Algorithmique de l'échafaudage

Annie Chateau

Introduction

Le séquençage produit des séquences courtes, les reads, qu'il faut *assembler* pour reconstituer la séquence génomique.

L'assemblage est un problème *difficile*, que l'on peut modéliser par le problème Shortest Superstring.

Problème : SSP est NP-complet

Introduction

Stratégies pour l'assemblage :

- Stratégie gloutonne
- Algorithme OLC (Overlap-Layout-Consensus)
- Approche graphes de De Bruijn

Introduction

Résultats : des ensembles de contigs de tailles variables, déconnectés les uns des autres

On n'a pas encore la séquence complète...

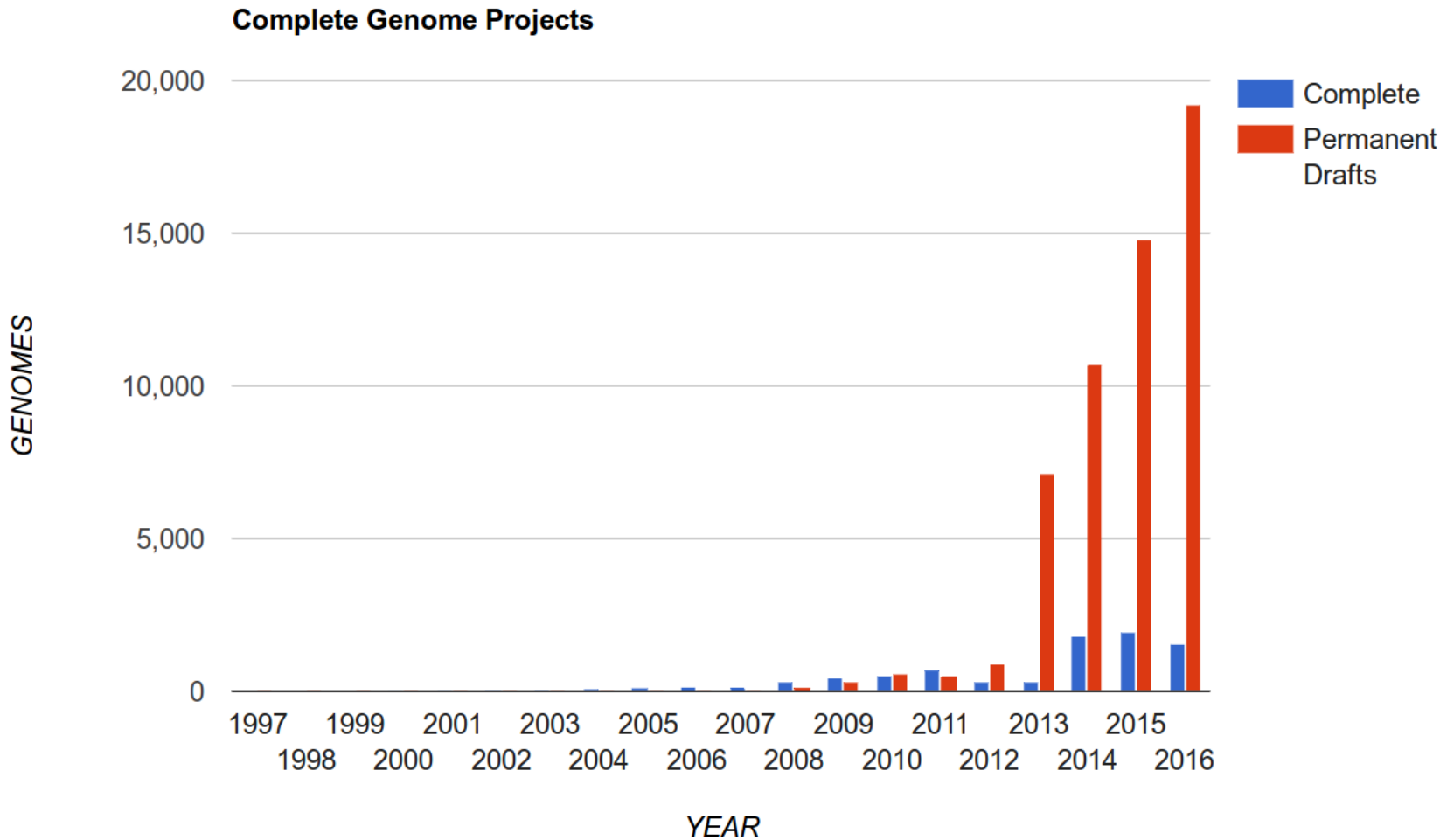
⇒ Besoin d'une étape supplémentaire pour produire des séquences de longueur comparables aux chromosomes.

Introduction

Motivations

- Pouvoir observer des phénomènes à l'échelle du génome
- Améliorer la qualité des génomes de référence
- Beaucoup de génomes sont à l'état de "brouillons" dans les bases de données

Introduction



https://gold.jgi.doe.gov/statistics

Le problème de l'échafaudage

Pour pouvoir déterminer l'ordre et l'orientation relative des contigs, on doit pouvoir disposer de :

- Informations entre les contigs
  - données d'appariement
  - données phylogénétiques
  - données de long reads
  - ...
- Un poids relatif à ces informations
  - le nombre de reads pairés
  - une mesure probabiliste
  - la profondeur de couverture
  - ...

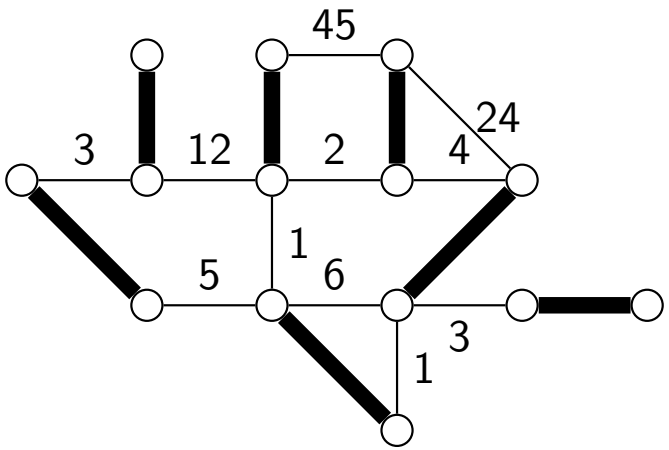
Le problème de l'échafaudage

On modélise les données sous forme d'un graphe $G = (V, E)$:

- **Sommets :** extrémités des contigs
- **Arêtes :**
  - entre deux extrémités d'un même contig (arêtes de contigs)
  - entre extrémités de contigs différents (arêtes inter-contigs)

Information de poids : $w : E \to \mathbb{R}$.

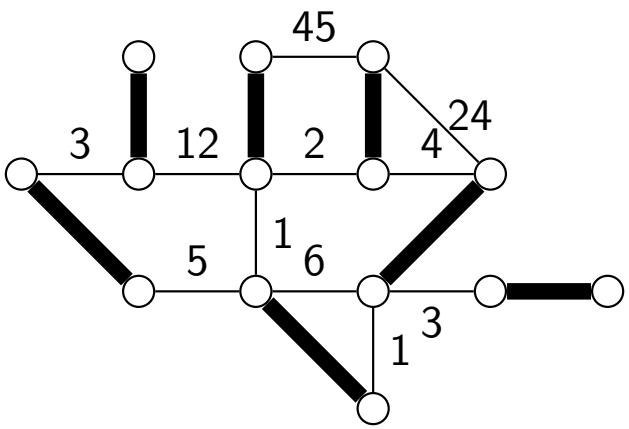
Le problème de l'échafaudage

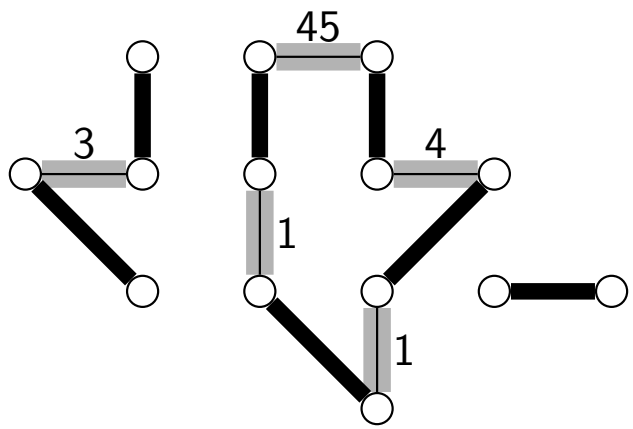


Graphe non orienté à $2n$ sommets, muni d'un couplage parfait (les arêtes de contig)

On travaille avec des paramètres structurels sur la solution : σ_p chemins (chromosomes linéaires) et σ_c cycles (chromosomes circulaires)

Le problème de l'échafaudage





$\sigma_p = 2$ et $\sigma_c = 2$

Formalisation du problème

	Question		Problème
Entrée	Décision	Optim.	

Formalisation du problème

	Question		Problème
Entrée	Décision	Optim.	
$G = (V, E)$, $w : E \to \mathbb{N}$, M^* couplage parfait, σ_p, σ_c, $k \in \mathbb{N}$			

Formalisation du problème

	Question		Problème
Entrée	Décision	Optim.	
$G = (V, E)$, $w : E \to \mathbb{N}$, M^* couplage parfait, σ_p, σ_c, $k \in \mathbb{N}$	Existe-t'il une collection S de σ_p chemins et σ_c cycles alternés couvrant G		SSCA (Strict Scaffolding)

Formalisation du problème

	Question		Problème
Entrée	Décision	Optim.	
$G = (V, E)$, $w : E \to \mathbb{N}$, M^* couplage parfait, σ_p, σ_c, $k \in \mathbb{N}$	Existe-t'il une collection S de σ_p chemins et σ_c cycles alternés couvrant G	telle que $w(S) \geq k$ (resp. $w(S) \leq k$)	MAX-SSCA (resp. MIN-SSCA)

Formalisation du problème

	Question		Problème
Entrée	Décision	Optim.	
$G = (V, E)$, $w : E \to \mathbb{N}$, M^* couplage parfait, σ_p, σ_c, $k \in \mathbb{N}$	Existe-t'il une collection S de σ_p chemins et σ_c cycles alternés couvrant G	telle que $w(S) \geq k$ (resp. $w(S) \leq k$)	
$G = (V, E)$, $w : E \to \mathbb{N}$, M^* couplage parfait, $\sigma_p, \sigma_c, k \in \mathbb{N}$, $m : E \to \mathbb{N}$			

Formalisation du problème

	Question		Problème
Entrée	Décision	Optim.	
$G = (V, E)$, $w : E \to \mathbb{N}$, M^* couplage parfait, σ_p, σ_c, $k \in \mathbb{N}$	Existe-t'il une collection S de σ_p chemins et σ_c cycles alternés couvrant G	telle que $w(S) \geq k$ (resp. $w(S) \leq k$)	
$G = (V, E)$, $w : E \to \mathbb{N}$, M^* couplage parfait, $\sigma_p, \sigma_c, k \in \mathbb{N}$, $m : E \to \mathbb{N}$	Existe-t'il une collection S de σ_p marches ouvertes et σ_c marches fermées alternées couvrant G		MSSCA (Multi Strict Scaffolding)

Formalisation du problème

	Question		Problème
Entrée	Décision	Optim.	
$G = (V, E)$, $w : E \to \mathbb{N}$, M^* couplage parfait, σ_p, σ_c, $k \in \mathbb{N}$	Existe-t'il une collection S de σ_p chemins et σ_c cycles alternés couvrant G	telle que $w(S) \geq k$ (resp. $w(S) \leq k$)	
$G = (V, E)$, $w : E \to \mathbb{N}$, M^* couplage parfait, $\sigma_p, \sigma_c, k \in \mathbb{N}$, $m : E \to \mathbb{N}$	Existe-t'il une collection S de σ_p marches ouvertes et σ_c marches fermées alternées couvrant G	telle que $w(S) \geq k$ (resp. $w(S) \leq k$)	MAX-MSSCA (resp. MIN-MSSCA)

Formalisation du problème

Entrée	Question		Problème
	Décision	Optim.	
$G = (V, E)$, $w : E \to \mathbb{N}$, M^* couplage parfait, σ_p, σ_c, $k \in \mathbb{N}$	Existe-t'il une collection S de $\leq\sigma_p$ chemins et $\leq\sigma_c$ cycles alternés couvrant G	telle que $w(S) \geq k$ (resp. $w(S) \leq k$)	SCA (Scaffolding) MAX-SCA (resp. MIN-SCA)
$G = (V, E)$, $w : E \to \mathbb{N}$, M^* couplage parfait, $\sigma_p, \sigma_c, k \in \mathbb{N}$, $m : E \to \mathbb{N}$	Existe-t'il une collection S de $\leq\sigma_p$ marches ouvertes et $\leq\sigma_c$ marches fermées alternées couvrant G	telle que $w(S) \geq k$ (resp. $w(S) \leq k$)	MSCA (Multi Scaffolding) MAX-MSCA (resp. MIN-MSCA)

Et maintenant ?

Une fois le problème formalisé, on va chercher à le classifier...

À votre avis ?

Et maintenant ?

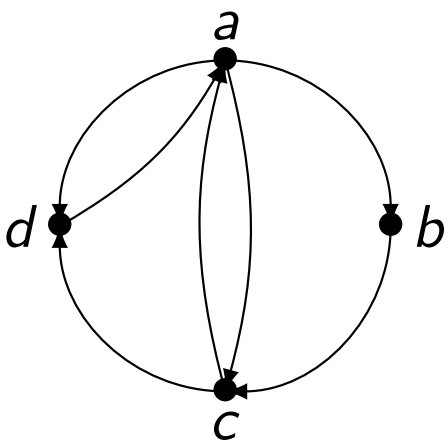
Une fois le problème formalisé, on va chercher à le classifier...

À votre avis ?

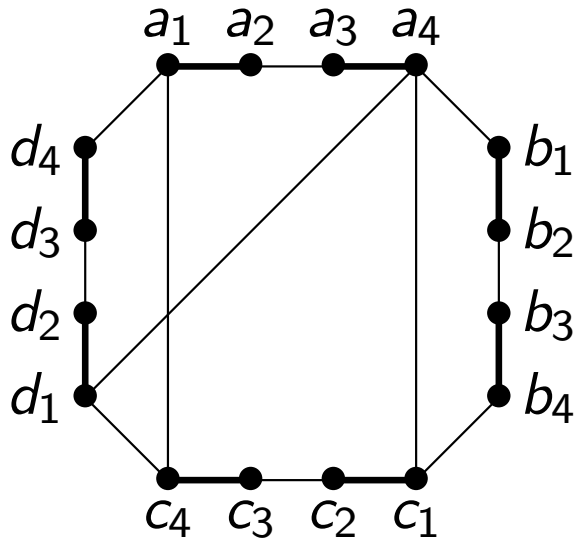
Ils sont tous NP-complets dans le cas général !

Complexité

Idée de la preuve : réduction depuis le TSP orienté



$\Rightarrow$

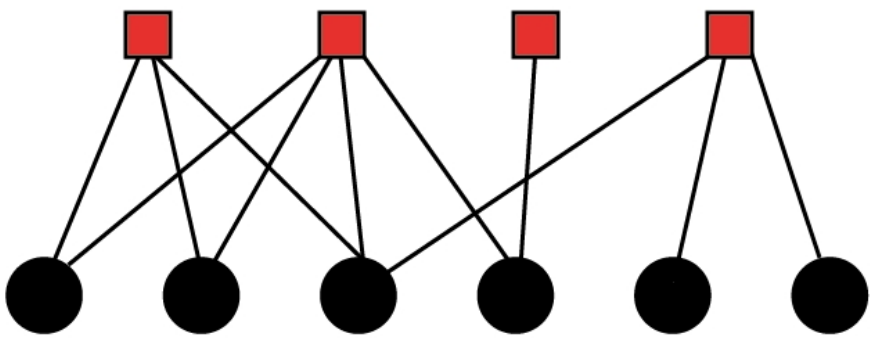


Complexité

Que faire ?

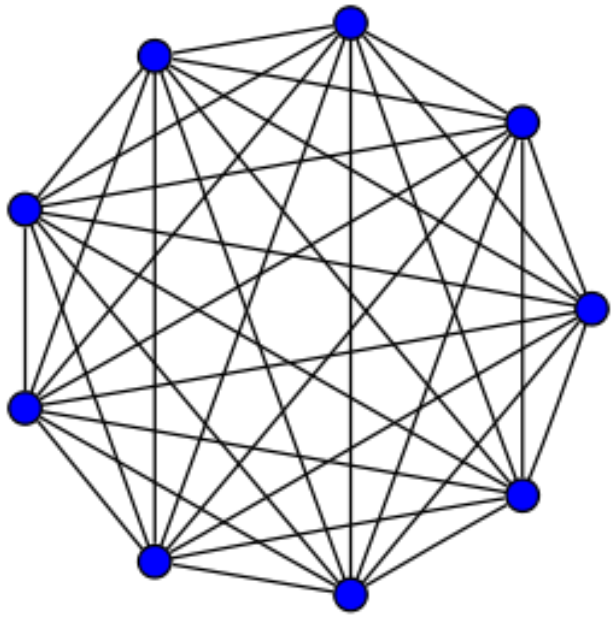
Première idée : Chercher des classes de graphes particulières où le problème pourrait devenir polynomial.

Exemple : le problème du Vertex Cover est NP-complet dans les graphes généraux, il devient polynomial dans les graphes *bipartis*



Complexité

Première idée : Chercher des classes de graphes particulières où le problème pourrait devenir polynomial.
Deuxième exemple : le problème du chemin hamiltonien est NP-complet dans les graphes généraux, il devient polynomial dans les graphes *complets*



Complexité

Les cas polynomiaux sont trop éloignés des graphes réels

Complexité

Les cas polynomiaux sont trop éloignés des graphes réels

Stratégies :

- Algorithmes polynomiaux, mais en essayant de contrôler l'écart à l'optimal

Complexité

Les cas polynomiaux sont trop éloignés des graphes réels

Stratégies :

- Algorithmes polynomiaux, mais en essayant de contrôler l'écart à l'optimal
- Algorithmes exacts mais FPT

Complexité

Les cas polynomiaux sont trop éloignés des graphes réels

Stratégies :

- Algorithmes polynomiaux, mais en essayant de contrôler l'écart à l'optimal
- Algorithmes exacts mais FPT
- Méthodes exactes avec résolution générale : CSP, PLNE

Conclusion

- Un problème NP-complet n'est pas forcément désespérant
- Les heuristiques peuvent donner de bons résultats en pratique
- Les méthodes exactes ne sont pas forcément à jeter aux orties
- Un problème NP-complet peut en cacher un autre !

Finition

Il reste toujours des trous, des incohérences, des ambiguïtés à lever.

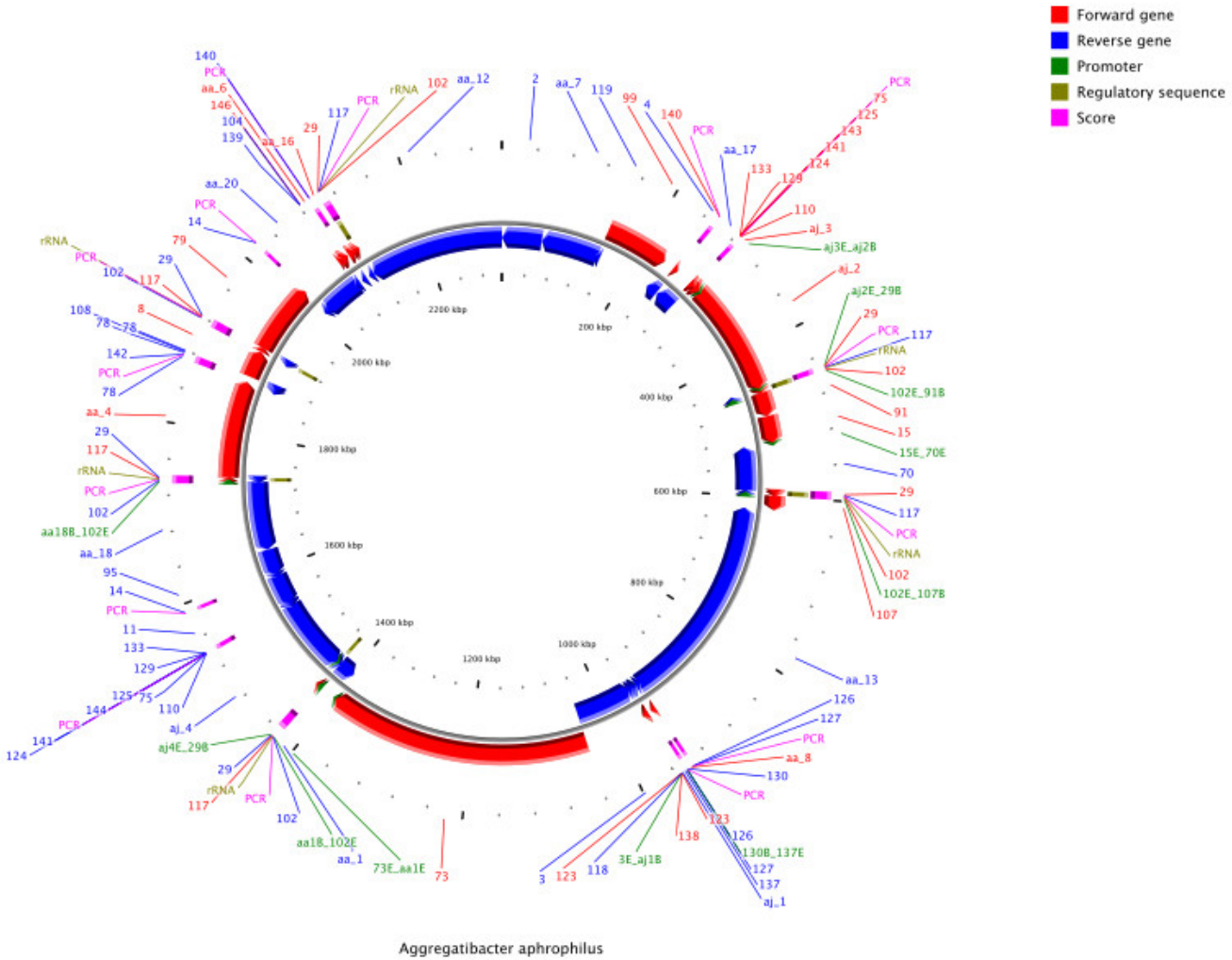
On réalise des expériences supplémentaires en laboratoire, on nettoie "à la main", afin de valider la correction de l'assemblage final.

C'est une opération très coûteuse, qui peut prendre des mois...

$\Rightarrow$ le temps et l'effort ne sont justifiés que pour les génomes "à haute priorité"

...les autres restent à l'état de brouillons (*draft genomes*)

Finition



Nagarajan et al. BMC Genomics 2010