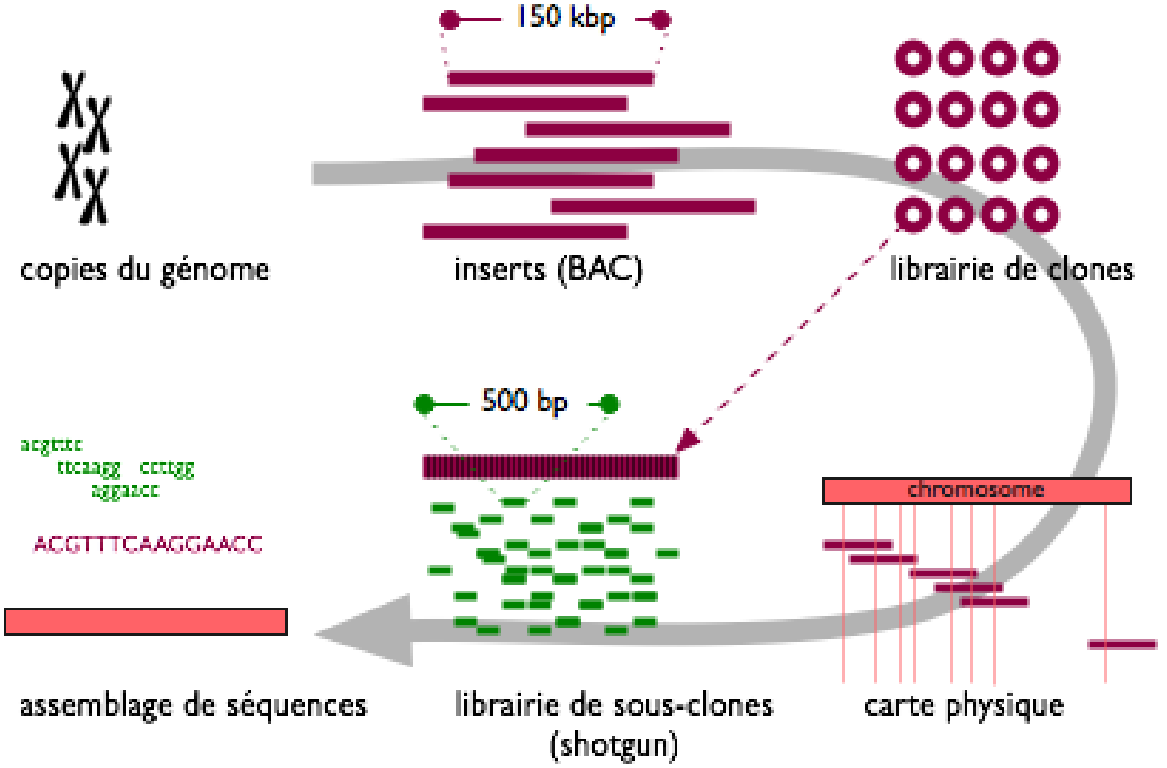
Séquençage d'ADN

ou comment lire le Livre en pièces

Approche hiérarchique



Séquençage d'un BAC

Le Livre du génome humain : 1 000 000 pages

longueur de l'insert de BAC : cca. 150000 pb
(cca 50 pages dans le Livre)

fragments shotgun :
les joindre à partir des chevauchements

```
s1: AATGCC
s2:    GCCTTACAC
s3:         ACACTG
s4:           ACTGAAGG
s5:              GAAGGTTTA
---------------------
B : AATGCCTTACACTGAAGGTTTA
```

TIGR assembler (1995)

une approche glouton utilisée pour assembler le génome de *H. influenzae* (1.8Mbp)

1. analyse de k-mers dans les fragments : chevauchements potentiels entre fragments avec k-mers partagés (score determiné par nombre de k-mers en commun)
2. identification de fragments avec régions répétées
3. initialisation de la séquence assemblée (contig) par un fragment
4. répéter : ajout du meilleur fragment à la séquence assemblée

En 2 : fragment avec trop de chevauchements potentiel=repeat

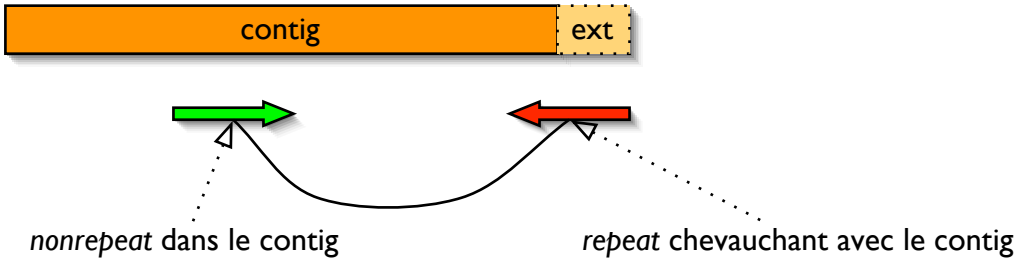
Sutton et al, *Genome Sci. Techn.* 1 : 9.

TIGR 2

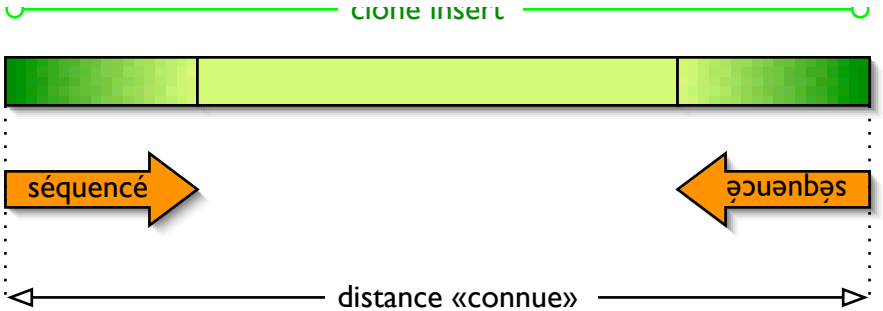
trouver meilleur fragment à ajouter : alignement local (Smith-Waterman), pour tous les fragments avec chevauchements potentiels

contig finit s'il n'y a plus de fragments à ajouter : à la frontière d'une région répétée ou à un vrai trou

extension dans la région répétée : utiliser des séquences shotgun appariées (*mate pairs*)



Mate pairs

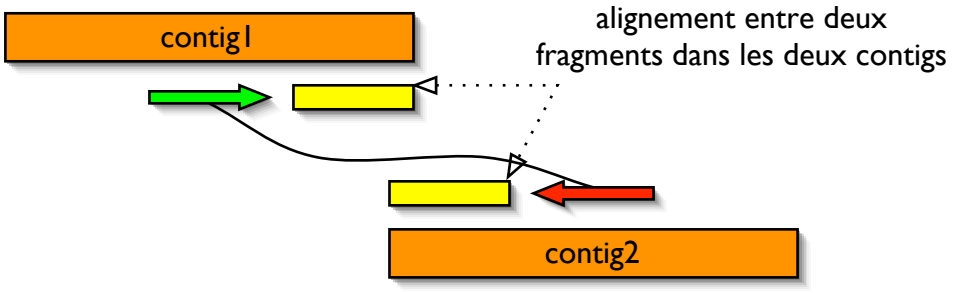


distances : 2k (M13), 10k (plasmid), 100k (BAC)

aident à orienter des contigs, à construire des ossatures (*scaffolds*), et à traverser des régions répétées

TIGR 3

joindre des contigs si évidence par chevauchements et mate pairs



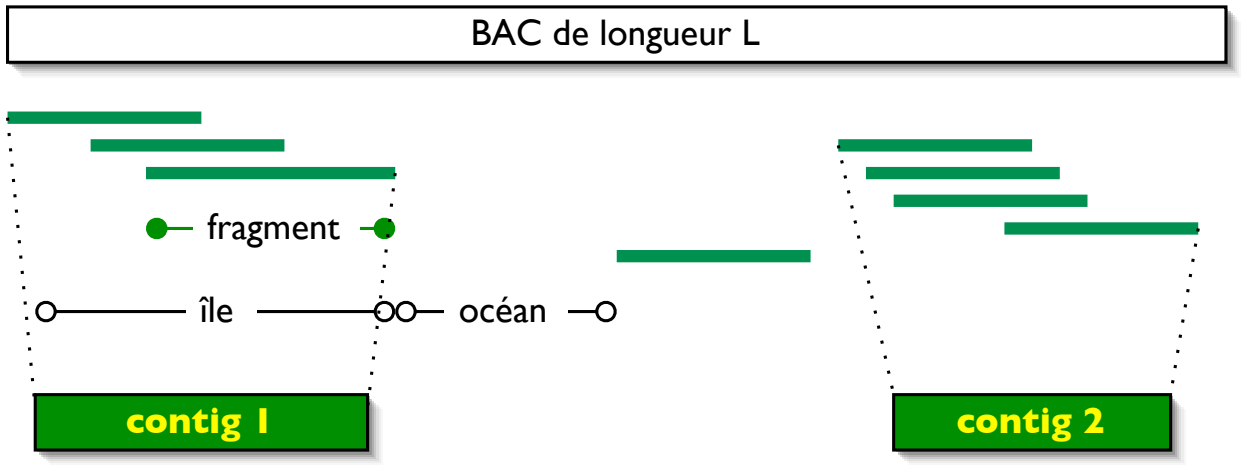
Assemblage : overlap-layout-consensus

Overlap : déterminer les chevauchements parmi les séquences shotgun

Layout : déterminer l'ordre des séquences shotgun

Consensus : déterminer la séquence des contigs

Couverture par fragments de shotgun



nombre de fragments : n

longueur d'un fragment : ℓ

longueur du BAC : L

couverture (*coverage*) : $c = n\ell/L$

Questions

- nombre de positions couvertes par au moins 1 (k) fragments
- nombre des îles
- nombre de fragments dans un île
- nombre des contigs (ou îles avec au moins k fragments)
- longueur d'un île

Couverture par ≥ 1 fragment

Thm. La probabilité qu'une position du BAC est couverte par au moins un fragment est $\approx (1 - e^{-c})$.

Preuve.

Probabilité qu'un fragment fixé couvre la position : $p = \ell/(L - \ell + 1) \approx \ell/L$

Probabilité qu'aucun fragment ne la couvre pas : $(1-p)^n \approx \left(1 - \frac{\ell}{L}\right)^n$.

Approximation : $(1 - a/x)^x \approx e^{-a}$.

Couverture par k fragments

Thm. Pour $k = 0, 1, \cdots$, la probabilité qu'une position du BAC est couverte par k fragments exactement est $p_k \approx \frac{c^k}{k!} e^{-c}$.

Preuve. Probabilité égale à $\binom{n}{k} p^k (1-p)^{n-k} \approx p_k$ (convergence vers la distribution Poisson avec paramètre $\lambda = np = c$).

Nombre des océans

Thm. Le nombre des océans est $\approx ne^{-c} = \frac{L}{\ell} c e^{-c}$.

Preuve. Probabilité qu'un fragment fixé est le dernier fragment d'un île observé :

$$p_{\text{dernier}} = \left(1 - \frac{\ell}{L}\right)^{n-1}.$$

+approximation comme avant

Espérance du nombre des océans = np.

Taille d'un île

Thm. L'espérance du nombre de fragments dans un île est $\approx e^c$.

Preuve.
Position du fragment définie par la position de son côté droit : variables aléatoires $X_1, X_2, \ldots, X_n$.

Fixons un fragment (X_1). Quelle est la position Y_1 du premier fragment après X_1 ?
Pout tout $h > 0$, probabilité que $Y_1 > X_1 + h\ell$ est

$$J(h) = \left(1 - \frac{h\ell}{L}\right)^{n-1} \approx \left(1 - \frac{ch}{n}\right)^n \approx e^{-ch}.$$

Taille d'un île — cont.

Soit M le nombre des fragments dans l'île.

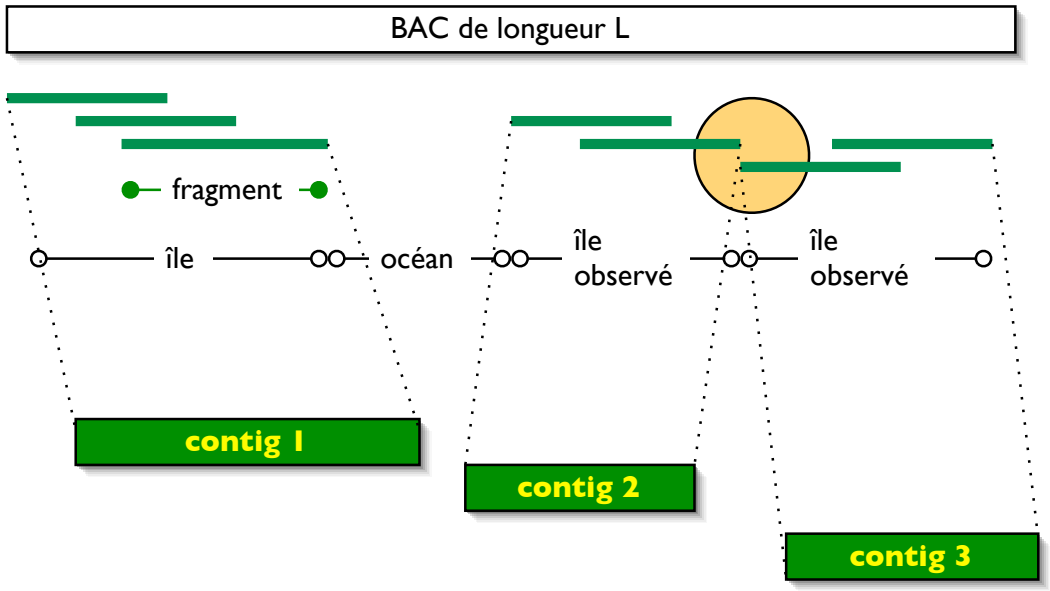
Considérons le premier fragment de l'île. Probabilité que c'est un île singulaire ($M = 1$) : $p_1 = J(1)$.

Probabilité que $M = k$: $p_k = \Big(1 - J(1)\Big)^{k-1} J(1)$ c'est la distribution géometrique !

Espérance de M est $1/J(1)$.

Modèle statistique pour chevauchements

chevauchement minimal : $\theta\ell$, $0 < \theta < 1$

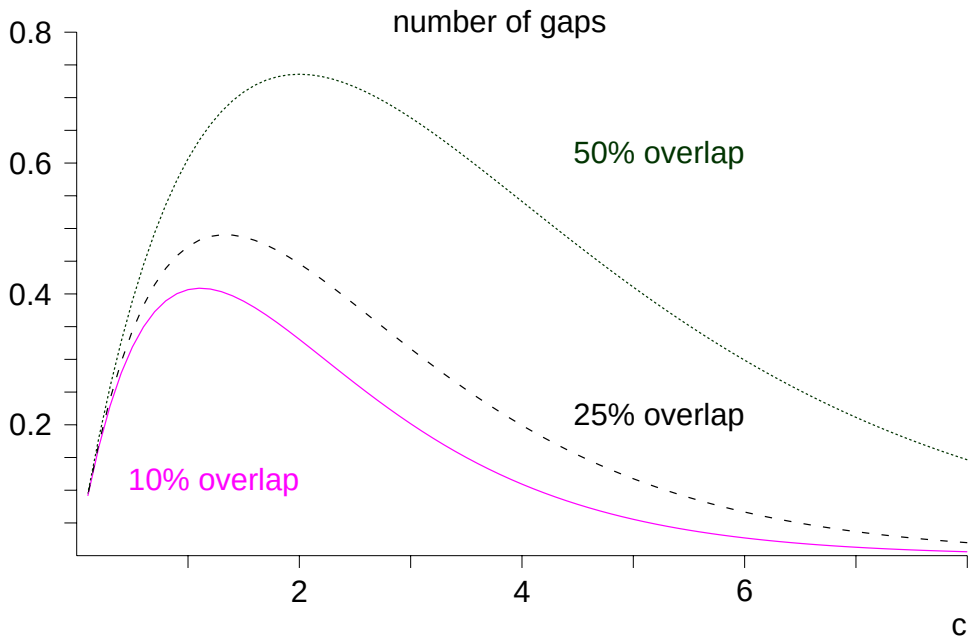


Chevauchements — cont.

Thm. Le nombre des océans est $\approx ne^{-c(1-\theta)} = \frac{L}{\ell} c e^{-c(1-\theta)}$.

Thm. L'espérance du nombre de fragments dans un île est $\approx e^{c(1-\theta)}$.

Chevauchements — cont.



couvertures typiques : 5X (“half shotgun”) 10X (“full shotgun”)
avec $\ell = 500$, $L = 150000$, $n = 1500$ or $n = 3000$.

Application : filtrage de contigs

compression de régions répétées lors d'assemblage

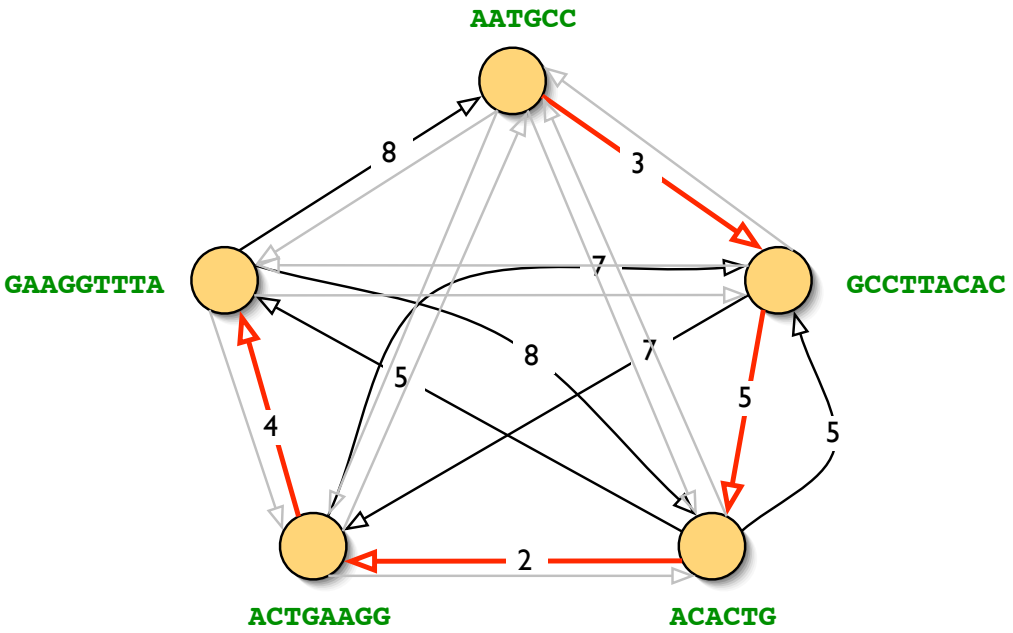
peut être identifiée par la densité de séquences shotgun dans le contig

log-likelihood ratio pour nombre de séquences k dans un contig de longueur $\rho\ell$:
- hypothèse 0 : pas de compression (couverture c)
- hypothèse alt. : compression (couverture $2c$)

$$\mathsf{LLR} = k \log 2 - \rho c \log e$$

Myers et al. Science 287 :2196

Graphe de chevauchements

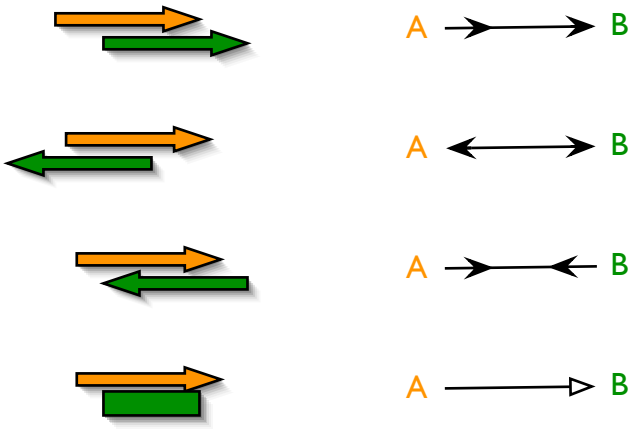


[problème d'orientation ignoré]

layout = chemin dans le graphe

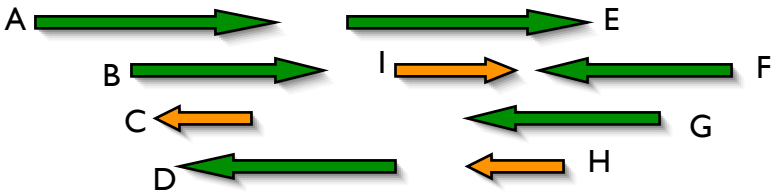
Layout

Catégories de chevauchements (orientation inconnue des fragments)



Myers, *J. Comput. Biol.* 2 : 275.

Exemple



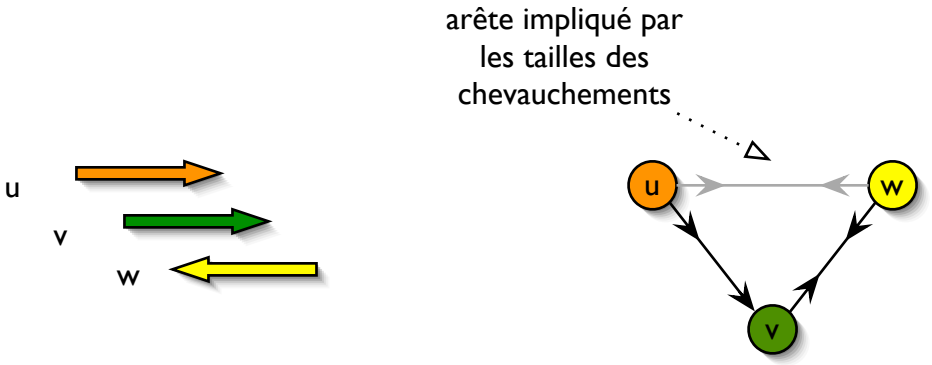
compter les flèches arrivant à un sommet : layout=chemin avec arêtes de chevauchements propres+ forêts avec arêtes de contention

Myers, *J. Comput. Biol.* 2 : 275.

Simplification du graphe

1. enlever les arêtes de contention

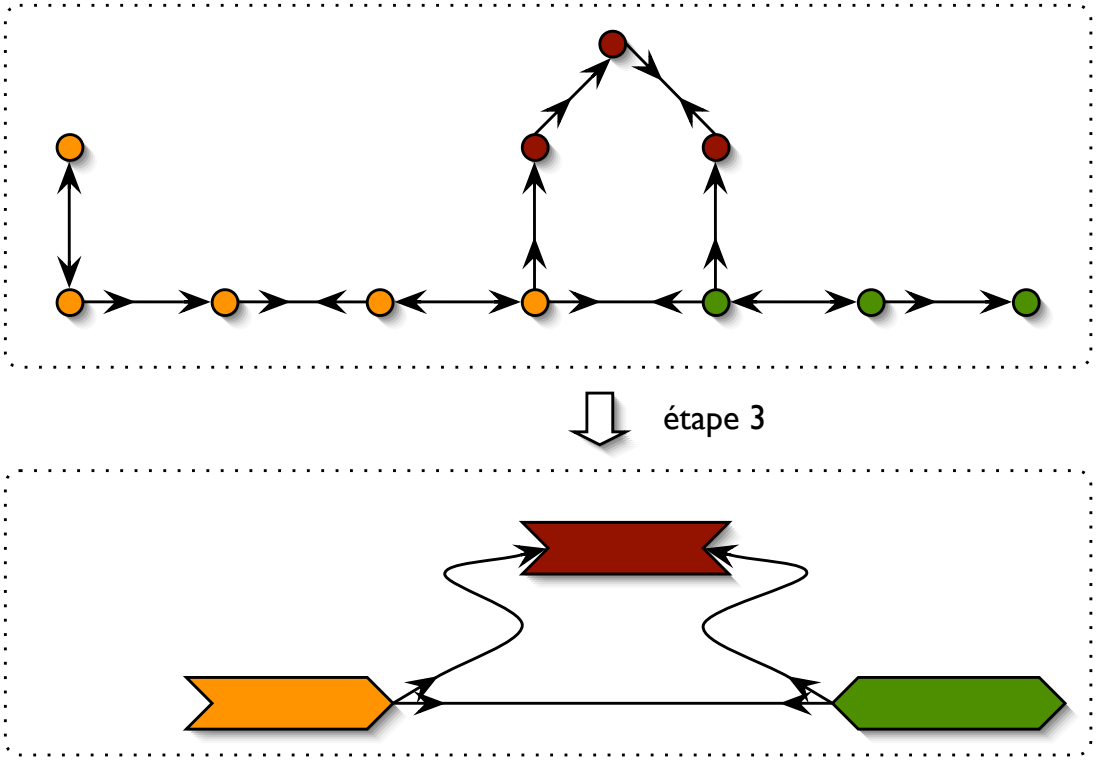
2. enlever des arêtes «transitifs» : si $u \rightleftharpoons v$, $v \rightleftharpoons w$ et $u \rightleftharpoons w$ sont des arêtes compatibles (orientation+taille de chevauchements), alors enlever $u \rightleftharpoons w$



Myers, *J. Comput. Biol.* 2 : 275.

Simplification du graphe 2

3. collapser des chemins : «super-sommets» ou contigs



Myers, *J. Comput. Biol.* 2 : 275.

Séquence de consensus

Le layout donne la position approximative de chaque fragment. Trouver la séquence de consensus : alignement multiple

Profile : enregistrer la fréquence de symbols dans l'alignement multiple

Joindre les séquences consécutives au contig dans l'ordre spécifié dans la phase *layout*, maintenir un profile dans le contig

Problème : alignement d'une séquence à un profile (dans une bande autour de la position approximative)

$\Rightarrow$ contigs