

Détection de domaines dans des séquences génomiques : un problème de couverture optimale

Philippe Veber Sébastien Tempel Rumen Andonov
Dominique Lavenier Jacques Nicolas

Projet SYMBIOSE, IRISA

21 Février 2007

Motivations

Formalisation

Modélisation

Contexte : “lire” (et comprendre ?) le génome

- ▶ comprendre la dépendance entre **séquence** et **fonctionnement** de l'organisme
- ▶ structure des génomes (= grammaire)

Contexte : “lire” (et comprendre ?) le génome

- ▶ comprendre la dépendance entre **séquence** et **fonctionnement** de l'organisme
- ▶ structure des génomes (= grammaire)
- ▶ essayons !

```
TTTATGATCCGATTCAATCTAAACCGTTCAATAAAA  
CCTTCAATCTAAACCGTTCAACAAAAATAAGGAATC  
AAATATGATCACATTTTCATCCCTAAAAACACTATAT  
TCAATAATATCCAAATCATATATTATGATTTTTTCAC  
TTTATAAGTTTAGGATATCAAATTTATTCAAATATT  
ACCGATTGTCCGCGGTAAACCGCGGGTTAAAACCTA
```

Contexte : “lire” (et comprendre ?) le génome

- ▶ comprendre la dépendance entre **séquence** et **fonctionnement** de l'organisme

- ▶ structure des génomes (= grammaire)

- ▶ essayons !

```
TTTATGATCCGATTCAATCTAAACCGTTCAATAAAA  
CCTTCAATCTAAACCGTTCAACAAAAATAAGGAATC  
AAATATGATCACATTTTCATCCCTAAAAACACTATAT  
TCAATAATATCCAAATCATATATTATGATTTTTTCAC  
TTTATAAGTTTAGGATATCAAATTTATTCAAATATT  
ACCGATTGTCCGCGGTAAACCGCGGGTTAAAACCTA
```

- ▶ **modularité** des génomes

- ▶ existence de briques élémentaires ?
- ▶ (re)composition de ces briques ?

Contexte : “lire” (et comprendre ?) le génome

- ▶ comprendre la dépendance entre **séquence** et **fonctionnement** de l'organisme

- ▶ structure des génomes (= grammaire)

- ▶ essayons !

```
TTTATGATCCGATTCAATCTAAACCGTTCAATAAAA  
CCTTCAATCTAAACCGTTCAACAAAAATAAGGAATC  
AAATATGATCACATTTTCATCCCTAAAAACACTATAT  
TCAATAATATCCAAATCATATATTATGATTTTTTCAC  
TTTATAAGTTTAGGATATCAAATTTATTCAAATATT  
ACCGATTGTCCGCGGTAAACCGCGGGTTAAAACCTA
```

- ▶ **modularité** des génomes

- ▶ existence de briques élémentaires ?
- ▶ (re)composition de ces briques ?

- ▶ étude des génomes par comparaisons

- ▶ inter-espèces = **conservation**
- ▶ au sein d'un génome = **séquences répétées**

Motivations : visualisation d'une famille de séquences

- ▶ soit une séquence d'intérêt
- ▶ on obtient une famille de séquences par recherche dans les génomes
- ▶ dans la famille, les séquences sont significativement similaires (p-valeur)
- ▶ comment **visualiser** leur ressemblance ?
 - ▶ faire apparaître les domaines en commun (= briques élémentaires)
 - ▶ **abstraction** des séquences en suites de **domaines**

Objectifs

- ▶ formaliser un type de représentation
 - ▶ séquence = suite de domaines
 - ▶ domaine = motif = séquence avec erreurs
 - ▶ approximation des séquences = couverture maximale avec le moins d'erreurs possible
 - ▶ représentation simple = utiliser le moins de domaines différents
- ▶ proposer un cadre algorithmique pour sa construction
 - ▶ en entrée : séquences, domaines autorisés
 - ▶ en sortie : codage des séquences en suite de domaines

Motivations

Formalisation

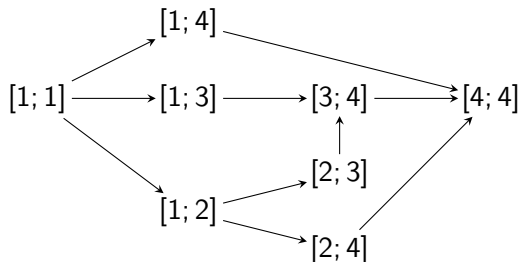
Modélisation

Graphe d'intervalle

Definition

Pour un entier n , le **graphe d'intervalle** $\mathcal{I}(n)$ est défini par :

- ▶ sommets :
 - ▶ intervalles $[i; j]$ pour $1 \leq i < j \leq n$,
 - ▶ $s = [1; 1]$ **entrée**,
 - ▶ $t = [n; n]$ **sortie**.
- ▶ arcs : $[i; j] \rightarrow [k; l]$ si $|k - j| < t$



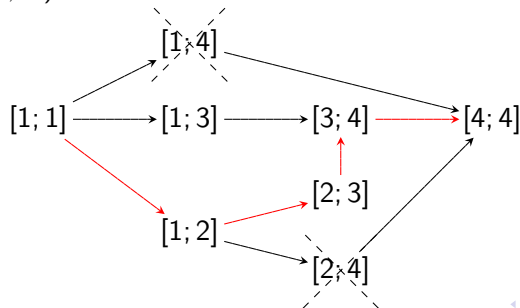
Graphe de couverture

Definition

Pour une séquence u , un ensemble de domaines D , le **graphe de couverture** $\mathcal{C}(u, D)$ est défini comme :

- ▶ le sous graphe de $\mathcal{I}(|u|)$
- ▶ obtenu en retirant les sommets $[i; j]$ tels que $u_i \dots u_j$ ne correspond à aucun domaine de D , sauf s et t

Une **couverture** de la séquence u est un chemin de s à t dans $\mathcal{C}(u, D)$.



Pondération

Le **graphe de couverture** est muni d'une **pondération des arcs** qui favorise les arcs :

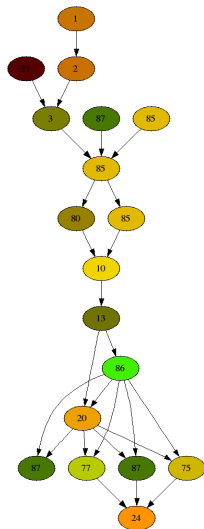
- ▶ dont les sommets ressemblent à des domaines dans D
- ▶ qui relie 2 intervalles proches sur la séquence

poids d'une couverture \equiv poids du chemin.

D est aussi muni d'une fonction de coût d_i $i \in D$

Un exemple

Le graphe de couverture pour
une séquence :



Le problème

Entrée :

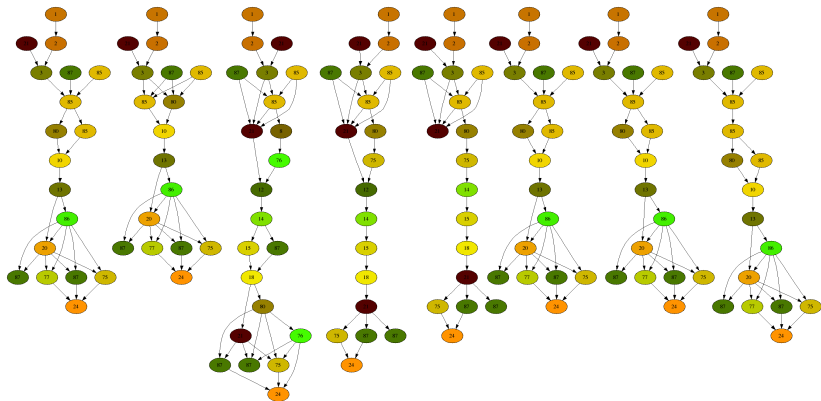
- ▶ ensemble de séquences
- ▶ ensemble de domaines

Sortie :

- ▶ une couverture pour chaque séquence
- ▶ ensemble des domaines utilisés
- ▶ vérifiant

poids des couvertures + poids des domaines est **minimal**

La même chose sur un dessin



Motivations

Formalisation

Modélisation

Un modèle linéaire

Variables

x_{ij}^r : couverture de s_r passe par l'arc $i \rightarrow j$

y_k : le domaine k est utilisé dans une couverture

Modèle Min

$$\sum_{r,i,j} c_{ij}^r x_{ij}^r + \sum_k d_k y_k$$

Subject to

$Ax = b$ (contraintes de flot)

$y_k \geq x_{ij}^r$ avec i ou j occurrence de k dans s_r

Les contraintes de flot

À la source

$$\sum_{s \rightarrow i} x_{si}^r = 1$$

Au puits

$$\sum_{i \rightarrow t} x_{it}^r = 1$$

Ailleurs

$$\sum_{j \rightarrow i} x_{ji}^r = \sum_{i \rightarrow j} x_{ij}^r$$

Quelques commentaires

- ▶ contraintes de flot \Rightarrow polytope entier (plus court chemin)
- ▶ source de complexité **partage des domaines**

Modèle

Min

$$\sum_{r,i,j} c_{ij}^r x_{ij}^l + \sum_k d_k y_k$$

Subject to

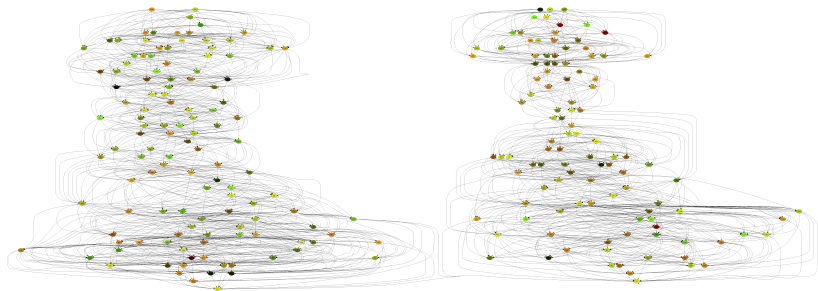
$Ax = b$ (contraintes de flot)

$y_k \geq x_{ij}^r$ avec i ou j occurrence de k dans s_r

Un résultat



Du travail, encore du travail



Quelques pistes

Algorithmique

- ▶ les variables y sont peu nombreuses (*tuning* CPLEX)
- ▶ les domaines en un seul exemplaire
- ▶ domaines souvent répétés **entre** séquences
- ▶ → *cost-splitting* + relaxation lagrangienne

Modélisation

- ▶ construire une représentation en contrôlant le compromis
 - ▶ précision
 - ▶ simplicité
- ▶ en jouant sur ce compromis (\approx taux de compression) :
 - ▶ hiérarchie de modèles

Merci de votre attention !

The screenshot shows the DomainOrganizer web application running in a Firefox browser. The browser's address bar shows the URL: `http://genoweb.univ-rennes1.fr/Serveur-GPO/outils_acces.php3?id_syndic=204`. The page header includes the logo for "Plateforme Bio-informatique GENOUEST" and a navigation menu with items: "La plateforme", "Outils", "Banques", "Séminaires", "Formations", and "Aide".

The main content area features the "DomainOrganizer" title and a descriptive paragraph: "Domain Organizer is a software package proposing a synthetic view of a set of DNA sequences by providing both a segmentation of them into domains and a classification on the basis of these domains. It aligns the sequences, finds the domains in the alignment and searches the distribution of each domain in sequences. After a classification step relatively to the presence or the absence of domains, the method results in a graphical view of a hierarchical clustering of the segmented sequences."

The interface is organized into several sections with input fields:

- User Parameter:** Includes an "Email (optional)" text input field.
- Sequences parameters:** Includes a "File containing sequences in fasta format" label and a "Browse..." button.
- Domain parameters:** Includes two input fields: "Minimal size of domains" (set to 20) and "Percentage of errors tolerated" (set to 25).
- Alignment parameters:** Includes a radio button for "Provide a custom alignment" with a "Browse..." button, and a radio button for "Use ClustalW" with two input fields: "Open Gap cost" (set to 20) and "Extension Gap cost" (set to 0.01).

A "Help" section on the right side of the page states: "A quick help appears here each time you enter a value."

The browser's taskbar at the bottom shows several open applications, including "pveber@hallux: ~/js...", "emacs@hallux", "pgfuseguide.pdf", "Gestionnaire de paq...", "Plateforme Bioinfor...", "Downloads", and "Inbox - Thunderbird".