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

AAATATGATCACATTTCATCCCTAAAAACACTATAT

TCAATAATATCCAAATCATATATTATGATTTTTCAC

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

  AAATATGATCACATTTCATCCCTAAAAACACTATAT

  TCAATAATATCCAAATCATATATTATGATTTTTCAC

  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

  AAATATGATCACATTTCATCCCTAAAAACACTATAT

  TCAATAATATCCAAATCATATATTATGATTTTTCAC

  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

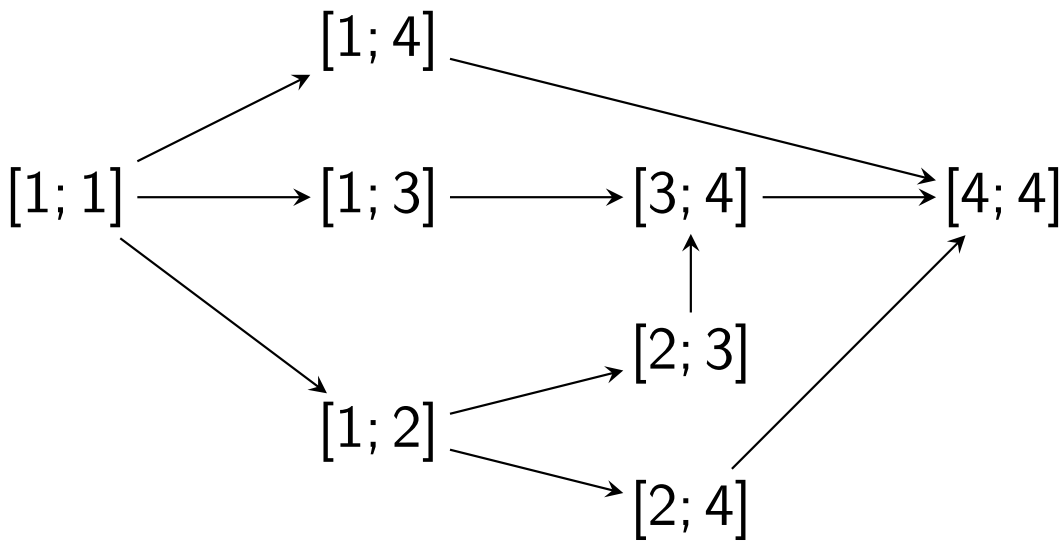
Modélisation

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 : $[i; j] \longrightarrow [k; l]$ si $|k - j| < t$



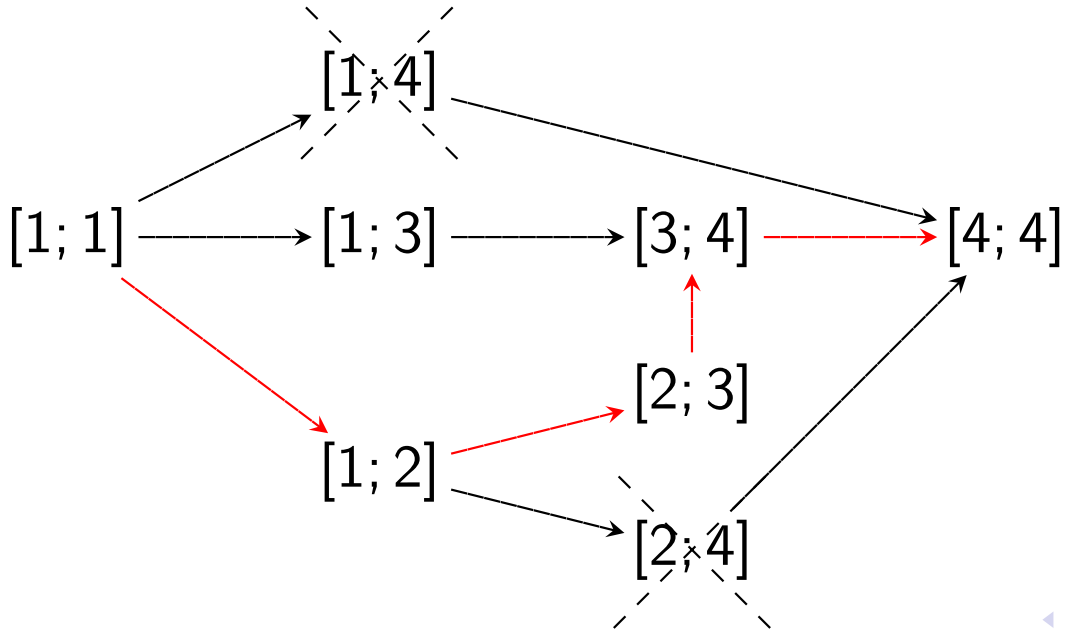
Graphe de couverture

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 :

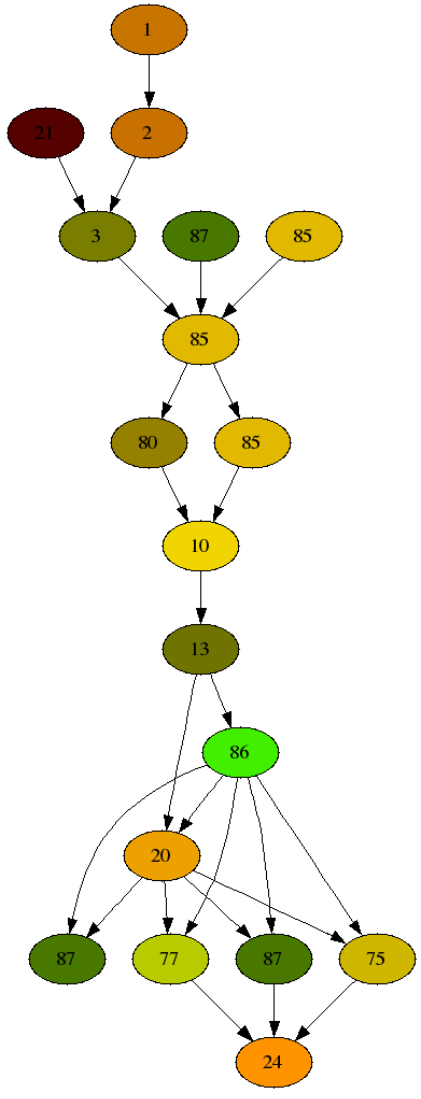
- dont les sommets ressemblent à des domaines dans D
- qui relient 2 intervalles proches sur la séquence

$$\text{poids d'une couverture} \equiv \text{poids du chemin.}$$

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 :

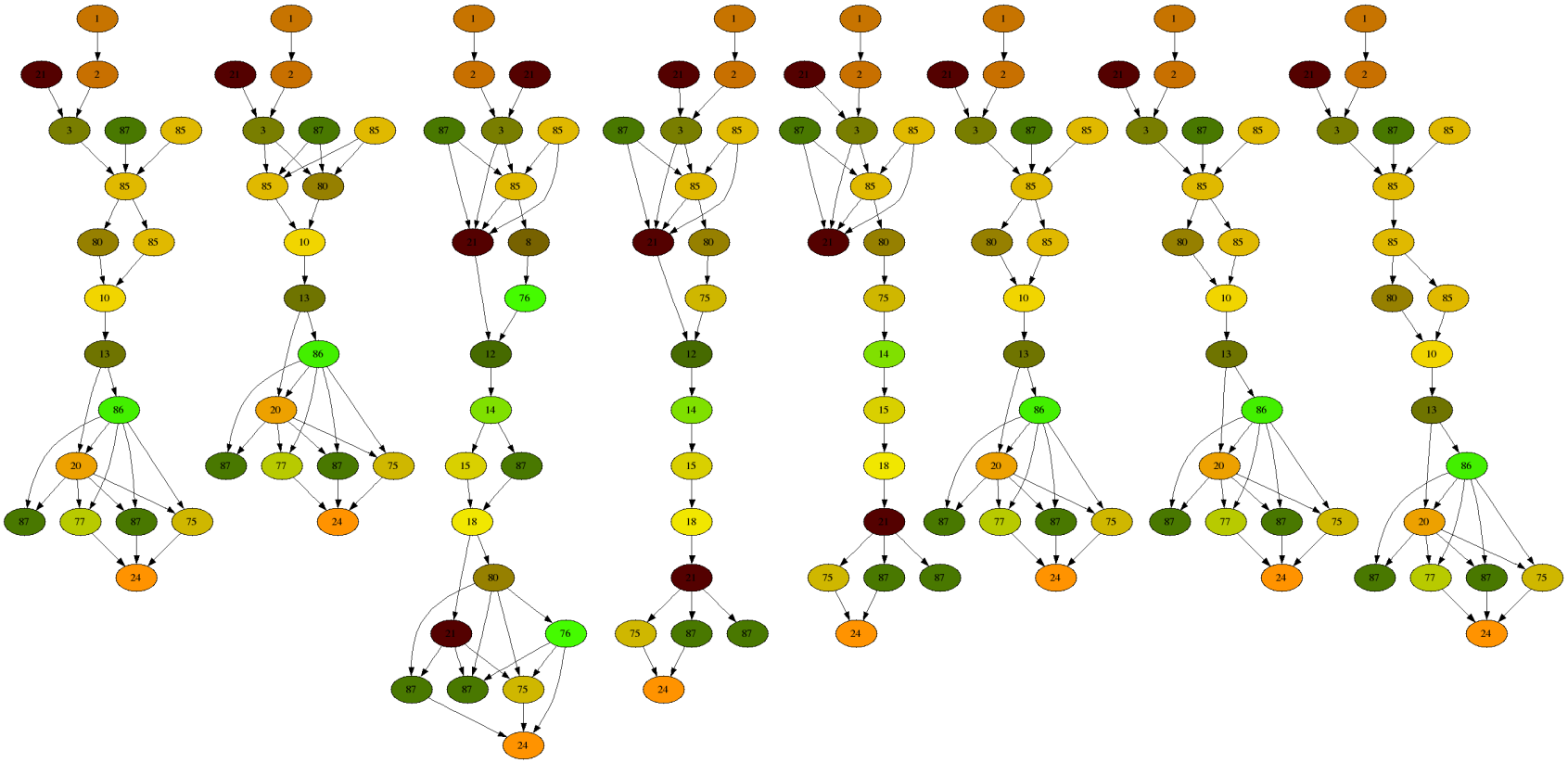
- ensemble de séquences
- ensemble de domaines

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 $i \longrightarrow j$

y_k : le domaine k est utilisé dans une couverture

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

Les contraintes de flot

À la source

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

Au puits

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

Ailleurs

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

Quelques commentaires

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

Modèle
Min

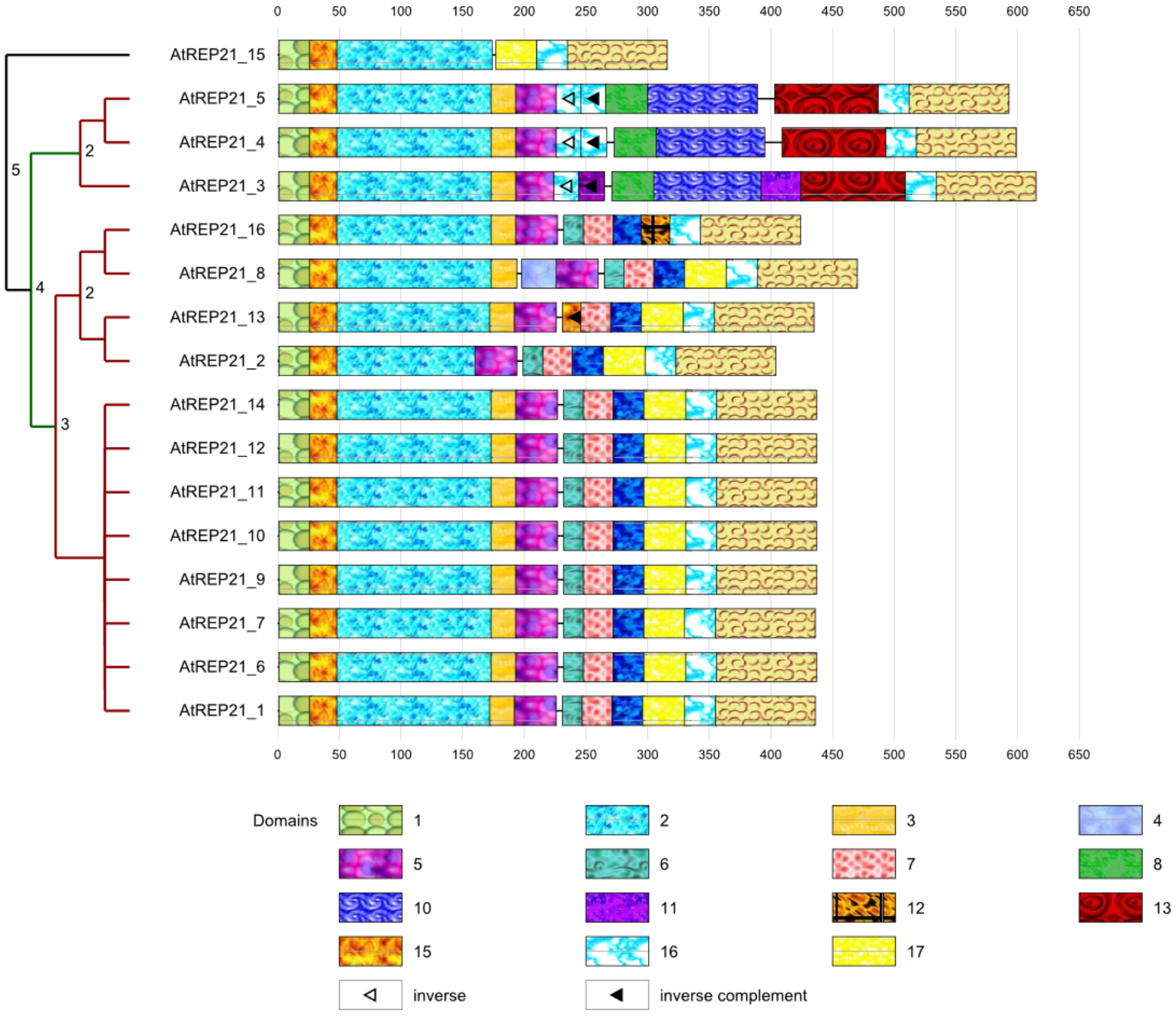
$$\sum_{r,i,j} c_{ij}^{r} x_{ij}^{l} + \sum_{k} d_k y_k$$

Subject to

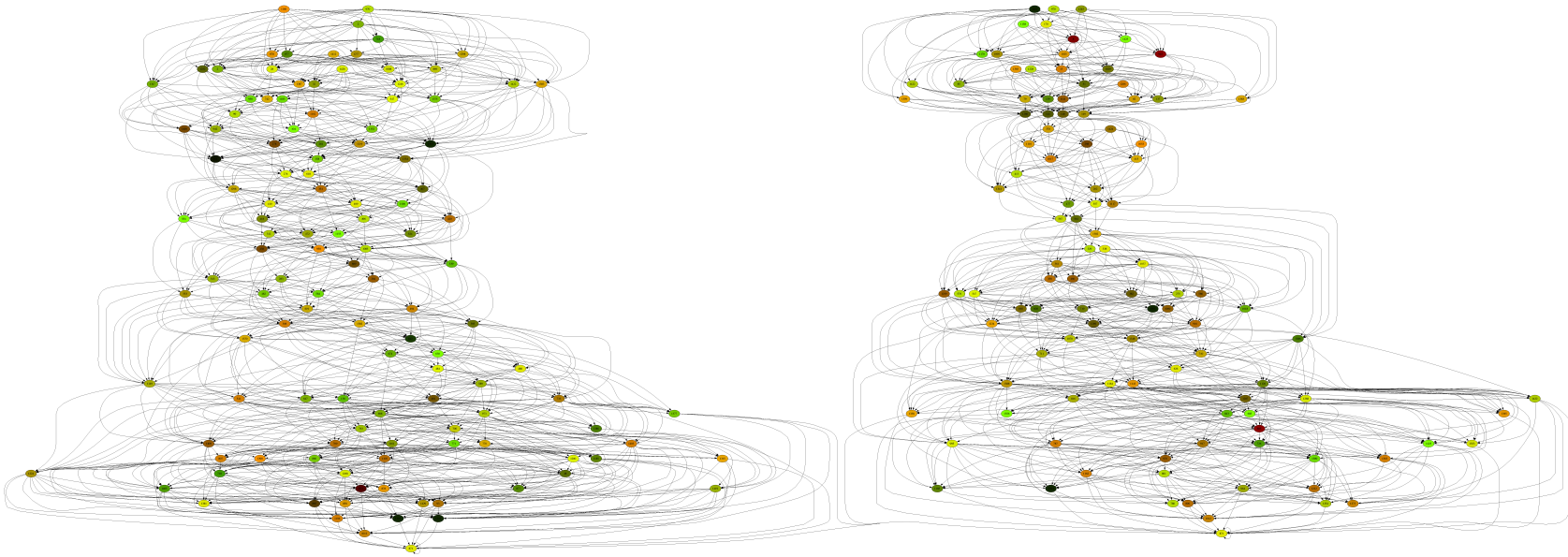
$Ax = b$ (contraintes de flot)

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

Un résultat


0 50 100 150 200 250 300 350 400 450 500 550 600 650
AtREP21_15
AtREP21_5
AtREP21_4
AtREP21_3
AtREP21_16
AtREP21_8
AtREP21_13
AtREP21_2
AtREP21_14
AtREP21_12
AtREP21_11
AtREP21_10
AtREP21_9
AtREP21_7
AtREP21_6
AtREP21_1
5
2
4
2
3
Domains
1
2
3
4
5
6
7
8
10
11
12
13
15
16
17
inverse
inverse complement

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
- $\longrightarrow$ *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 !