Développement d'un modèle d'alignements flexibles pour la reconnaissance de repliements des protéines

Guillaume Collet

INRA, unité Mathématique, Informatique et Génome

26 juin 2006

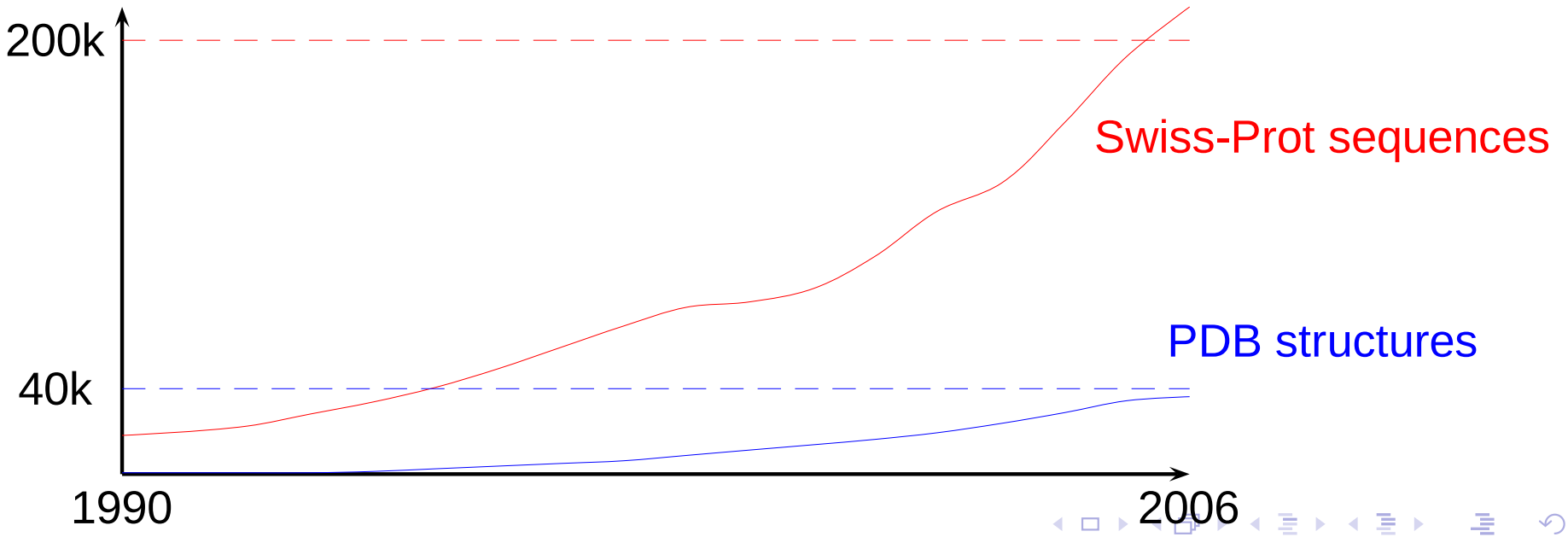
Contexte

Connaissances croissantes

- des données génomiques
- du nombre de séquences de protéines

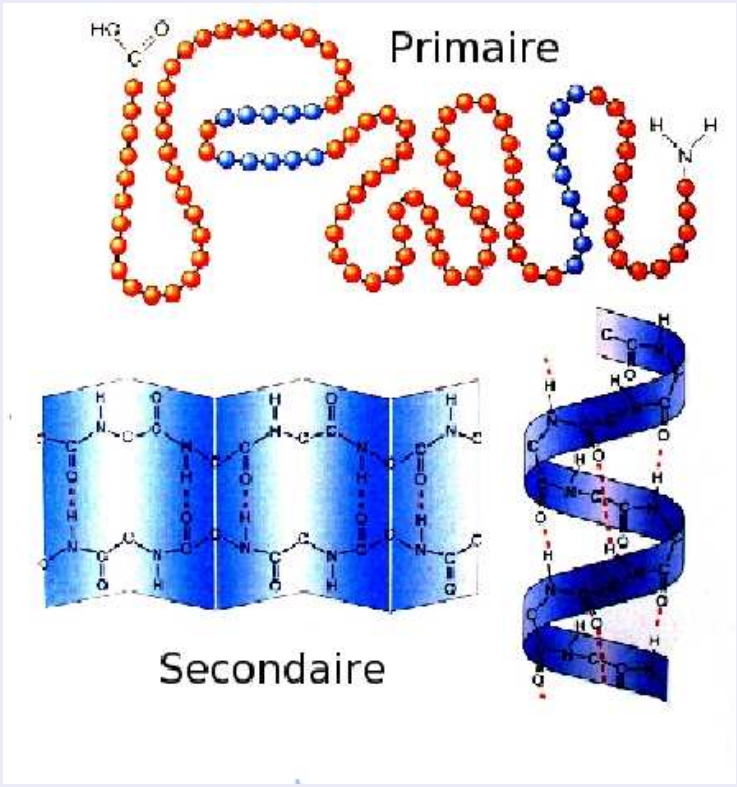
Néanmoins

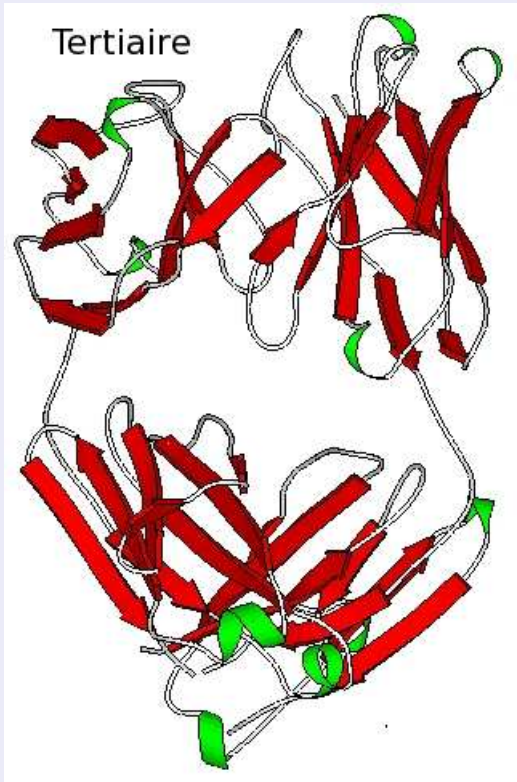
- Faible croissance du nombre de structures de protéines connues



Les protéines

Niveaux de description structurelle





Les méthodes de prédiction de structures de protéines

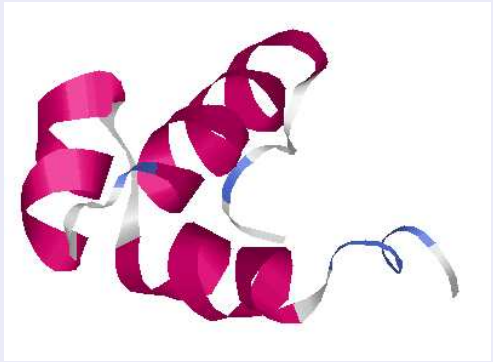
Méthodes utilisant une structure support

- Méthodes par homologie
  - Identité de séquence $>20\% =$ homologie entre les deux protéines
  - On en déduit donc la structure de la protéine recherchée
- Méthodes par reconnaissance de repliements
  - Identité de séquence $<20\% =$ on ne sait pas
  - Comparaison avec des structures connues

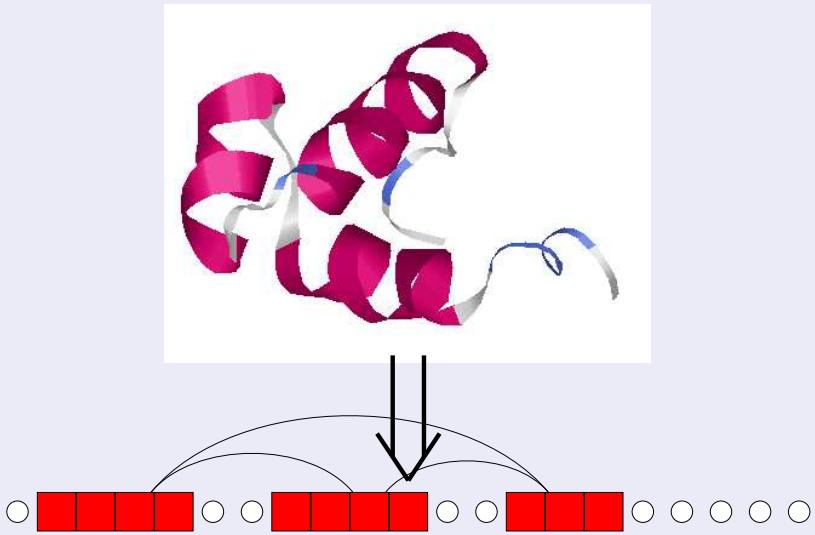
Méthodes par reconstruction

- Méthodes *ab initio*
  - Pas d'identité de séquence
  - Modélisation des atomes et des interactions physico-chimiques
- Méthodes *de novo*
  - Reconstruction à partir d'un ensemble de fragments de structures

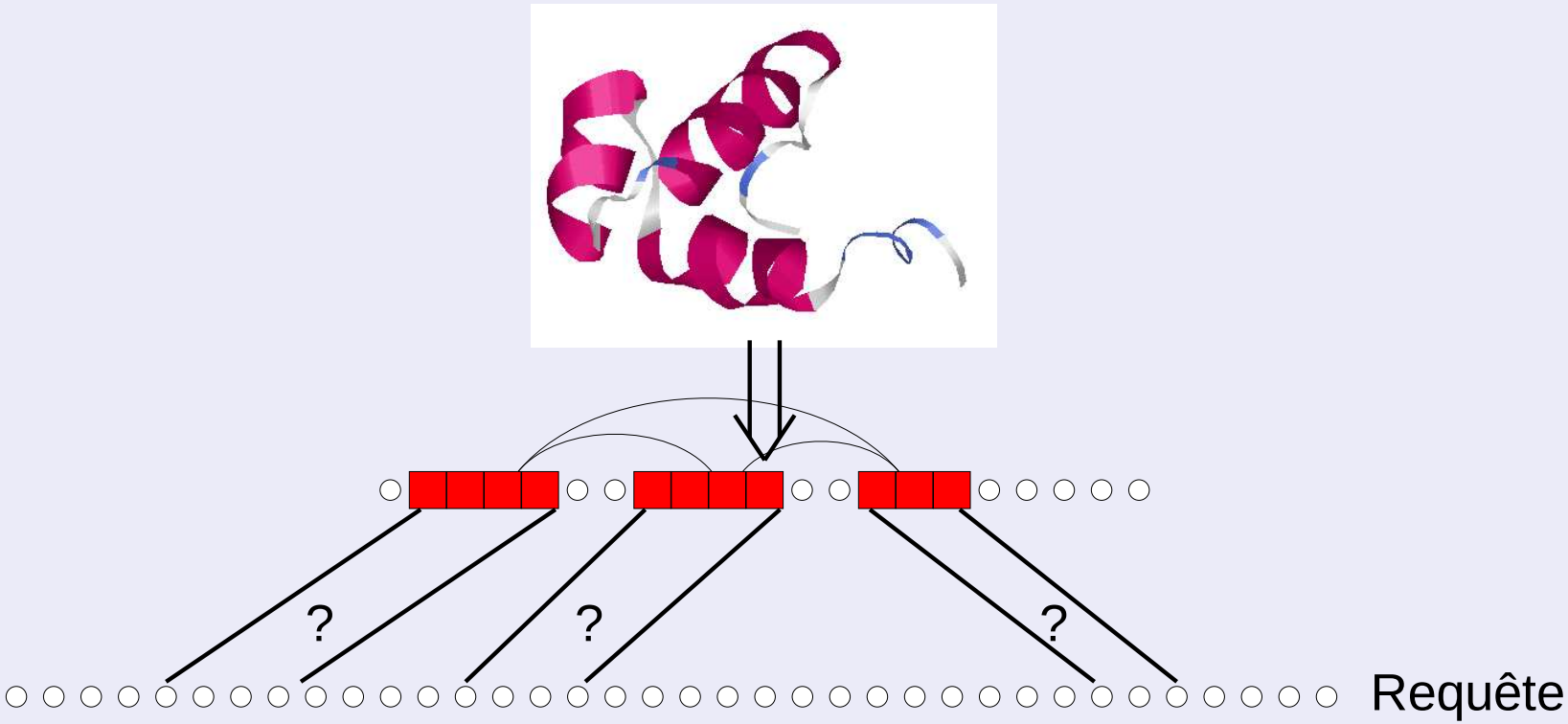
Principe de la reconnaissance de repliements de protéines



Principe de la reconnaissance de repliements de protéines



Principe de la reconnaissance de repliements de protéines



Une fonction de *fitness* permet de connaître le score de l'alignement

Principe de la reconnaissance de repliements de protéines

Les composantes

- Une banque de structures de protéines ($\Rightarrow$ des *cœurs*)
- Une fonction de score pour évaluer la qualité d'un alignement
- Une méthode de recherche du meilleur alignement
- Une méthode de normalisation des scores

Definition

Un cœur est constitué :

- De blocs = régions structurellement conservées
- D'arcs = acides aminés en contacts

Etat des lieux

Ce que nous savons

- Les structures sont plus conservées que les séquences au cours de l'évolution.
- Certaines structures sont très proches à quelques variations près.

Ce que nous avons

- Un logiciel d'alignement de structures sur des séquences protéiques : FROST
- Un système de prototypage des algorithmes d'alignement

Objectifs

Objectifs d'ordre biologique

- Détecter les homologues structuraux ayant un faible taux d'identité de séquence (< 20%)
- Etre capables d'omettre certains blocs s'ils diminuent la qualité de l'alignement.

Objectifs d'ordre informatique

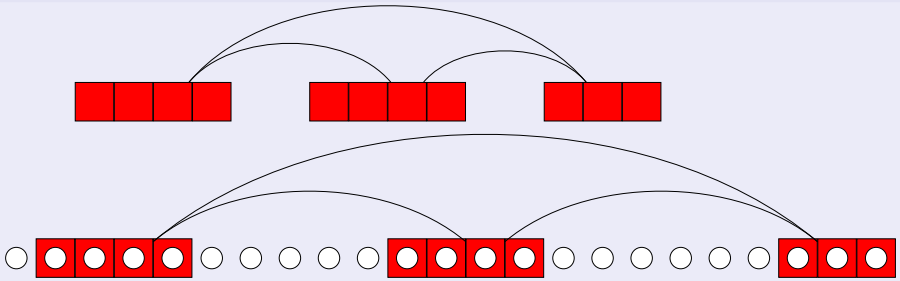
- Comprendre la complexité du problème pour trouver la modélisation la plus efficace en temps de calcul
- Résoudre le problème en un temps raisonnable.

Les difficultés rencontrés

- Plusieurs modélisations possibles
- Les problème sont de plus en plus grands
  - Pour une séquence de 100 AA sur un coeur de 50 AA (5 blocs) :
  - en global : 6 471 002 solutions possibles
  - en flexible : 9 361 459 solutions possibles
- Le temps de calcul n'est pas forcément lié à la taille du problème à résoudre (ex : la programmation dynamique)

Les alignements flexibles

Principe

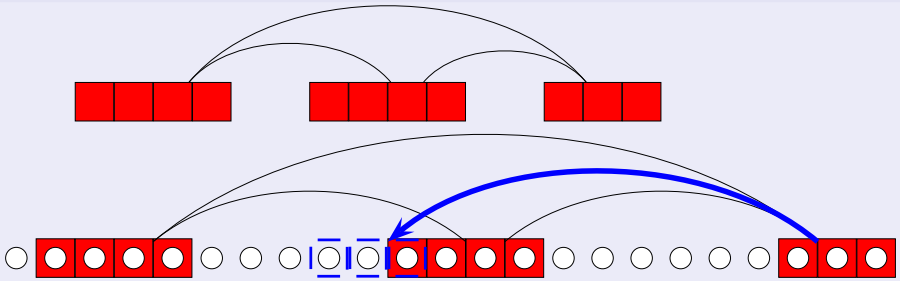


Modélisations

- Laisser les blocs "disparaître" $\Rightarrow$ modèle compact
- Les transformer en blocs fictifs $\Rightarrow$ modèle enrichi

Les alignements flexibles

Principe

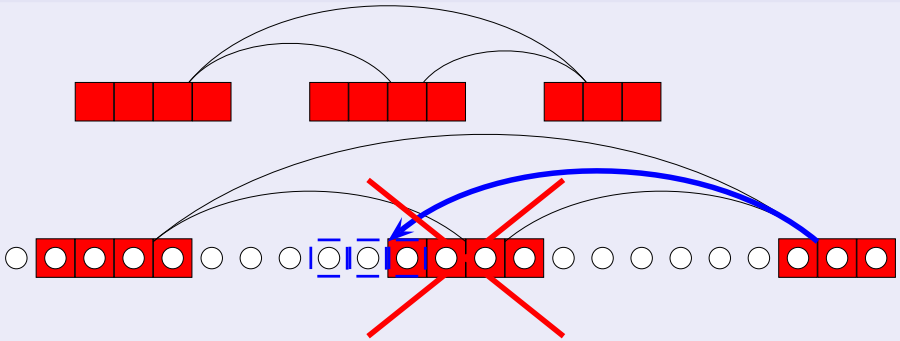


Modélisations

- Laisser les blocs "disparaître" $\Rightarrow$ modèle compact
- Les transformer en blocs fictifs $\Rightarrow$ modèle enrichi

Les alignements flexibles

Principe

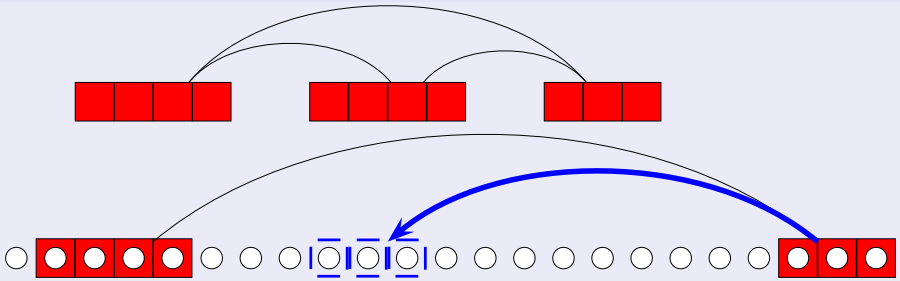


Modélisations

- Laisser les blocs "disparaître" $\Rightarrow$ modèle compact
- Les transformer en blocs fictifs $\Rightarrow$ modèle enrichi

Les alignements flexibles

Principe

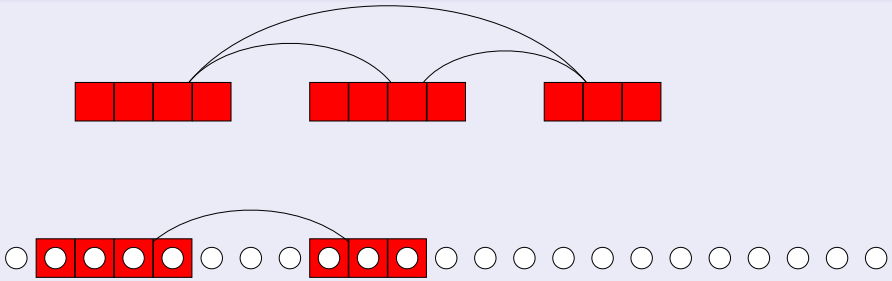


Modélisations

- Laisser les blocs "disparaître" $\Rightarrow$ modèle compact
- Les transformer en blocs fictifs $\Rightarrow$ modèle enrichi

Les alignements flexibles

Principe

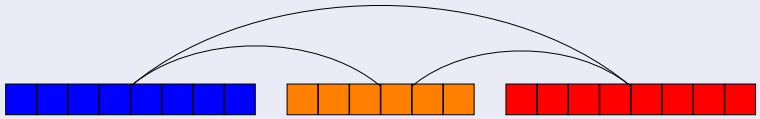


Modélisations

- Laisser les blocs "disparaître" $\Rightarrow$ modèle compact
- Les transformer en blocs fictifs $\Rightarrow$ modèle enrichi

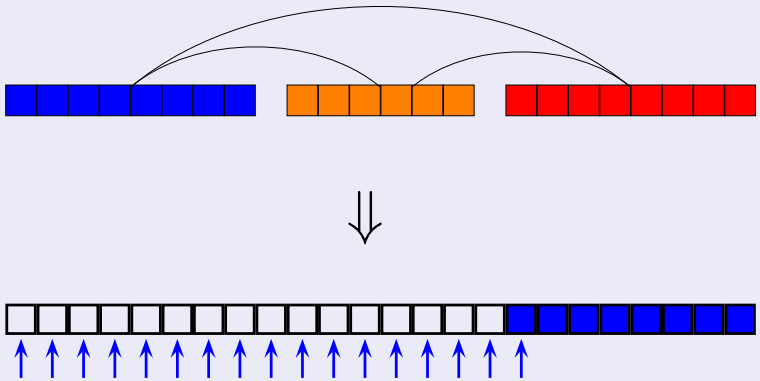
Modèle mathématique

Le graphe de flots



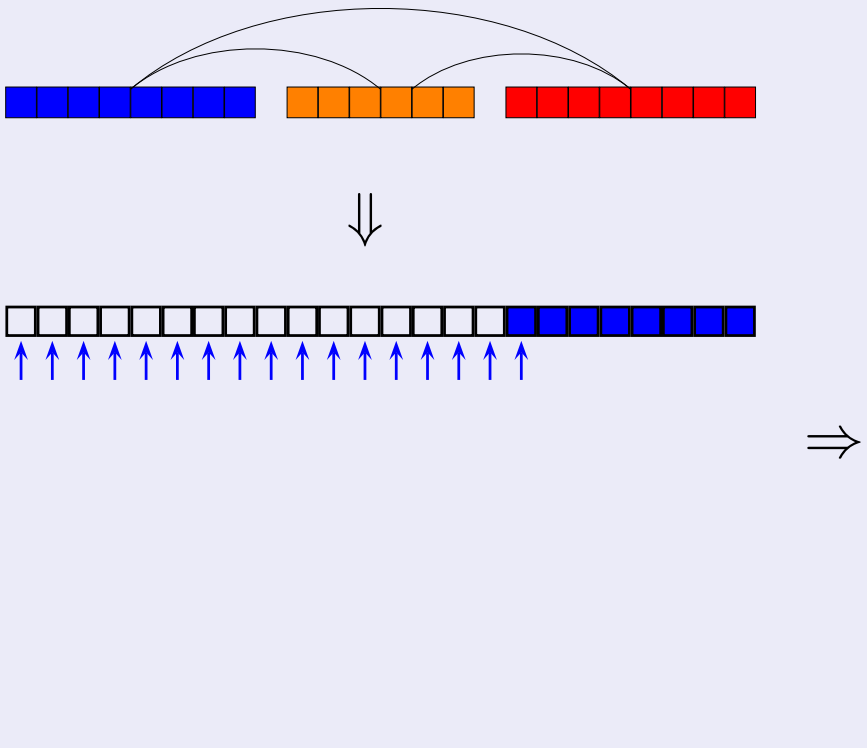
Modèle mathématique

Le graphe de flots



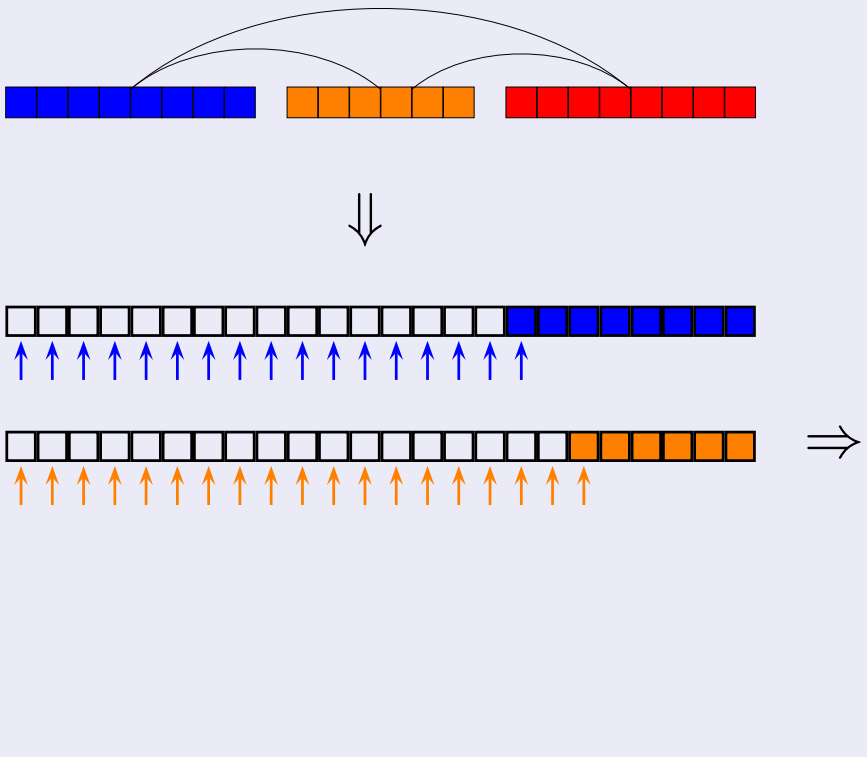
Modèle mathématique

Le graphe de flots



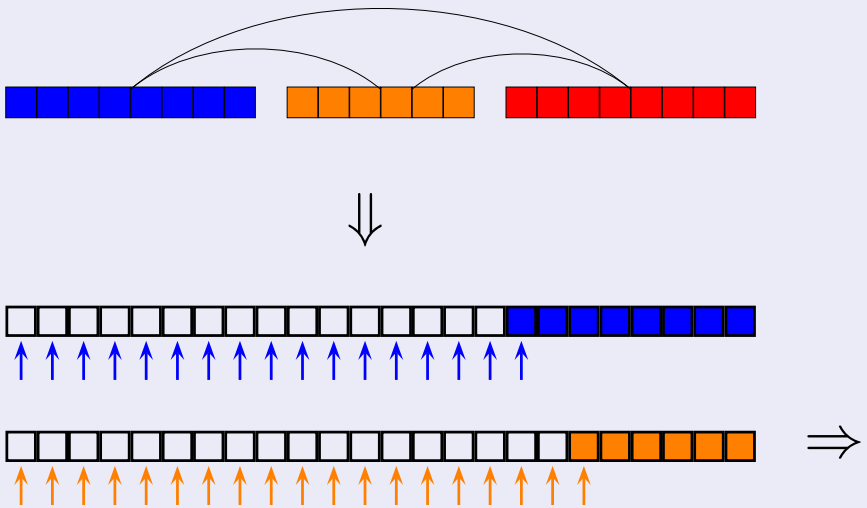
Modèle mathématique

Le graphe de flots



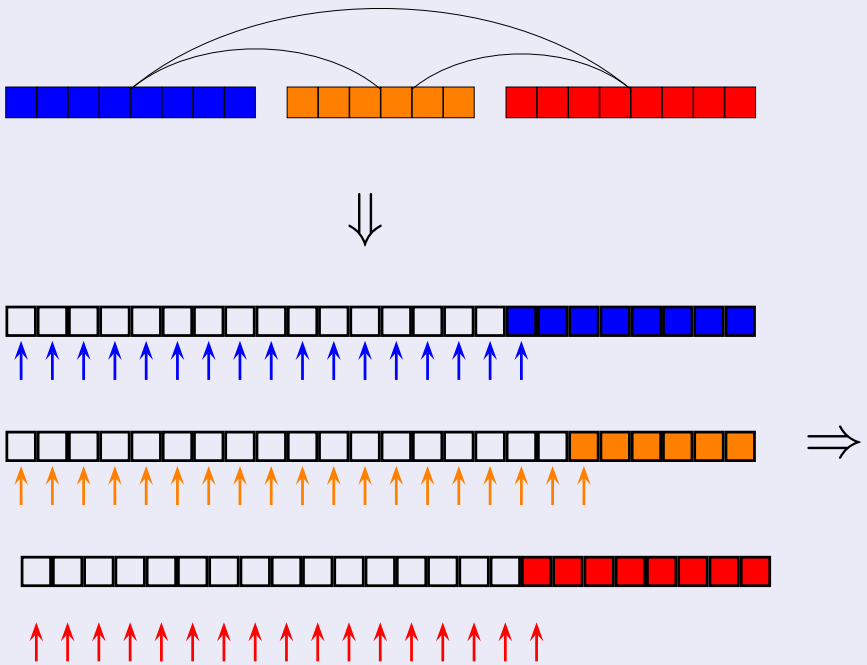
Modèle mathématique

Le graphe de flots



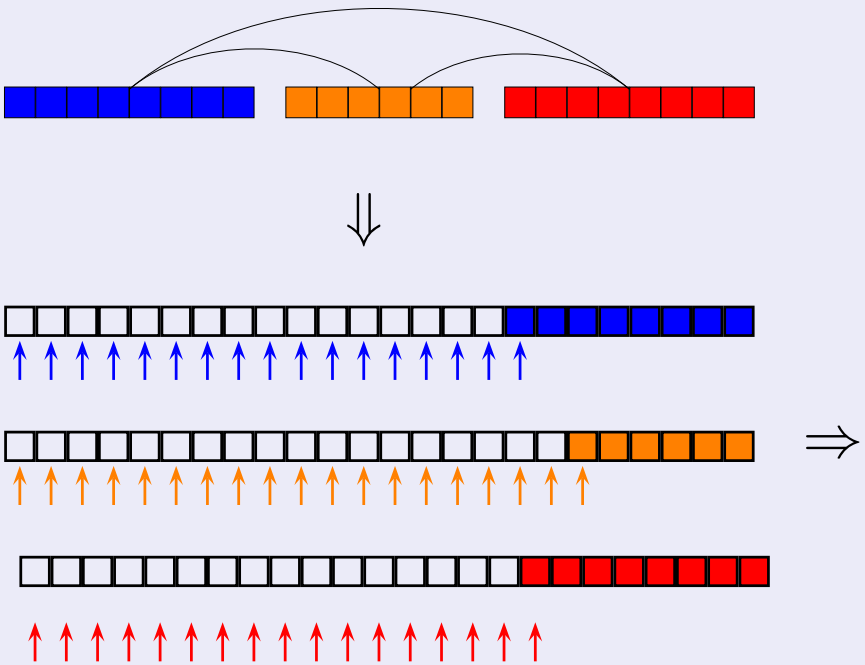
Modèle mathématique

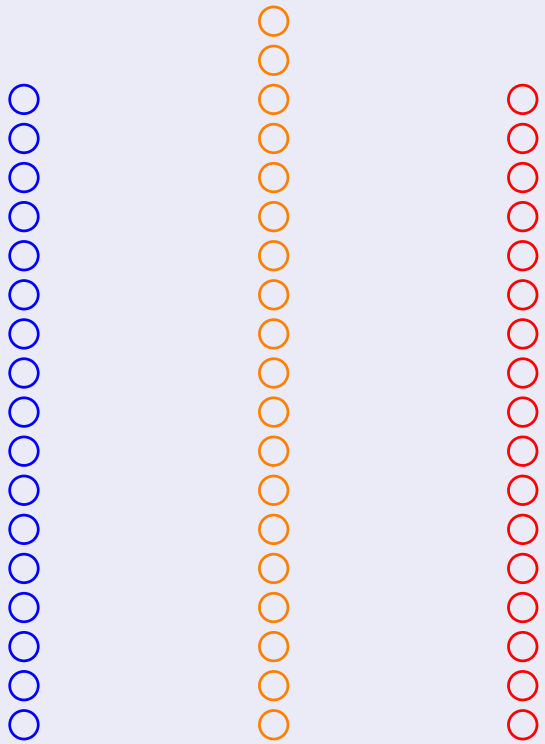
Le graphe de flots



Modèle mathématique

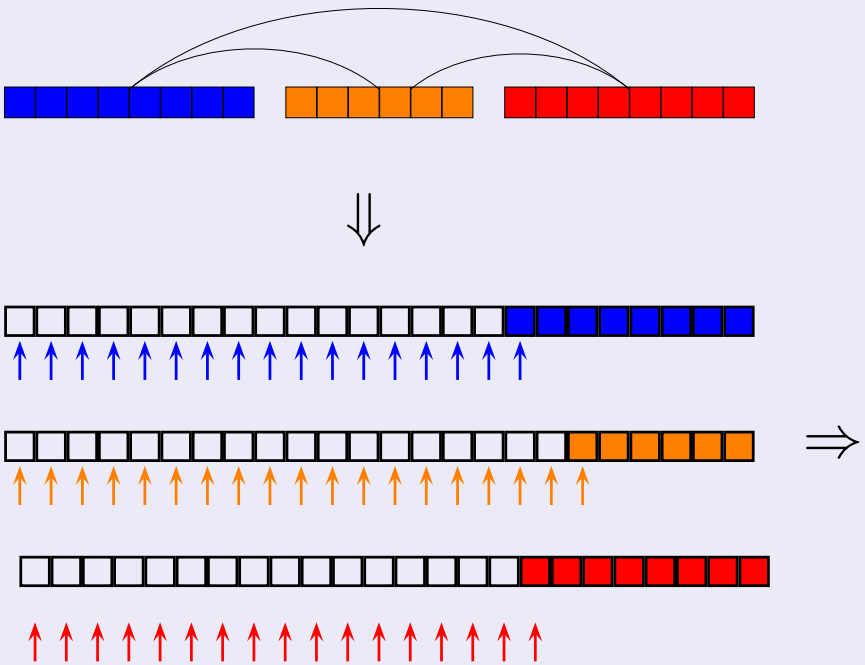
Le graphe de flots

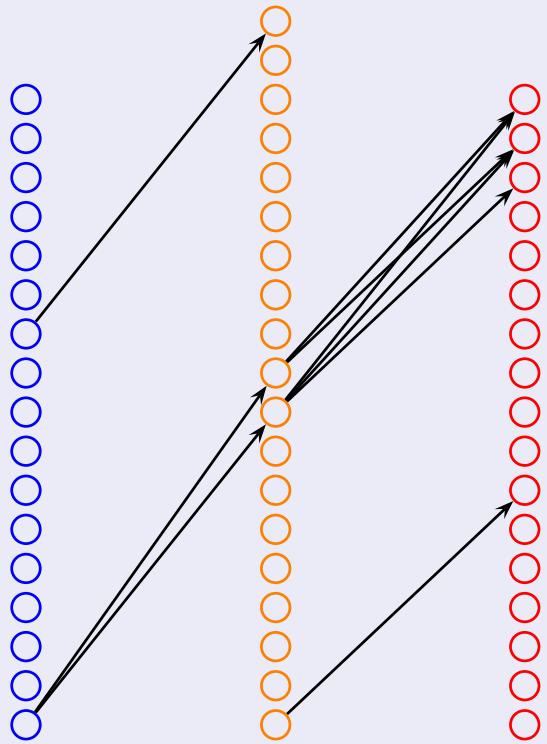




Modèle mathématique

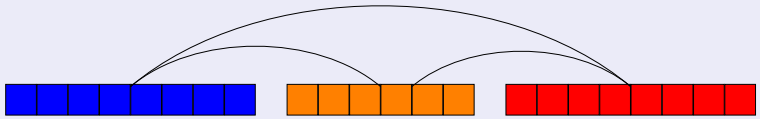
Le graphe de flots

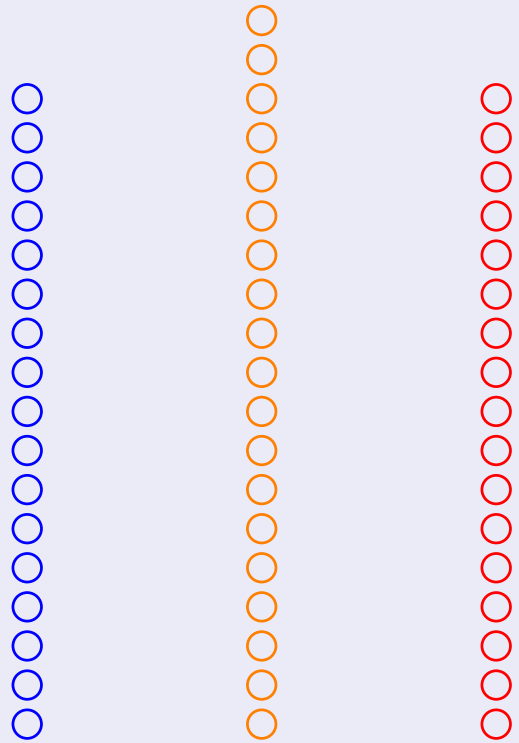




Modèle mathématique

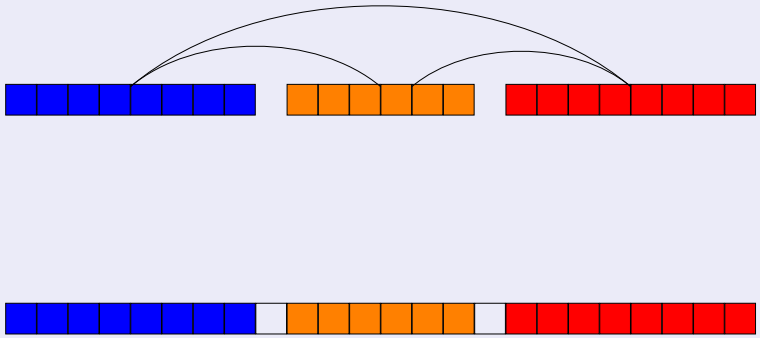
Exemples d'alignements

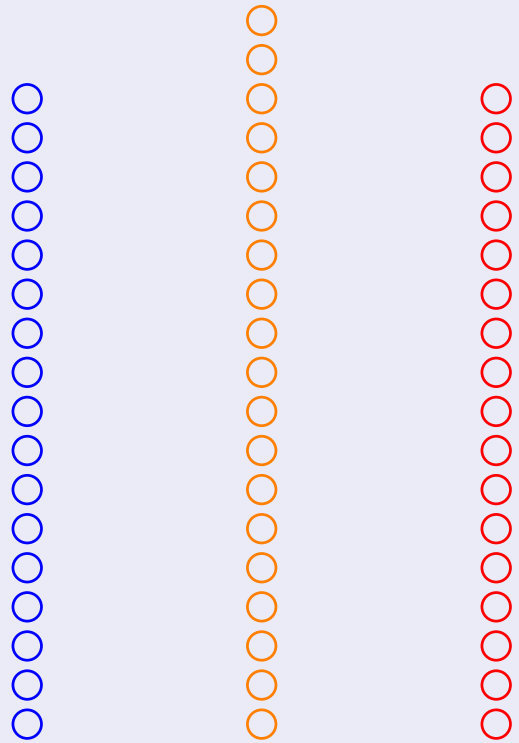




Modèle mathématique

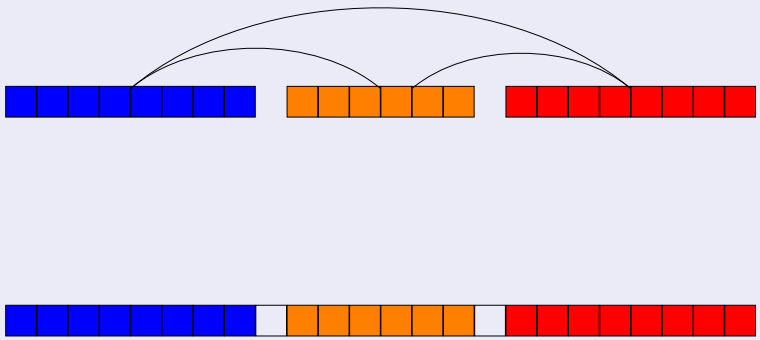
Exemples d'alignements

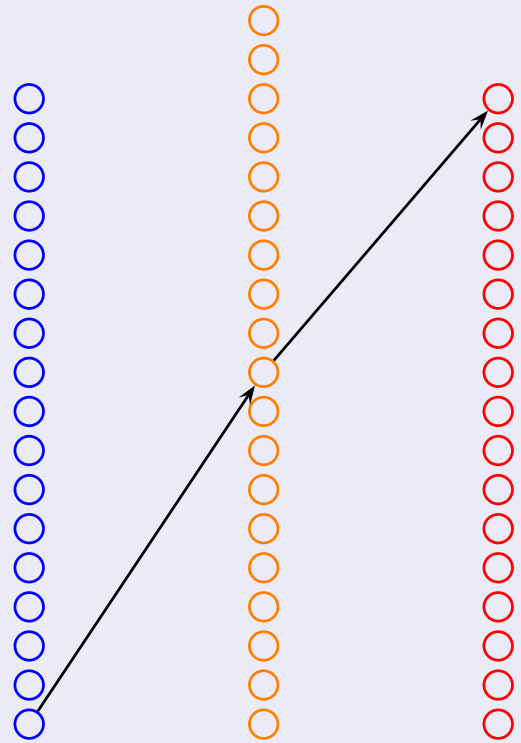




Modèle mathématique

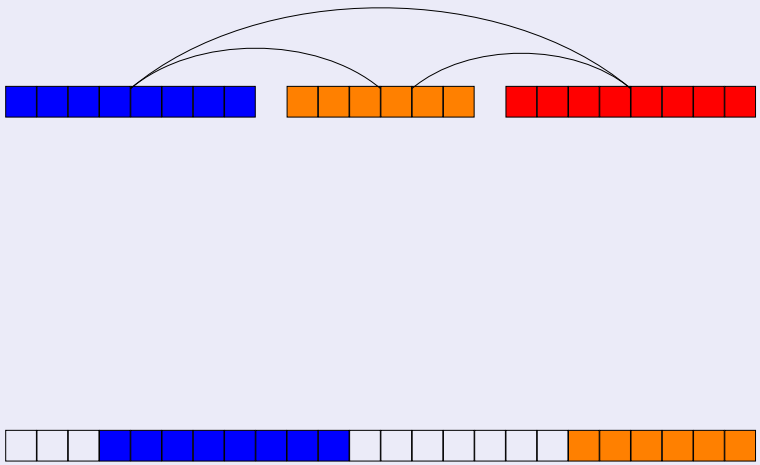
Exemples d'alignements

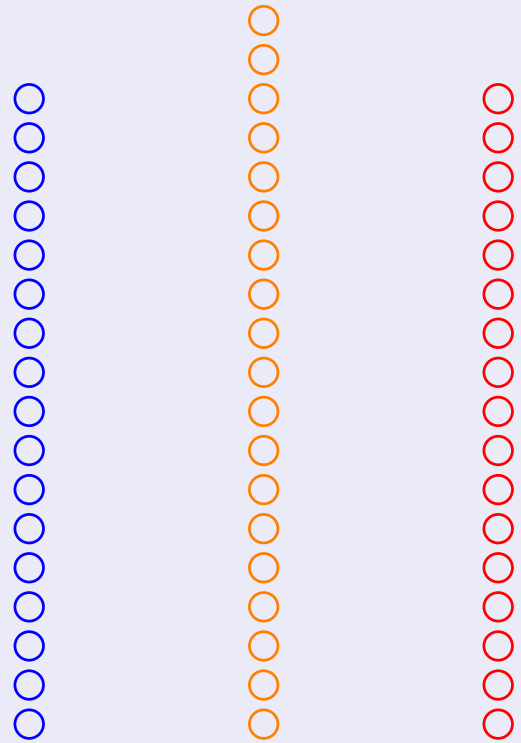




Modèle mathématique

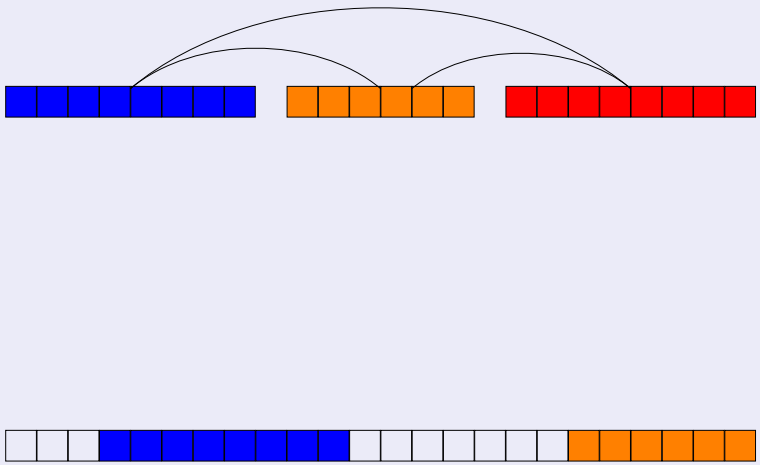
Exemples d'alignements

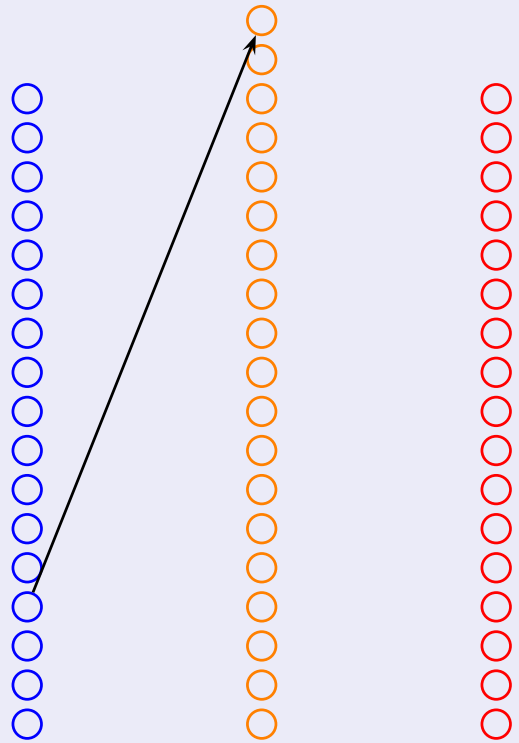




Modèle mathématique

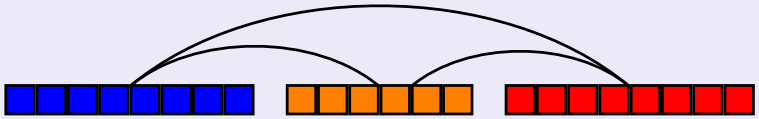
Exemples d'alignements

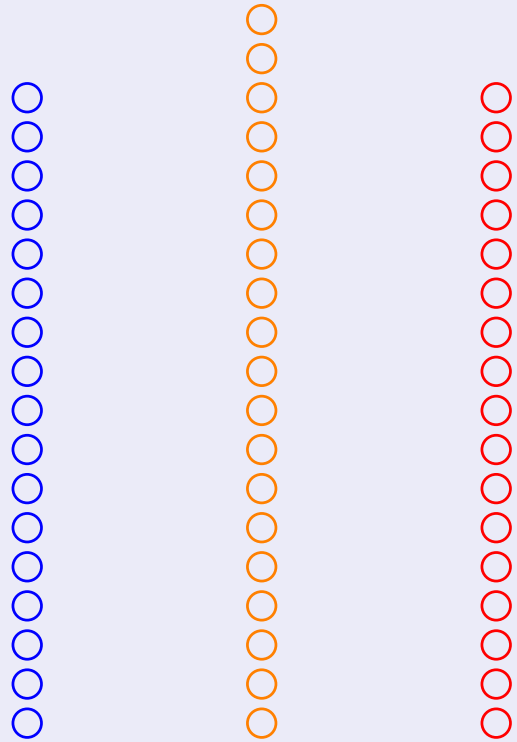




Modèle mathématique

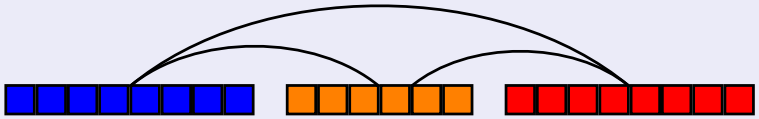
Les variables du modèle





Modèle mathématique

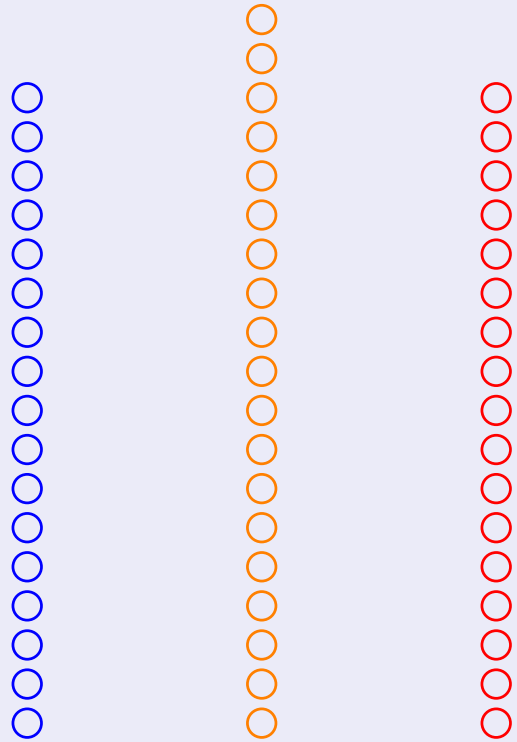
Les variables du modèle



Les variables $Y_{ij} \Rightarrow$ les nœuds du graphes

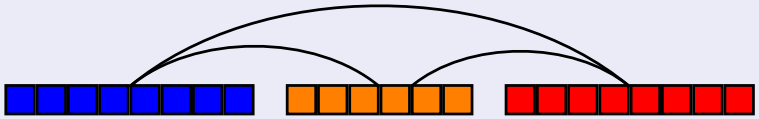
Contrainte :

Une seule ou aucune position pour chaque bloc



Modèle mathématique

Les variables du modèle



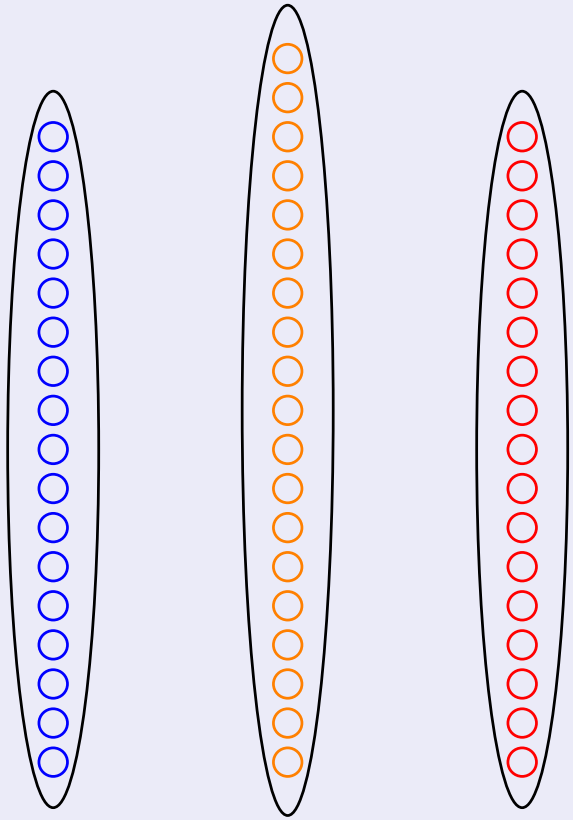
Les variables $Y_{ij} \Rightarrow$ les nœuds du graphes

Contrainte :

Une seule ou aucune position pour chaque bloc

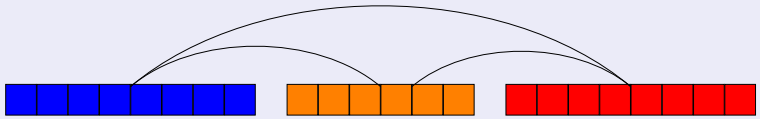
$$\Downarrow$$

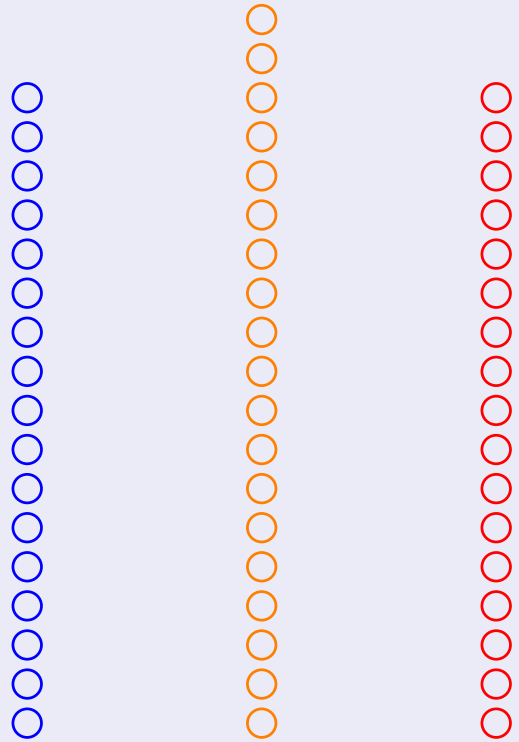
$$\sum_{j=1}^{\tilde{n}_i} Y_{ij} \leq 1$$



Modèle mathématique

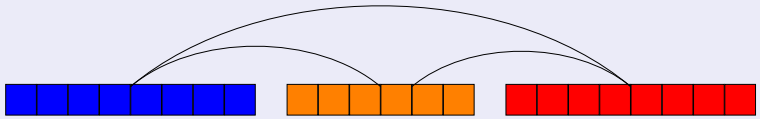
Les variables du modèle





Modèle mathématique

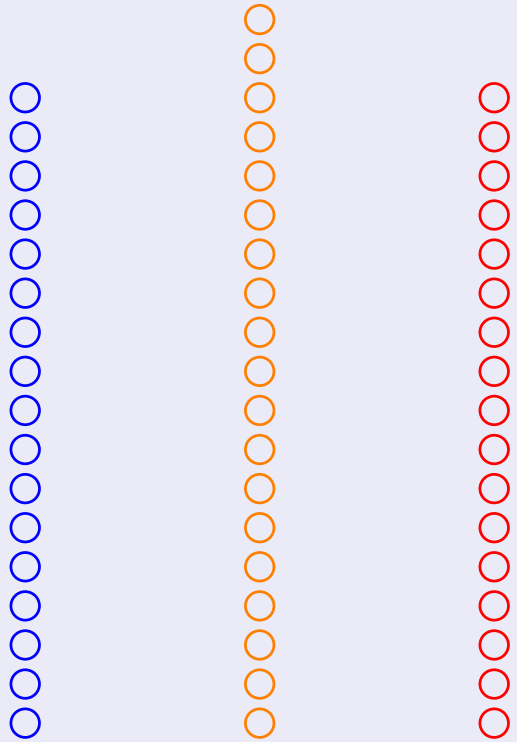
Les variables du modèle



Contrainte :

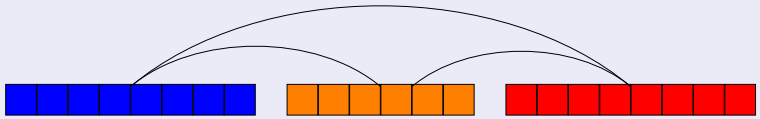
l'ordre des blocs est conservé

pas de chevauchement



Modèle mathématique

Les variables du modèle



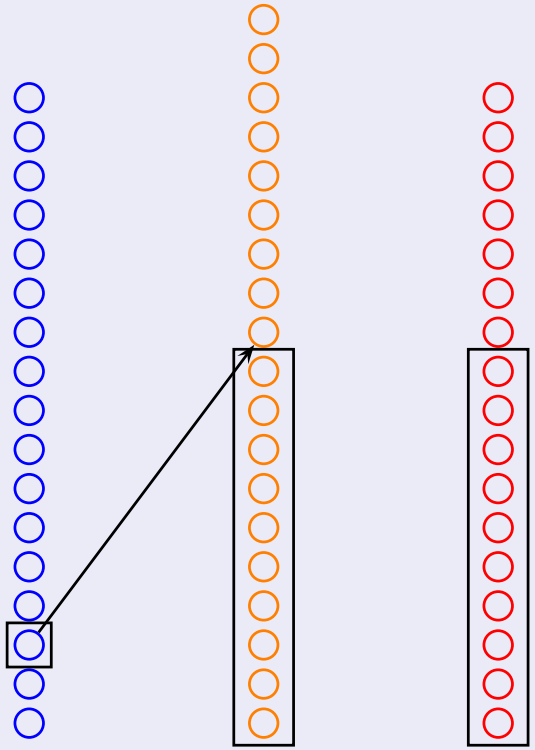
Contrainte :

l'ordre des blocs est conservé

pas de chevauchement

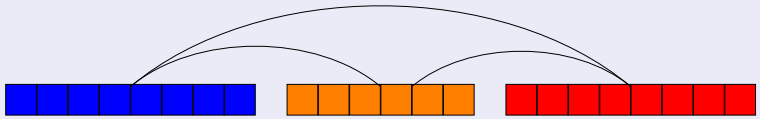
$$\Downarrow$$

$$Y_{ij} + \sum_{l=1}^{j+L_i-1} Y_{kl} \leq 1$$

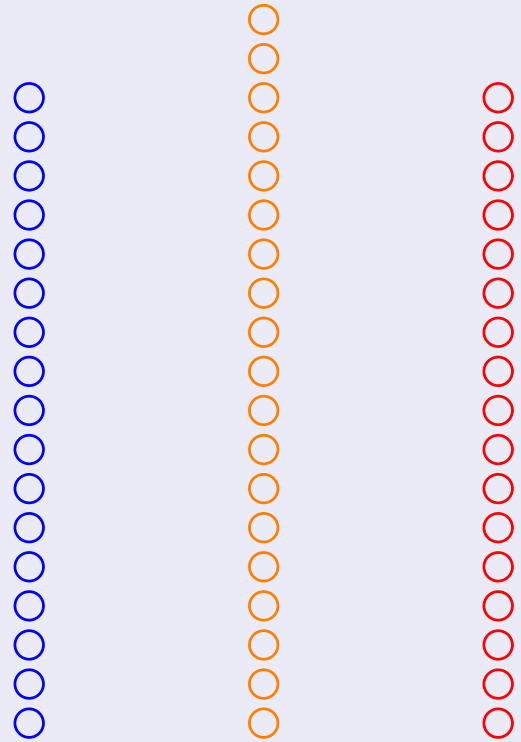


Modèle mathématique

Les variables du modèle

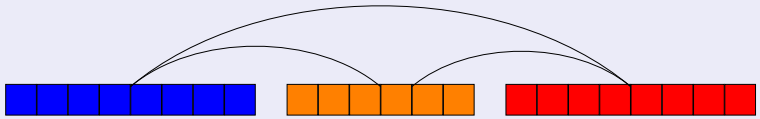


1 interaction = 1 variable Z_{ijkl}



Modèle mathématique

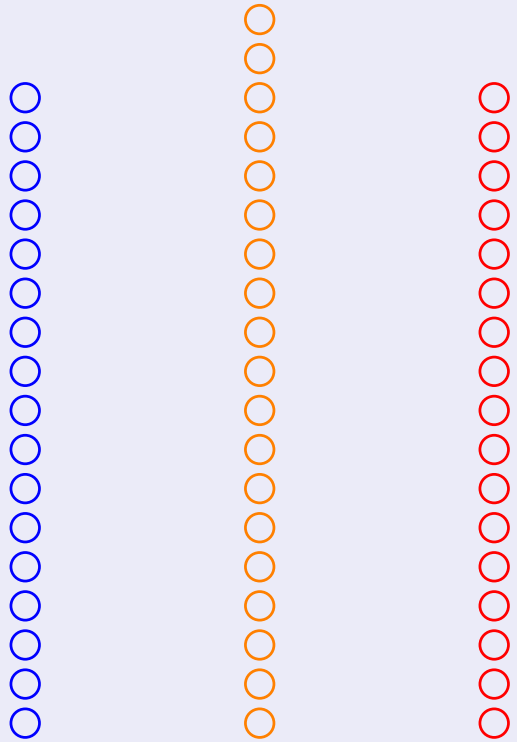
Les variables du modèle



1 interaction = 1 variable Z_{ijkl}

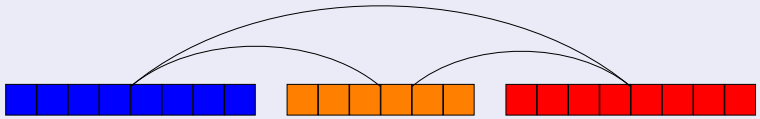
Contrainte :

Une seule interaction sortante pour un nœud



Modèle mathématique

Les variables du modèle



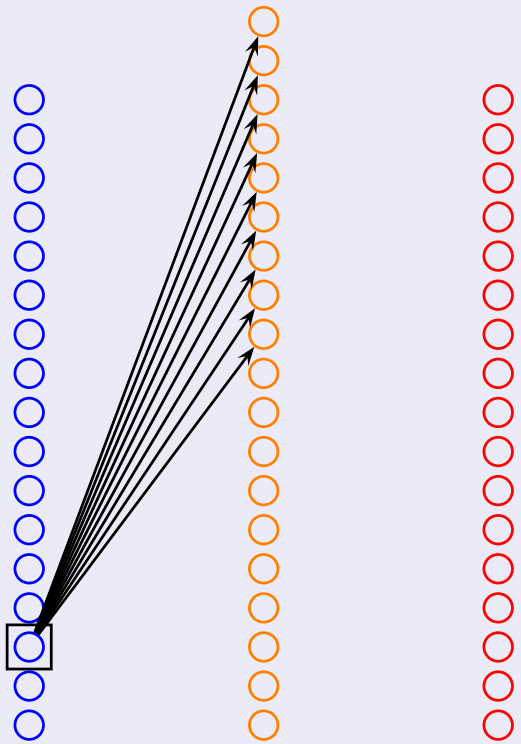
1 interaction = 1 variable Z_{ijkl}

Contrainte :

Une seule interaction sortante pour un nœud

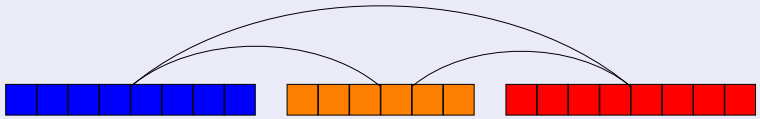
$$\Downarrow$$

$$Y_{ij} \geq \sum_{l=j+L_i}^{\tilde{n}_k} Z_{ijkl}$$



Modèle mathématique

Les variables du modèle



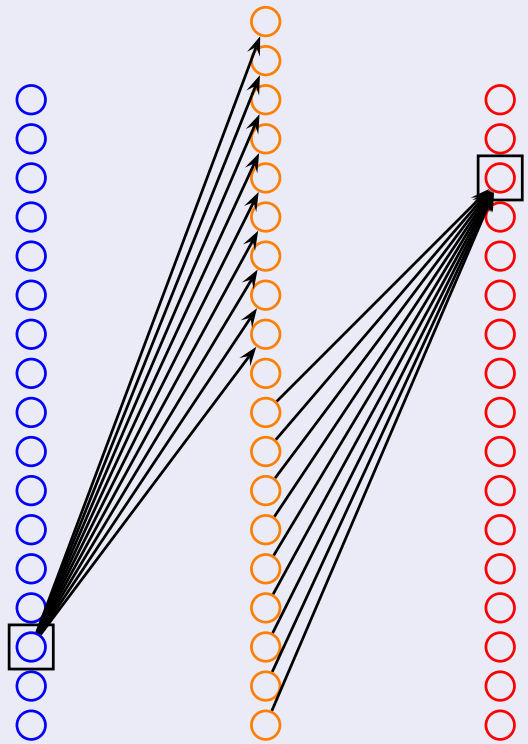
1 interaction = 1 variable Z_{ijkl}

Contrainte :

Une seule interaction entrante dans un nœud

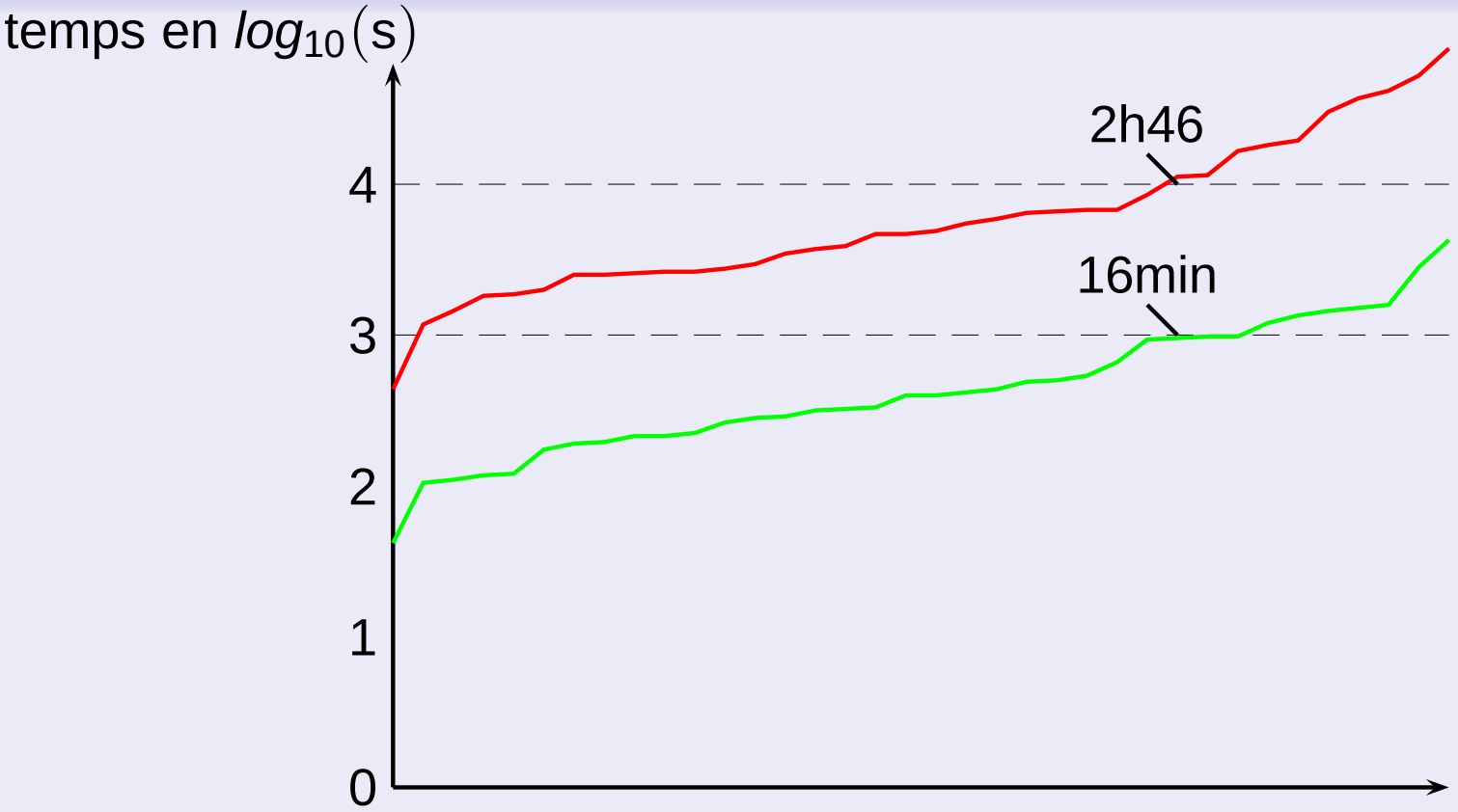
$$\Downarrow$$

$$Y_{ij} \geq \sum_{l=j+L_i}^{\tilde{n}_k} Z_{ijkl}$$



Comparaison en temps de calculs

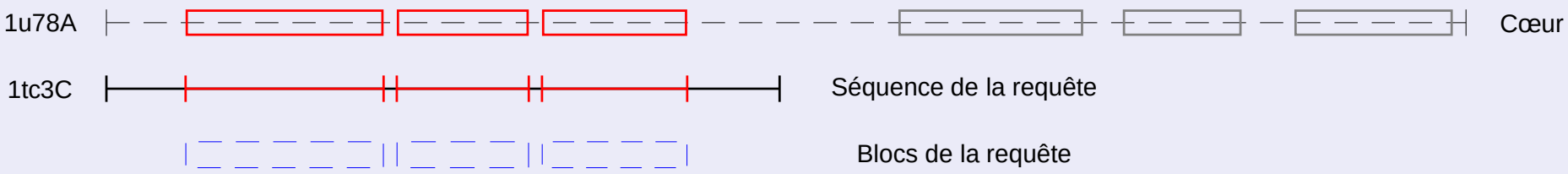
Temps de calculs des modèles flexibles



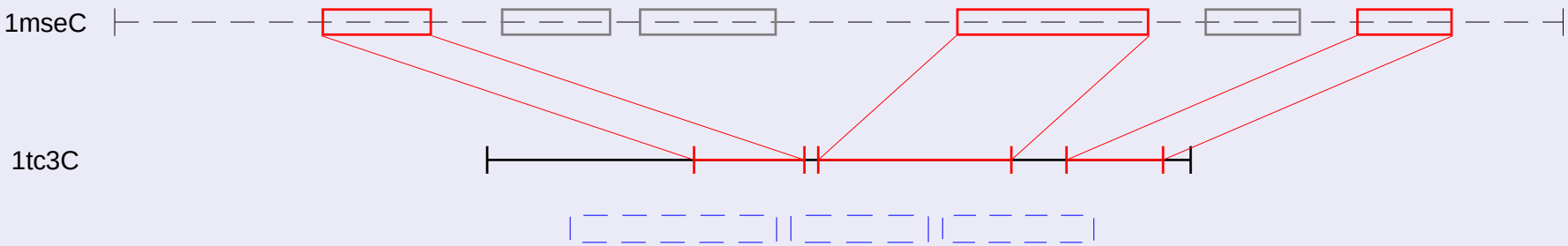
- En rouge les temps du modèle compact.
- En vert les temps du modèle enrichi.

Résultats avec alignement flexible

Des blocs sont omis

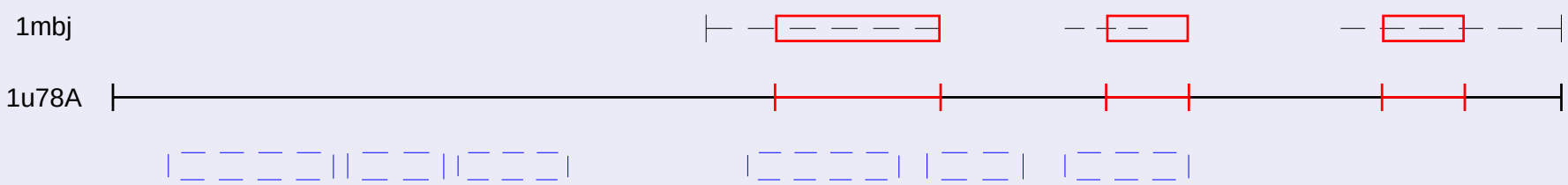


Mais dispersion de la séquence

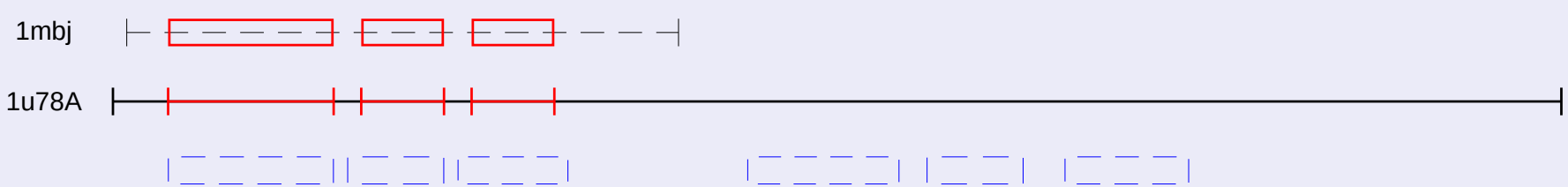


Résultats avec alignement semi-global

Alignement sans poids de gaps



Alignement avec poids de gaps



Résultats sur les scores normalisés

Scores plus discriminants

Séquence : 1gvdA	Cœur : 1ityA	Cœur : 1bl0A
Alignement global	26.94	0.17
Alignement semi-global	35.53	0.17
Alignement flexible	67.56	2.81

TAB.: Scores normalisés des alignements de la séquence 1gvdA sur les cœurs 1ityA et 1bl0A.

La compétition CASP

- Compétition internationale de prédiction de structures des protéines
- Prédiction en "aveugle"
- Deux compétitions parallèles :
  - Prédictions complètement automatiques (serveur de prédiction)
  - Prédictions semi-automatiques

Conclusion et perspectives

Conclusion

- Les alignements flexibles sont intégrés
- Les scores sont plus discriminants

Perspectives

- Un algorithme dédié plus efficace
- Intégrer les poids de gaps
- Compétition CASP

Merci de votre attention

Exemple de cœur