

Compression de séquences d'A.D.N. à base de grammaires minimales

Matthieu Perrin

M.I.T.1, ENS Cachan/Bretagne et Université Rennes 1

3 septembre 2010

Stage effectué du 10 juin 2010 au 9 juillet 2010 à l'IRISA, encadré par François Coste (responsable) et Matthias Gallé, projet Symbiose, IRISA/INRIA Rennes-Bretagne Atlantique.

- 1 Introduction
- 2 Techniques d'encodage
 - Codage à taille fixe
 - Encodeur arithmétique
- 3 Encodage d'une grammaire
 - Linéarisation
 - Encodage par niveaux
 - Ancienneté
 - Anti-dictionnaire
- 4 Conclusion

A.D.N.

- Mot sur quatre bases azotées : A, C, G, T ;
- “Programme” des êtres vivants ;
- Structure mal connue ;
- Contient des répétitions.

Structure

A C T G T T G A C T G T A G C T G

Structure

A C T G T T G A C T G T A G C T G



$$\left\{ \begin{array}{l} 0 \rightarrow 1 T G 1 A G C T G \\ 1 \rightarrow A C T G T \end{array} \right.$$

Structure

A C T G T T G A C T G T A G C T G



$$\begin{cases} 0 \rightarrow 1 T G 1 A G C T G \\ 1 \rightarrow A C T G T \end{cases}$$


$$\begin{cases} 0 \rightarrow 1 T G 1 A G 2 \\ 1 \rightarrow A 2 T \\ 2 \rightarrow C T G \end{cases}$$

Structure

A C T G T T G A C T G T A G C T G

↓

$$\left\{ \begin{array}{l} 0 \rightarrow 1 T G 1 A G C T G \\ 1 \rightarrow A C T G T \end{array} \right.$$

↓

$$\left\{ \begin{array}{l} 0 \rightarrow 1 T G 1 A G 2 \\ 1 \rightarrow A 2 T \\ 2 \rightarrow C T G \end{array} \right.$$

↓

$$\left\{ \begin{array}{l} 0 \rightarrow 1 3 1 A G 2 \\ 1 \rightarrow A 2 T \\ 2 \rightarrow C 3 \\ 3 \rightarrow T G \end{array} \right.$$

Grammaires

$$\left\{ \begin{array}{l} 0 \rightarrow 131AG2 \\ 1 \rightarrow A2T \\ 2 \rightarrow C3 \\ 3 \rightarrow TG \end{array} \right.$$

Vocabulaire

Terminaux : A,C,G,T

Non-terminaux : 1,2,3

Axiome (ou séquence) : 0

Règles de dérivation : $0 \rightarrow 131AG2, 1 \rightarrow A2T, \dots$

Codage

Comment encoder une grammaire

?

- 1 Introduction
- 2 Techniques d'encodage
 - Codage à taille fixe
 - Encodeur arithmétique
- 3 Encodage d'une grammaire
 - Linéarisation
 - Encodage par niveaux
 - Ancienneté
 - Anti-dictionnaire
- 4 Conclusion

Codage

But

- Représentation des données sous forme de bits,
- Représentation uniquement décodable,
- Représentation la plus concise possible.

Codage à taille fixe

$$\left\{ \begin{array}{l} A \rightarrow 00 \\ C \rightarrow 01 \\ G \rightarrow 10 \\ T \rightarrow 11 \end{array} \right.$$

ACTGTTGACTGTAGCT



00 01 11 10 11 11 10 00 01 11 10 11 00 10 01 11

Et avec un symbole de plus ?

$$\left\{ \begin{array}{l} A \rightarrow 000 \\ C \rightarrow 001 \\ G \rightarrow 010 \\ T \rightarrow 011 \\ X \rightarrow 100 \end{array} \right.$$

Encodeur arithmétique

Encodeur arithmétique

- Nombre de symboles au choix,
- Taille des codes de chaque symbole non-entière,
- Encodeur entropique,
- Le modèle peut être changé au cours de l'encodage.

Encodeur arithmétique

Encodeur arithmétique

- Nombre de symboles au choix,
- Taille des codes de chaque symbole non-entière,
- Encodeur entropique,
- Le modèle peut être changé au cours de l'encodage.

Encodeur arithmétique adaptatif

- Distribution initiale équiprobable,
- Le modèle s'adapte en fonction des symboles rencontrés.

- 1 Introduction
- 2 Techniques d'encodage
 - Codage à taille fixe
 - Encodeur arithmétique
- 3 Encodage d'une grammaire
 - Linéarisation
 - Encodage par niveaux
 - Ancienneté
 - Anti-dictionnaire
- 4 Conclusion

Linéarisation

$$\left\{ \begin{array}{l} 0 \rightarrow 131AG2 \\ 1 \rightarrow AT \\ 2 \rightarrow C3 \\ 3 \rightarrow TG \end{array} \right.$$

↓

$$0 \rightarrow 131AG2/1 \rightarrow AT/2 \rightarrow C3/3 \rightarrow TG$$

Linéarisation

$$\left\{ \begin{array}{l} 0 \rightarrow 131AG2 \\ 1 \rightarrow A2T \\ 2 \rightarrow C3 \\ 3 \rightarrow TG \end{array} \right.$$

↓

131AG2/A2T/C3/TG

Encodage par niveaux

Principe

$$\begin{aligned}E_0 &= \Sigma \\E_{i+1} &= \{\gamma \in \Gamma, rhs(\gamma) \in \bigcup_{k=0}^i E_k\} \\E'_0 &= E_0 \\E'_{i+1} &= E_{i+1} - E_i\end{aligned}$$

Encodage par niveaux

}	0	→	717a524g6c336
	1	→	ac
	2	→	gt
	3	→	ag
	4	→	12
	5	→	23
	6	→	33
	7	→	245

ac/gt/ag/#12/23/33/#245/#717a524g6c336

M.T.F.

Alphabet initial	Symbole lu	Symbole codé	Alphabet final
{a, c, g, t}	a	0	{a, c, g, t}
{a, c, g, t}	a	0	{a, c, g, t}
{a, c, g, t}	a	0	{a, c, g, t}
{a, c, g, t}	c	1	{c, a, g, t}
{c, a, g, t}	c	0	{c, a, g, t}
{c, a, g, t}	g	2	{g, c, a, t}
{g, c, a, t}	g	0	{g, c, a, t}
{g, c, a, t}	g	0	{g, c, a, t}
{g, c, a, t}	a	2	{a, g, c, t}
{a, g, c, t}	a	0	{a, g, c, t}
{a, g, c, t}	t	3	{t, a, g, c}
{t, a, g, c}	t	0	{t, a, g, c}
{t, a, g, c}	t	0	{t, a, g, c}

Nouveau modèle

Trois tableaux

Pour chaque symbole σ ,

Adaptativité (a) : Nombre d'occurrences déjà rencontrées de σ

Autorisation (b) : Booléen au choix

Ancienneté (c) : Nombre de symboles différents rencontrés depuis la dernière occurrence de σ

Probabilité proportionnelle à

$$(a + k * (n - c)) * b$$

n : nombre de symboles

k : choisi pour favoriser l'un des aspects.

Anti-dictionnaire

Formule

$$\forall \gamma = \gamma_0 \dots \gamma_{|\gamma|} \in \Gamma, \forall \delta = \delta_0 \dots \delta_{|\delta|} \in \Gamma, \forall 0 \leq i \leq |\delta| - |\gamma|,$$
$$(\delta_i = \gamma_0, \dots, \delta_{i+|\gamma|-1} = \gamma_{|\gamma|-1}) \Rightarrow P(\delta_{i+|\gamma|} = \gamma_{|\gamma|}) = 0$$

Anti-dictionnaire

Formule

$$\forall \gamma = \gamma_0 \dots \gamma_{|\gamma|} \in \Gamma, \forall \delta = \delta_0 \dots \delta_{|\delta|} \in \Gamma, \forall 0 \leq i \leq |\delta| - |\gamma|,$$

$$(\delta_i = \gamma_0, \dots, \delta_{i+|\gamma|-1} = \gamma_{|\gamma|-1}) \Rightarrow P(\delta_{i+|\gamma|} = \gamma_{|\gamma|}) = 0$$

a c g / g t / c a c ?

- 1 Introduction
- 2 Techniques d'encodage
 - Codage à taille fixe
 - Encodeur arithmétique
- 3 Encodage d'une grammaire
 - Linéarisation
 - Encodage par niveaux
 - Ancienneté
 - Anti-dictionnaire
- 4 Conclusion

Résultats avec des grammaires m.c.

Sequence	XM	Sequitur	A.A.C	Améliorations
CHMPXX	1.6577	-	2.0931	-
CHNTXX	1.6068	2.12	2.2179	2.1726
HEHCMVCG	1.8426	2.12	2.2119	-
HUMDYSTROP	1.9031	2.34	2.2884	2.1900
HUMGHCSA	0.9828	1.86	1.6732	1.6146
HUMHBB	1.7513	-	2.1760	2.1222
HUMHDAB	1.6671	-	2.1948	2.1329
HUMHPRTB	1.7361	-	2.1990	2.13326
MPOMTCG	8768	-	2.2012	-
MTPACG	1.8447	2.16	2.1457	2.0895
VACCG	1.7649	2.11	2.1008	2.0609

