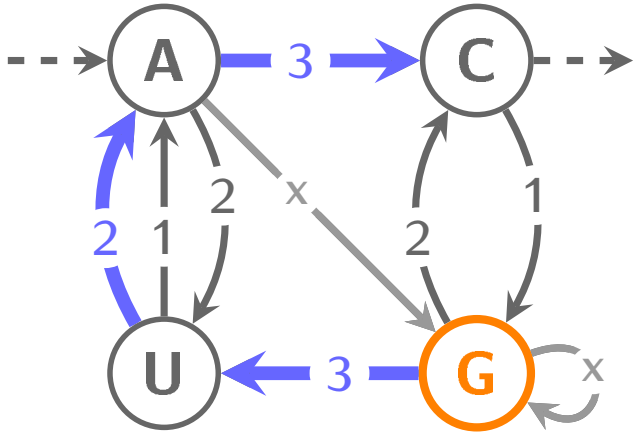
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

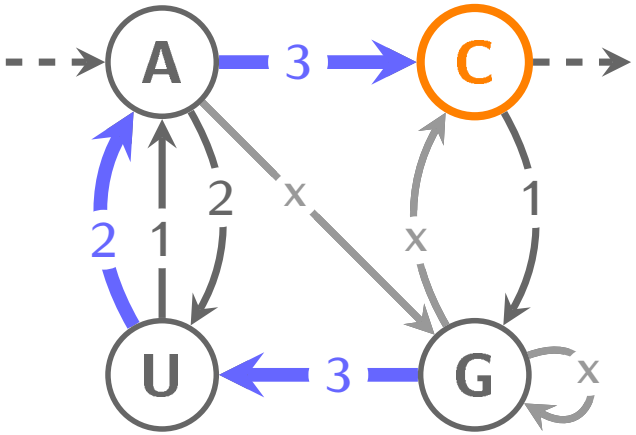
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

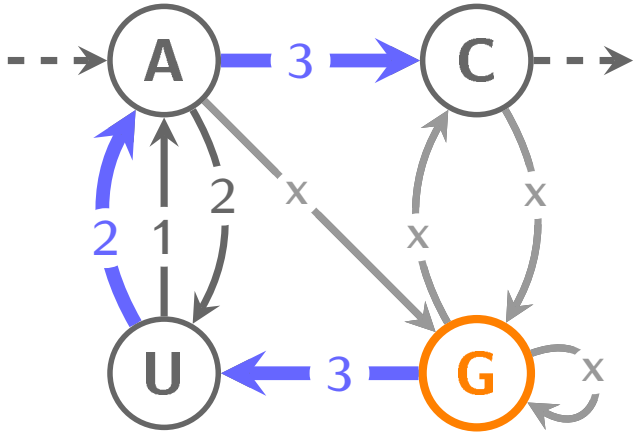
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

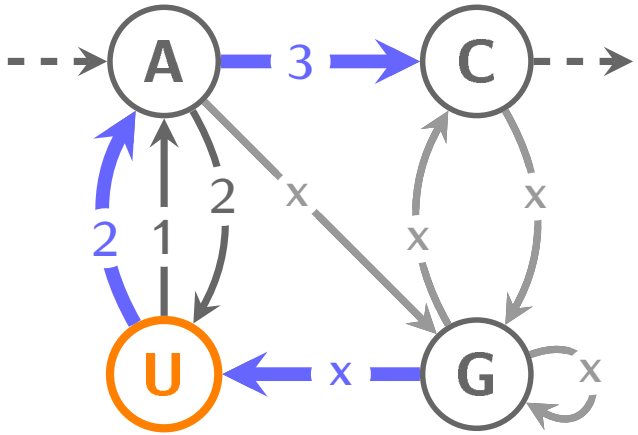
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

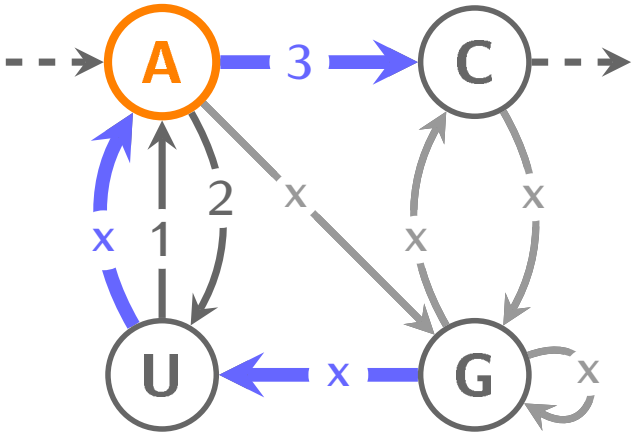
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

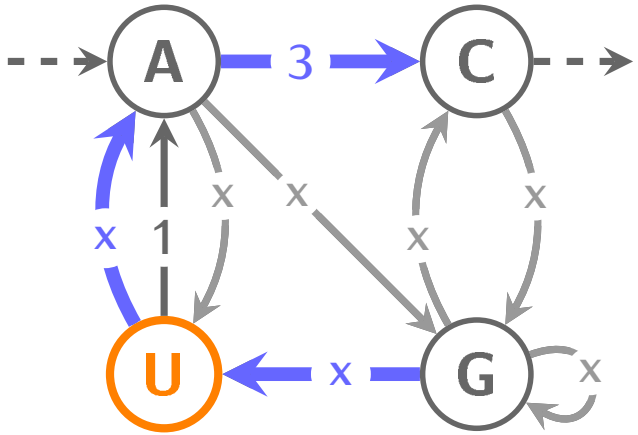
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

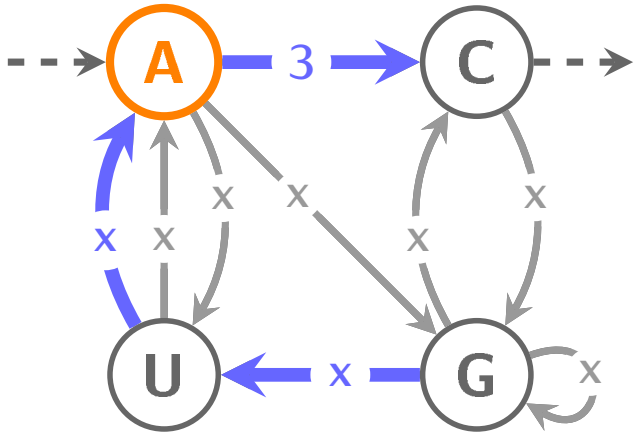
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

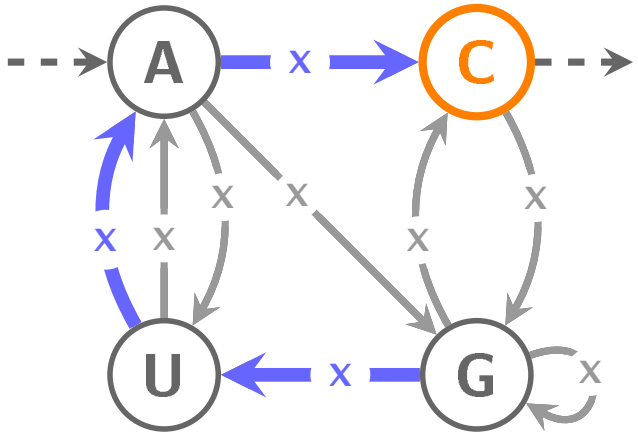
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

$$\mathbb{P}(c) = \frac{1}{|\mathcal{T}_{w_{-1}}|} \times \frac{1}{\prod_{b \in \mathcal{B}} (d^+(b) - 1)!}$$

$$= \frac{1}{\#\text{Chemins Eulériens}}$$

[de Bruijn/van Aardenne-Ehrenfest/Smith/Tutte 51]

Rem. : Aussi uniforme sur les mots car chaque mot représenté par $\prod_{b \in \mathcal{B}} d^+(b)!$ chemins eulériens.

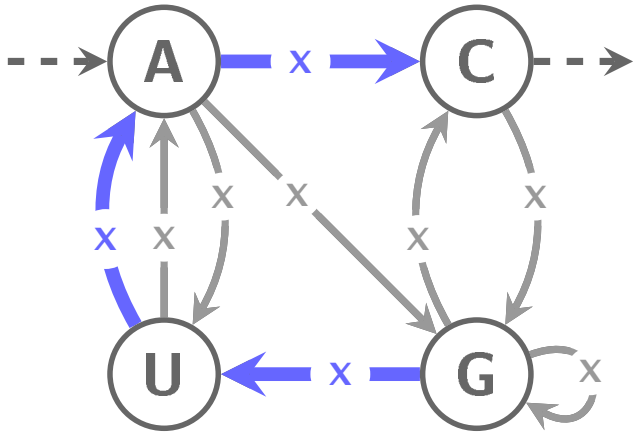
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

$$\mathbb{P}(c) = \frac{1}{|\mathcal{T}_{w_{-1}}|} \times \frac{1}{\prod_{b \in \mathcal{B}} (d^+(b) - 1)!}$$

$$= \frac{1}{\#\text{Chemins Eulériens}}$$

[de Bruijn/van Aardenne-Ehrenfest/Smith/Tutte 51]

Rem. : Aussi uniforme sur les mots car chaque mot représenté par $\prod_{b \in \mathcal{B}} d^+(b)!$ chemins eulériens.

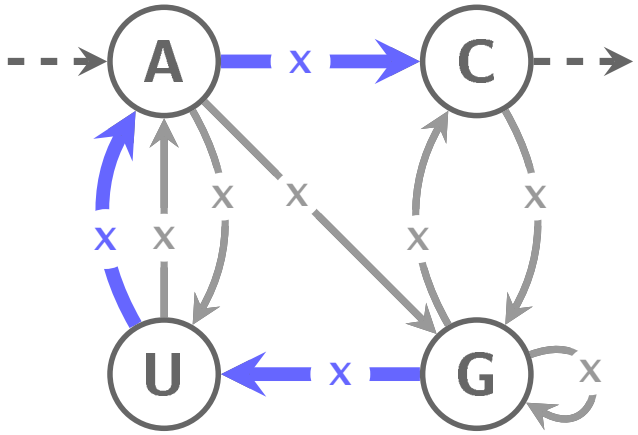
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

$$\mathbb{P}(c) = \frac{1}{|\mathcal{T}_{w_{-1}}|} \times \frac{1}{\prod_{b \in \mathcal{B}} (d^+(b) - 1)!}$$

$$= \frac{1}{\#\text{Chemins Eulériens}}$$

[de Bruijn/van Aardenne-Ehrenfest/Smith/Tutte 51]

Rem. : Aussi uniforme sur les mots car chaque mot représenté par $\prod_{b \in \mathcal{B}} d^+(b)!$ chemins eulériens.

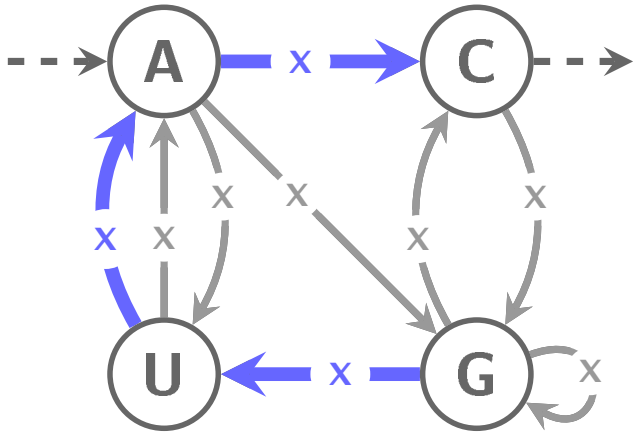
Génération aléatoire sous contrainte de dinucléotides

Algorithme [Kandel/Matias/Unger/Winkler 96] :

1. Engendrer un arbre couvrant $t \in \mathcal{T}_{w_{-1}}$ (enraciné en w_{-1}) par une marche *arrière* et *premier passage* ;
2. Marche avant sur permutation des arêtes sortantes, sauf celles de t, prises en dernier.

$\Rightarrow$ Chemin Eulérien uniforme en temps linéaire

Exemple : ACGUAUAGGC



AGGCGUAUAC

$$\mathbb{P}(c) = \frac{1}{|\mathcal{T}_{w_{-1}}|} \times \frac{1}{\prod_{b \in \mathcal{B}} (d^+(b) - 1)!}$$
$$= \frac{1}{\#\text{Chemins Eulériens}}$$

[de Bruijn/van Aardenne-Ehrenfest/Smith/Tutte 51]

Rem. : Aussi uniforme sur les mots car chaque mot représenté par $\prod_{b \in \mathcal{B}} d^+(b)!$ chemins eulériens.

Combiner les contraintes

- Mutations synonymes : conservation de la protéine
  $\Rightarrow$ Automate fini (Tirages indépendants)
- Stabilité globale des ARNm préservée par l'évolution
  $\Rightarrow$ Chemins Eulériens

Combiner les deux ? Dur dur ...

- Chemins Eulériens : Pas de formalisme grammatical
- Synonymie brise la symétrie $\rightarrow$ Plus uniforme

Proposition pragmatique [Burge et Katz 03] :

- Échange de paires de codons préservant les dinuc. ...
- ... **ou** pouvant être compensée par un $2^{\text{ième}}$ échange ;
- Sans remise au niveau des codons échangés.

Erronée (non-uniforme, non-ergodique) en théorie et pratique.

Combiner les contraintes

- Mutations synonymes : conservation de la protéine
  $\Rightarrow$ Automate fini (Tirages indépendants)
- Stabilité globale des ARNm préservée par l'évolution
  $\Rightarrow$ Chemins Eulériens

Combiner les deux ? Dur dur ...

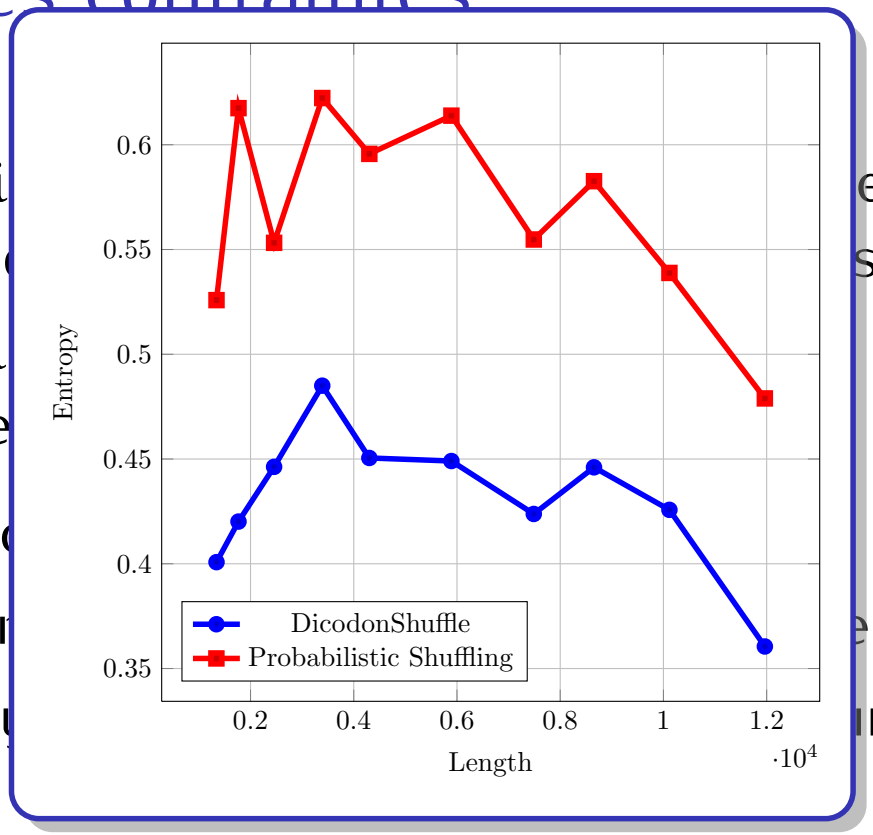
- Chemins Eulériens : Pas de formalisme grammatical
- Synonymie brise la symétrie $\rightarrow$ Plus uniforme

Proposition pragmatique [Burge et Katz 03] :

- Échange de paires de codons préservant les dinuc. ...
- ... **ou** pouvant être compensée par un $2^{\text{ième}}$ échange ;
- Sans remise au niveau des codons échangés.

Erronée (non-uniforme, non-ergodique) en théorie et pratique.

Combiner les contraintes



- Mutati … e la protéine
  ⇒ Aut … s)
- Stabili … par l'évolution
  ⇒ Che …

Combiner l …

- Chemin … e grammatical
- Synon … niforme

Proposition pragmatique [Burge et Katz 03] :

- Échange de paires de codons préservant les dinuc. …
- … **ou** pouvant être compensée par un $2^{\text{ième}}$ échange ;
- Sans remise au niveau des codons échangés.

Erronée (non-uniforme, non-ergodique) en théorie et pratique.

Génération de Boltzmann multidimensionnelle

Génération de Boltzmann [Duchon/Flajolet/Louchard/Schaeffer 04]

- Idée : Engendrer des objets selon une distribution

$$\mathbb{P}(w \mid x) = \frac{x^{|w|}}{S(x)}$$

où $S(z)$ la s. g. de la classe, et x un paramètre magique ;

- Classes spécifiables : implémentations génériques et $\exists x \in [0, \rho]$ tel que le rejet soit efficace $\rightarrow \mathcal{O}(1)$ à $\mathcal{O}(n)$

Boltzmann multidimensionnel [Bodini/P. 10]

- Idée : Généralisation à k paramètres additifs ;
- Engendrer selon une distribution pondérée :

$$\mathbb{P}(w \mid x_1, x_2 \ldots) = \frac{\prod_{i=1}^{k} x_i^{|w|_i}}{S(x_1, x_2 \ldots)}$$

où $S(u_1, \ldots)$ s. g. multivariée, et $x_1, \ldots$ param. magiques ;

- Distrib. concentrées $\Rightarrow$ Rejet très efficace $\rightarrow \mathcal{O}(1)$ à $\mathcal{O}(n^{k/2})$

Génération de Boltzmann multidimensionnelle

Génération de Boltzmann [Duchon/Flajolet/Louchard/Schaeffer 04]

- Idée : Engendrer des objets selon une distribution

$$\mathbb{P}(w \mid x) = \frac{x^{|w|}}{S(x)}$$

  où $S(z)$ la s. g. de la classe, et x un paramètre magique ;
- Classes spécifiables : implémentations génériques et $\exists x \in [0, \rho]$ tel que le rejet soit efficace $\rightarrow \mathcal{O}(1)$ à $\mathcal{O}(n)$

Boltzmann multidimensionnel [Bodini/P. 10]

- Idée : Généralisation à k paramètres additifs ;
- Engendrer selon une distribution pondérée :

$$\mathbb{P}(w \mid x_1, x_2 \ldots) = \frac{\prod_{i=1}^{k} x_i^{|w|_i}}{S(x_1, x_2 \ldots)}$$

  où $S(u_1, \ldots)$ s. g. multivariée, et $x_1, \ldots$ param. magiques ;
- Distrib. concentrées $\Rightarrow$ Rejet très efficace $\rightarrow \mathcal{O}(1)$ à $\mathcal{O}(n^{k/2})$

Séquences codantes par Boltzmann en 16 dimensions

Considérons la s. g. des séquences synonymes d'un ARNm m

$$S(u_{\mathrm{AA}}, u_{\mathrm{AU}} \dots u_{\mathrm{GG}}) = \sum_{w' \in Syn(m)} \prod_{b,b' \in \mathcal{B}^2} u_{bb'}^{|w'|_{bb'}}$$

1. Calculer $\mathbf{x} = \{x_{bb'}\}_{b,b' \in \mathcal{B}^2}$ tel que

$$u_{bb'} \cdot \frac{\partial S(u_{\mathrm{AA}}, \dots, u_{\mathrm{GG}})}{\partial u_{b,b'}} \Bigg|_{\substack{u_{\mathrm{AA}} = x_{\mathrm{AA}} \\ u_{\mathrm{AU}} = x_{\mathrm{AU}} \\ \dots}} = |m|_{bb'}, \forall b, b' \in \mathcal{B}^2$$

2. Engendrer w selon la distribution pondérée par $\mathbf{x}$

$$\mathbb{P}(w \mid x_{\mathrm{AA}}, \dots, x_{\mathrm{GG}}) = \frac{\prod_{b,b' \in \mathcal{B}^2} u_{bb'}^{|w|_{bb'}}}{S(x_{\mathrm{AA}}, \dots, x_{\mathrm{GG}})}$$

3. Accepter uniquement les mots w tels que

$$\frac{|w|_{bb'}}{|m|_{bb'}} \in [1 - \varepsilon(n), 1 + \varepsilon(n)], \forall (b, b') \in \mathcal{B}^2$$

Séquences codantes par Boltzmann en 16 dimensions

Considérons la s. g. des séquences synonymes d'un ARNm m

$$S(u_{\mathrm{AA}}, u_{\mathrm{AU}} \dots u_{\mathrm{GG}}) = \sum_{w' \in Syn(m)} \prod_{b,b' \in \mathcal{B}^2} u_{bb'}^{|w'|_{bb'}}$$

1. Calculer $\mathbf{x} = \{x_{bb'}\}_{b,b' \in \mathcal{B}^2}$ tel que

$$u_{bb'} \cdot \frac{\partial S(u_{\mathrm{AA}}, \dots, u_{\mathrm{GG}})}{\partial u_{b,b'}} \Bigg|_{\substack{u_{\mathrm{AA}} = x_{\mathrm{AA}} \\ u_{\mathrm{AU}} = x_{\mathrm{AU}} \\ \dots}} = |m|_{bb'}, \forall b, b' \in \mathcal{B}^2$$

2. Engendrer w selon la distribution pondérée par $\mathbf{x}$

$$\mathbb{P}(w \mid x_{\mathrm{AA}}, \dots, x_{\mathrm{GG}}) = \frac{\prod_{b,b' \in \mathcal{B}^2} u_{bb'}^{|w|_{bb'}}}{S(x_{\mathrm{AA}}, \dots, x_{\mathrm{GG}})}$$

3. Accepter uniquement les mots w tels que

$$\frac{|w|_{bb'}}{|m|_{bb'}} \in [1 - \varepsilon(n), 1 + \varepsilon(n)], \forall (b, b') \in \mathcal{B}^2$$

Séquences codantes par Boltzmann en 16 dimensions

Considérons la s. g. des séquences synonymes d'un ARNm m

$$S(u_{\mathrm{AA}}, u_{\mathrm{AU}} \dots u_{\mathrm{GG}}) = \sum_{w' \in Syn(m)} \prod_{b,b' \in \mathcal{B}^2} u_{bb'}^{|w'|_{bb'}}$$

1. Calculer $\mathbf{x} = \{x_{bb'}\}_{b,b' \in \mathcal{B}^2}$ tel que

$$u_{bb'} \cdot \left. \frac{\partial S(u_{\mathrm{AA}}, \dots, u_{\mathrm{GG}})}{\partial u_{b,b'}} \right|_{\substack{u_{\mathrm{AA}} = x_{\mathrm{AA}} \\ u_{\mathrm{AU}} = x_{\mathrm{AU}} \\ \dots}} = |m|_{bb'}, \forall b, b' \in \mathcal{B}^2$$

2. Engendrer w selon la distribution pondérée par $\mathbf{x}$

$$\mathbb{P}(w \mid x_{\mathrm{AA}}, \dots, x_{\mathrm{GG}}) = \frac{\prod_{b,b' \in \mathcal{B}^2} u_{bb'}^{|w|_{bb'}}}{S(x_{\mathrm{AA}}, \dots, x_{\mathrm{GG}})}$$

3. Accepter uniquement les mots w tels que

$$\frac{|w|_{bb'}}{|m|_{bb'}} \in [1 - \varepsilon(n), 1 + \varepsilon(n)], \forall (b, b') \in \mathcal{B}^2$$

Séquences codantes par Boltzmann en 16 dimensions

- Engendrer w selon la distribution pondérée par $\mathbf{x}$

$$\mathbb{P}(w \mid x_{\mathrm{AA}}, \ldots, x_{\mathrm{GG}}) = \frac{\prod_{b,b' \in \mathcal{B}^2} u_{bb'}^{|w|_{bb'}}}{S(x_{\mathrm{AA}}, \ldots, x_{\mathrm{GG}})}$$

Méthode récursive : ARNm m en entrée $\rightarrow$ Triplets $t_1, \ldots, t_{|m|/3}$

$$\mathcal{Z}_{c,i} = \sum_{w' \in Syn(m_{[i:]})} \prod_{b,b' \in \mathcal{B}^2} x_{bb'}^{|c.w'|_{bb'}}$$

$=$ Poids total des ARNm synonymes avec les triplets $[i, |m|/3]$, précédés par une lettre $c \in \mathcal{B} \cup \{\varepsilon\}$

$$= \begin{cases} \displaystyle\sum_{(b_1,b_2,b_3) \in Syn(t_i)} 1_{c \neq \varepsilon}(x_{cb_1}) \cdot x_{b_1 b_2} \cdot x_{b_2 b_3} \cdot \mathcal{Z}_{b_3, i+1} & \text{if } i \leq |m|/3 \\ 1 & \text{if } i > |m|/3 \end{cases}$$

Génération : Choisir, à chaque étape, un triplet avec probabilité proportionnelle à sa contribution à $\mathcal{Z}_{\star,\star}$ $\rightarrow \mathcal{O}(n)$

Séquences codantes par Boltzmann en 16 dimensions

- Calculer $\mathbf{x} = \{x_{bb'}\}_{b,b' \in \mathcal{B}^2}$ tel que

$$u_{bb'} \cdot \frac{\partial S(u_{\mathrm{AA}}, \ldots, u_{\mathrm{GG}})}{\partial u_{b,b'}}\Bigg|_{\substack{u_{\mathrm{AA}}=x_{\mathrm{AA}} \\ u_{\mathrm{AU}}=x_{\mathrm{AU}} \\ \ldots}} = |m|_{bb'}, \forall b, b' \in \mathcal{B}^2$$

Descente de gradients : Dérivées par *pointage partiel*

$$\mathcal{Z}^{\bullet aa'}_{c,i} = \sum_{w' \in Syn(m_{[i:]})} |w'|_{a,a'} \prod_{b,b' \in \mathcal{B}^2} x_{bb'}^{|c.w'|_{bb'}}$$

$$= \begin{cases} \displaystyle\sum_{(b_1,b_2,b_3) \in Syn(t_i)} \Big(1_{c \neq \varepsilon}(x_{cb_1})\, x_{b_1 b_2}\, x_{b_2 b_3}\, \mathcal{Z}^{\bullet aa'}_{b_3,i+1} \\ + |c.b_1.b_2.b_3|_{aa'}\, 1_{c \neq \varepsilon}(x_{cb_1})\, x_{b_1 b_2}\, x_{b_2 b_3}\, \mathcal{Z}_{b_3,i+1} \Big) & \text{if } i \leq |m|/3 \\ 0 & \text{if } i > |m|/3 \end{cases}$$

Séquences codantes par Boltzmann en 16 dimensions

- Accepter uniquement les mots w tels que

$$\frac{|w|_{bb'}}{|m|_{bb'}} \in [1 - \varepsilon(n), 1 + \varepsilon(n)], \forall (b, b') \in \mathcal{B}^2$$

Rem. : Génération uniforme à composition fixée en dinuc.

Analyse de complexité : Délicate car ...

- Complexité au pire infinie ... en moyenne
- Asymptotique mal définie (taille finie)
- Convergence typique vers Lois Normales 16-dim.
  Moyennes/variances en $\Theta(n)$
  $\rightarrow$ Composition exacte : Rejet en $\Theta(n^8)$
  $\rightarrow$ Composition approchée : Rejet en $\Theta(1)$, $\epsilon(n) \in \Omega(\frac{1}{\sqrt{n}})$

Séquences codantes par Boltzmann en 16 dimensions

- Accepter uniquement les mots w tels que

$$\frac{|w|_{bb'}}{|m|_{bb'}} \in [1 - \varepsilon(n), 1 + \varepsilon(n)], \forall (b, b') \in \mathcal{B}^2$$

Rem. : Génération uniforme à composition fixée en dinuc.

Analyse de complexité : Délicate car ...

- Complexité au pire infinie ... en moyenne
- Asymptotique mal définie (taille finie)
- Convergence typique vers Lois Normales 16-dim.
  Moyennes/variances en $\Theta(n)$
  $\rightarrow$ Composition exacte : Rejet en $\Theta(n^8)$
  $\rightarrow$ Composition approchée : Rejet en $\Theta(1)$, $\epsilon(n) \in \Omega(\frac{1}{\sqrt{n}})$

Séquences codantes par Boltzmann en 16 dimensions

- Accepter uniquement les mots w tels que

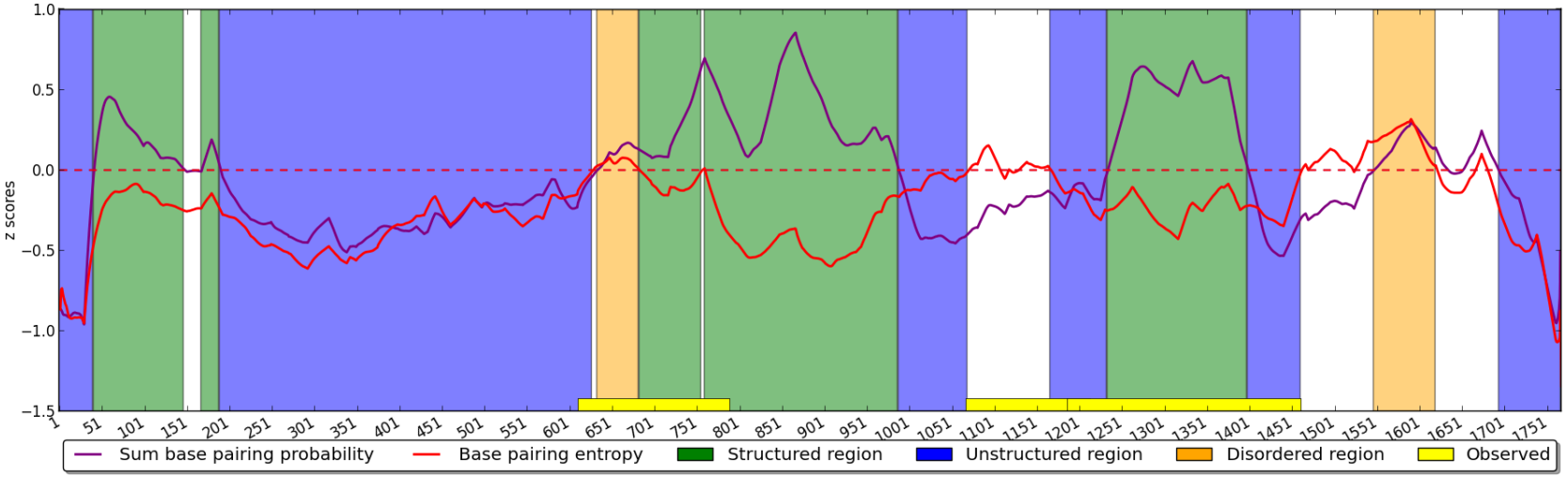
$$\frac{|w|_{bb'}}{|m|_{bb'}} \in [1-\varepsilon(n), 1+\varepsilon(n)], \forall (b,b') \in \mathcal{B}^2$$

Rem. : Génération uniforme à composition fixée en dinuc.

Analyse de complexité : Délicate car ...

- Complexité au pire infinie ... en moyenne
- Asymptotique mal définie (taille finie)
- Convergence typique vers Lois Normales 16-dim.
  Moyennes/variances en $\Theta(n)$
  $\rightarrow$ Composition exacte : Rejet en $\Theta(n^8)$
  $\rightarrow$ Composition approchée : Rejet en $\Theta(1)$, $\epsilon(n) \in \Omega(\frac{1}{\sqrt{n}})$

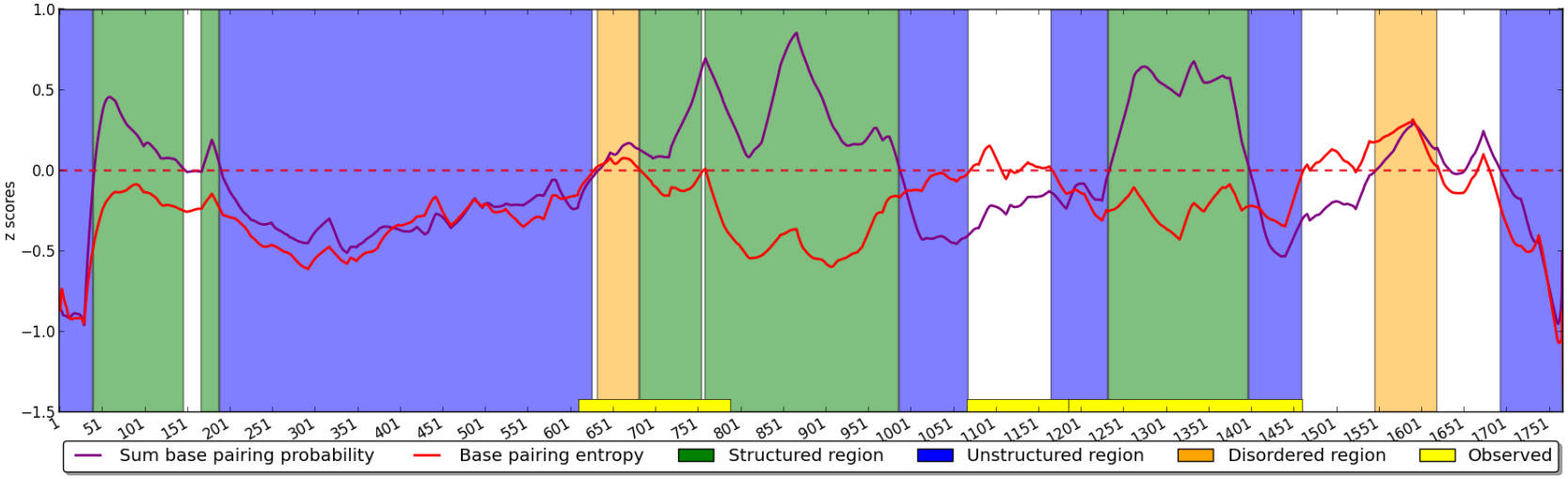
Conclusion



http ://csb.cs.mcgill.ca/sparcs/

- Génération de mélanges préservant les dinucléotides et des contraintes locales → un problème complexe !
- Boltzmann multidimensionnel (en dim. 16) passe à l'échelle, et permet l'étude de séquences de taille $\simeq 1000$!
- Analyse de l'algorithme à finaliser
- Structures combinatoires alternatives ???

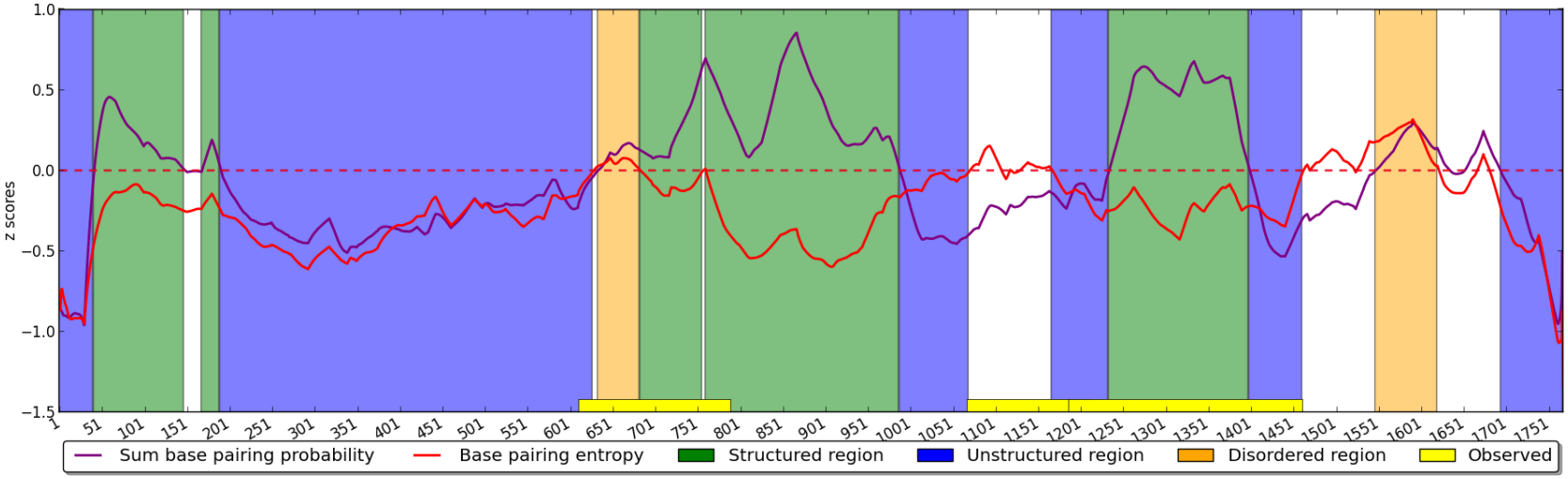
Conclusion



http ://csb.cs.mcgill.ca/sparcs/

- Génération de mélanges préservant les dinucléotides et des contraintes locales → un problème complexe !
- Boltzmann multidimensionnel (en dim. 16) passe à l'échelle, et permet l'étude de séquences de taille ≈ 1000.
- Analyse de l'algorithme à finaliser
- Structures combinatoires alternatives ???

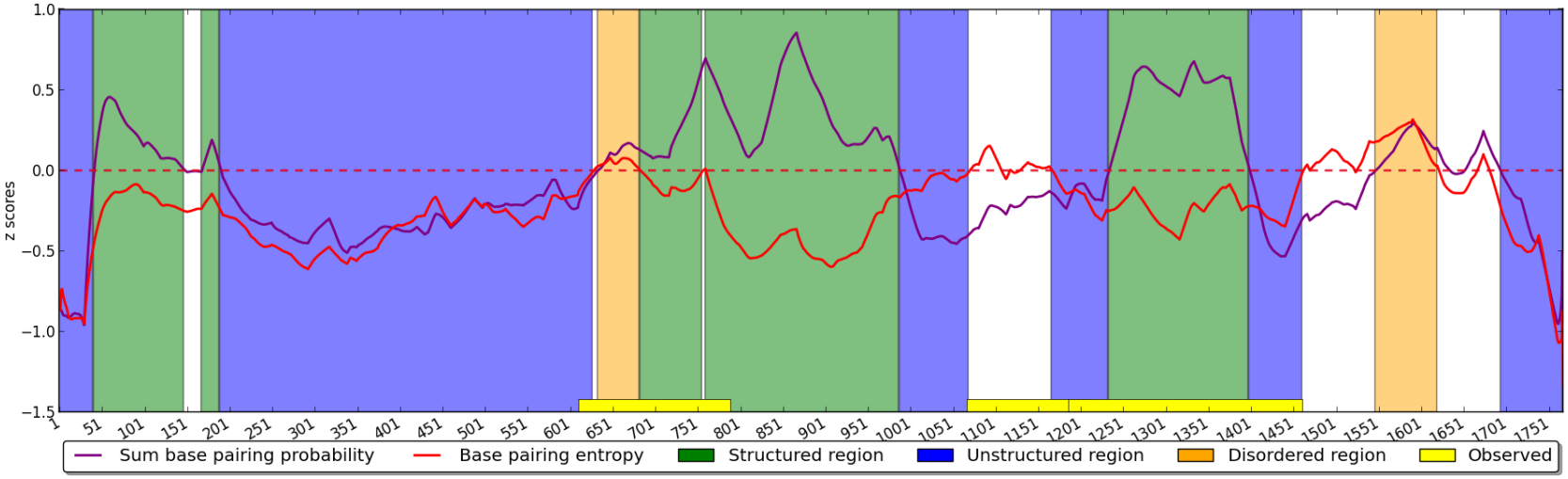
Conclusion



http ://csb.cs.mcgill.ca/sparcs/

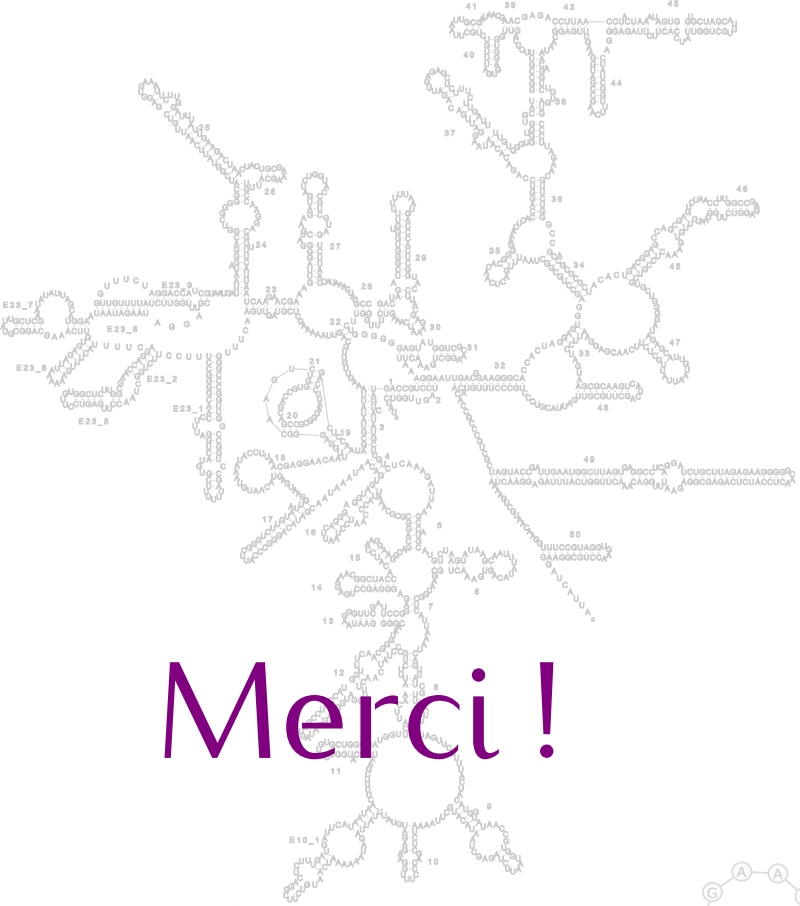
- Génération de mélanges préservant les dinucléotides et des contraintes locales → un problème complexe !
- Boltzmann multidimensionnel (en dim. 16) passe à l'échelle, et permet l'étude de séquences de taille ≈ 1000.
- Analyse de l'algorithme à finaliser
- Structures combinatoires alternatives ???

Conclusion



http ://csb.cs.mcgill.ca/sparcs/

- Génération de mélanges préservant les dinucléotides et des contraintes locales → un problème complexe !
- Boltzmann multidimensionnel (en dim. 16) passe à l'échelle, et permet l'étude de séquences de taille ≈ 1000.
- Analyse de l'algorithme à finaliser
- Structures combinatoires alternatives ???



Merci !