Alignements multiples locaux et partiels de séquences protéiques à partir de paires de fragments alignables

Olivier Baldellon

Alignements multiples locaux et partiels de séquences protéiques à partir de paires de fragments alignables

Olivier Baldellon,
École Normale Supérieure de Cachan, Antenne de Bretagne
Équipe Symbiose
Encadrant: François Coste

25 septembre 2008

Table des matières

Introduction

Les protéines sont des composants indispensables au vivant. En effet, parmi leurs nombreux rôles, elles participent au transport de l'oxygène avec l'hémoglobine des globules rouges ou à la défense immunitaire avec les anticorps. Une protéine est une chaîne de plusieurs centaines d'acides aminés dont la fonction est définie par sa structure 3D et la position des acides aminés le long de cette structure. Un des sujets majeurs de la bioinformatique est ce qu'on appelle l'alignement de protéines, c'est à dire l'identification des parties homologues entre plusieurs protéines. L'intérêt principal des alignements est que deux protéines possédant de nombreuses parties communes ont souvent des fonctions similaires. Ainsi les alignements permettent de découvrir des familles de protéines ou de prédire la fonction d'une protéine nouvellement découverte.

Les premiers alignements de la littérature étaient représentés par des matrices. En 2000, l'article [2] définit de manière exacte le problème d'alignement matriciel en s'appuyant sur une formalisation à base de graphe. Le problème ainsi posé s'appelle le problème GMT (Generalized Maximum Trace). L'idée de représenter la solution sous forme de graphe va permettre la considération de nouveaux types alignements plus riches que ceux proposés dans le problème GMT qui se limitait à certains graphes bien particuliers.

Ainsi des articles comme ceux de l'algorithme d'ABA [4, 3] vont considérer des graphes où les croisements d'arêtes seront autorisés. Les solutions obtenues ne sont plus forcement représentables sous forme de matrice, mais peuvent contenir des informations biologiques plus intéressantes.

Cependant ces articles travaillant avec des alignements à base de graphe n'ont jamais clairement défini le type de solution recherché. Seul des algorithmes se basant sur une heuristique étaient proposés.

Le but de ce stage était donc de pallier à ce manque et de chercher à définir rigoureusement à la manière de l'article [2] le type de solution recherchée. Nous avons donc introduit une notion de localité permettant de définir cette nouvelle famille de graphe. Intuitivement ces graphes correspondent aux alignements qui sont localement matriciels. Plus concrètement, si l'on regarde le graphe d'alignement "de près", les alignements sont les mêmes que dans [2] mais si l'on prends du "recul" alors on autorise les duplications (deux positions d'une même séquence alignées ensemble) et les croisements.

Nous avons ensuite cherché à formuler le problème en programmation linéaire puis de le résoudre à l'aide du logiciel Cplex.

1 État de l'art

1.1 Définitions

Avant de commencer à rentrer dans le détail, définissons plus précisément les termes que nous aurons l'occasion d'employer.

On appelle *séquence* protéique, ou plus simplement séquence un mot sur l'alphabet des acides aminés. On appelle *position* un couple (S, i) où S est une séquence et i un entier naturel strictement inférieur à la taille de S. Intuitivement la position (S, i) correspond à la i^{eme} lettre de S. Par exemple si on considère la séquence $S = ABBA$, S contient deux lettres, A et B, mais quatre positions, la première $(S, 0)$ et la quatrième $(S, 3)$ associées à la lettre A, la deuxième $(S, 1)$ et la troisième $(S, 2)$ associées à la lettre B. On appelle *fragment* un ensemble de positions contiguës d'une même séquence. Par abus de langage on dira de la position (S, i) qu'elle appartient à S, de même on dira qu'un fragment est inclus dans une séquence.

1.2 Problème de la trace maximale généralisée (GMT problem)

Alignements matriciels et traces

Les alignements matriciels ont été définis rigoureusement dans l'article [2] de la manière suivante : «Soit $S = \{S_1, ...S_k\}$ un ensemble de k mots sur un alphabet Σ et soit $\hat{\Sigma}$ l'ensemble $\Sigma \cup \{'-'\}$. Un alignement matriciel de S est un ensemble $\hat{S} = \{\hat{S}_1, ..., \hat{S}_k\}$ vérifiant deux propriétés. Premièrement, tous les $\hat{S}_i$ ont la même longueur. Deuxièmement, S_i et $\hat{S}_i$ sont identiques si l'on ne tient pas compte des symboles $'-'$.» On parle d'alignement matriciel car si on note n la taille de $\hat{S}_i$ alors on peut identifier $\hat{S}$ à une matrice de n colonnes et de k lignes dont le contenu de la case située à la colonne c et à la ligne l est le l^{ieme} caractère de $\hat{S}_c$.

Si deux positions p et q sont dans la même colonne, on dit qu'elles sont alignées, on dit de même que le couple (p, q) est réalisé.

On considère alors un graphe non-orienté k-parties $G = (E, V)$ défini de la manière suivante : l'ensemble V des sommets est formé par l'ensemble des positions des séquences de S, l'ensemble E des arêtes est constitué d'un sous-ensemble de $V \times V$ choisi préalablement par l'utilisateur. On appelle G *graphe complet d'alignement* [1]

On nomme *trace* le sous-ensemble de E constitué des couples de positions (p, q) réalisés.

[1] Dans l'article initial, le terme employé est graphe d'alignement, cependant nous utiliserons une autre terminologie pour éviter de futures ambiguïtés.

Graphes et chemins mixtes

L'article [2] introduit la notion de graphe complet étendu[2]. Avant de définir précisément cette notion nous devons déjà définir la notion de graphe mixte. Un *graphe mixte* est un triplet $G = (V, E, A)$ où V est un ensemble de sommets, E un ensemble d'arêtes (non orientées) sur V et A un ensemble d'arcs (qui peuvent être vus comme des arêtes orientées).

On appelle *graphe complet étendu* le graphe d'alignement complet auquel on ajoute les arcs de la forme $((S, i), (S, i + 1))$ avec S une séquence et i un entier inférieur $|S| - 1$ où $|S|$ est la taille de S.

Un chemin sur un graphe mixte est une suite de la forme $v_1, e_1, v_2, e_2, ..., v_k$, où pour tout i, v_i est un sommet et e_i une arête ou un arc vérifiant $e_i = (v_1, v_{i+1})$.

On appelle *chemin mixte* un chemin contenant au moins une arête et au moins un arc.

Énoncé du problème

On définit un bloc comme un sous ensemble de E vérifiant la propriété suivante : «Si une arête d'un bloc est réalisée, alors toutes les arêtes du bloc sont réalisées». De plus à chaque bloc b on associe un poids ω_b. On définit le poids d'une union de blocs distincts deux à deux, comme la somme des poids des blocs.

Problème de la trace maximale généralisée (GMT) 1
Étant donné un graphe complet étendu, et une partition de ce graphe en bloc. Trouver l'union de blocs de poids maximal ne contenant pas de cycle mixte.

1.3 Relation d'ordre

L'article [1] donne une caratérisation des traces à partir de relation d'ordre permettant une nouvelle formulation au problème GMT.

On définit la relation d'ordre strict $<$ sur deux positions d'une même séquence en posant $(S, i) < (S, j)$ si et seulement si $i < j$. On peut alors définir une relation sur des composantes connexes de notre graphe en posant $C_1 \preceq C_2$ si et seulement si il existe deux positions d'une même séquence $p \in C_1$ et $q \in C_2$ tel que $p < q$.

Propriété 1.1
Soit G un graphe, les deux propositions suivantes sont équivalentes :
- *G ne possède pas de cycle mixte,*
- *la fermeture transitive de la relation $\preceq$ est une relation d'ordre strict sur G.*

[2] L'article original parlait de graphe d'alignement étendu.

Le lecteur pourra trouver une démonstration de cette propriété dans [1]. On peut donc reformuler le problème de la trace maximale généralisée :

Problème de la trace maximale généralisée (GMT) 2
Étant donné un graphe complet étendu G, et une partition de ce graphe en bloc. Trouver l'union de bloc de poids maximal tel que la fermeture transitive de la relation $\preceq$ soit une relation d'ordre strict sur G.

1.4 ABA

À partir du moment où l'on représente les alignements avec des graphes, on a accès à une représentation beaucoup plus expressive. Les croisements par exemple ne peuvent être représentés sous forme de matrice. Il faut garder à l'esprit cependant que tous les graphes ne représentent pas forcément des alignements pertinents.

Décrivons succintement le principe d'ABA.

- On aligne chaque couple de séquences avec Proda, un algorithme d'alignement matriciel.
- On combine tous les alignements obtenus pour obtenir un alignement multiple.
- On supprime tous les cycles mixtes de taille inférieure à une taille donnée par des corrections successives (ici la taille des cycle mixtes est le nombre d'arêtes contenues dans le cycle).

En fait, les croisements dans des graphes d'alignements sont caractérisés par des cycles mixtes. En supprimant les petits cycles mixtes, ABA fait en sorte que lorsque l'on regarde le graphe «localement» il n'y ait pas de croisement.

Nous essaierons dans la suite de ce rapport de donner un sens plus pertinent à la notion de localité.

2 Formalisation des alignements de fragments de protéines

2.1 Définitions et hypothèses

Définitions

Soit $\mathcal{S}$ un ensemble de séquences. Soit $G_{\mathcal{S}}$ le *graphe complet non-orienté* associé à $\mathcal{S}$ dont les sommets sont les positions des séquences de $\mathcal{S}$. On dit que A est un alignement de la paire de fragments $\{f_1, f_2\}$ si et seulement si A est un sous-graphe de $G_{\mathcal{S}}$ dont les arêtes sont dans $f_1 \times f_2 \cup f_2 \times f_1$. On appelle *appariement local* tout couple de la forme $(\{f_1, f_2\}, A)$ où $\{f_1, f_2\}$ est une paire de fragments et où A est un alignement de la paire $\{f_1, f_2\}$.

On appelle *arc intra-fragment* tout triplet de la forme (p_1, p_2, f) où p_1 et p_2 sont deux positions successives (c'est à dire $p_1 = (S, i) \Rightarrow p_2 = (S, i+1)$) de f.

Hypothèses

On suppose que l'on dispose d'une famille de séquences $\mathcal{S}$, et d'un ensemble $\mathcal{L}$ d'appariements locaux.

Par la suite on dira que deux fragments sont alignables si et seulement si il existe un appariement local A de $\{f_1, f_2\}$ tel que $(\{f1, f2\}, A) \in \mathcal{L}$. On dira aussi que deux positions p et q sont alignables si et seulement il existe un appariement local L=$(\{f_p, f_q\}, A)$ de $\mathcal{L}$ tel que $(p, q) \in A$ .On notera alors $[p, q]_L$. Ainsi de même que des algorithmes comme ABA partaient d'alignements de séquences deux à deux, nous partons d'alignements locaux de séquences (i.e. alignements de fragments) deux à deux. Nous chercherons ensuite le meilleur alignement respectant certaines contraintes que nous verrons plus tard que l'on peut construire à partir de ces alignements locaux.

2.2 Différentes familles d'alignements

Dans la suite on appellera graphe d'appariement, tout graphe non-orienté dont les sommets sont les positions des séquences de $\mathcal{S}$. On appellera alignement simple, ou plus simplement alignement, un graphe d'appariement transitif.

Si (p, q) est une arête d'un graphe d'appariement G , on dira que p et q sont alignés dans G. On dira aussi que l'arête (p, q) est réalisée dans G. Si $L = (\{f_1, f_2\}, A)$ est un appariement local dont toutes les arêtes de A sont réalisées dans un graphe d'appariement G, on dira que L est réalisé dans G.

Appariement fragment-compatible

On dira qu'un graphe d'appariement est $\mathcal{L}$-fragment-compatible ou plus simplement fragment-compatible, si pour toute arêtes (p, q) du graphe, il existe un appariement local $L \in \mathcal{L}$ tel que $[p, q]_L$.

Appariement de fragment

On appelle graphe d'appariement de $\mathcal{L}$-fragment ou tout simplement appariement de fragment, tout appariement pouvant s'écrire comme une union d'éléments de $\mathcal{L}$. Cette définition est équivalente à la suivante : «Pour tout couple (p, q) de positions alignées, il existe un appariement local L tel que $[p, q]_L$ et tel que A soit inclus dans l'ensemble des arêtes du graphe. »

Cette approche consiste à aligner non plus seulement des positions mais des fragments.

Alignement induit

On définit un *alignement induit* comme étant la fermeture transitive d'un appariement.

Appariement fragment-transitif

On peut enfin définir un dernier type de graphe, les alignements fragment-transitifs. On dit d'un alignement qu'il est fragment transitif si pour tout couple d'appariements locaux $(\{f_1, f_2\}, A)$, $(\{f_1, f_3\}, B)$ réalisé, il existe un appariement local $(\{f_2, f_3\}, C)$ réalisé. Autrement dit un appariement fragment transitif est un alignement où la relation d'alignement sur les fragments est transitive.

2.3 Graphe mixte

Soit G_A un appariement de fragment. On définit le graphe d'appariement étendu de G_A, en ajoutant à ce dernier un ensemble d'arc intra-fragment $Arc(G_A)$ défini de la manière suivante : une arête a appartient à $Arc(G_A)$ si et seulement si en notant $a = (p_1, p_2, f)$, il existe un appariement local réalisé dans G_A de la forme $(\{f, g\}, A)$ où g est un fragment quelconque. La différence fondamentale avec le graphe d'alignement étendu de Kececioglu******, est que contrairement à ce dernier qui ne dépend que des hypothèses de départ, notre graphe d'appariement étendu dépend d'un graphe d'appariement. De plus les arcs ne sont plus intra-séquence mais intra-fragment ce qui permet d'ajouter une notion de localité.

Problème de l'alignement de fragment maximum (PAFM) 1
Soit un ensemble de séquence $\mathcal{S}$.
Soit $\mathcal{L}$ un ensemble d'appariement locaux sur $\mathcal{S}$.

Soit ρ une fonction qui à un graphe d'appariement de fragment associe un score.
Trouver le graphe d'appariement fragment-transitif de $\mathcal{L}$-fragment dont le graphe d'alignement étendu ne possède pas de cycle mixte et qui maximise la fonction ρ.

Cependant on peut imaginer d'autres problèmes plus aisés à résoudre en simplifiant les contraintes liées à la transitivité :

Problème 2.1
Soit un ensemble de séquence $\mathcal{S}$.
Soit $\mathcal{L}$ un ensemble d'appariements locaux sur $\mathcal{S}$.
Soit ρ une fonction qui à un graphe d'appariement de fragment associe un score.
Trouver le graphe d'alignement de $\mathcal{L}$-fragment dont le graphe d'alignement étendu ne possède pas de cycle mixte et qui maximise la fonction ρ.

Problème 2.2
Soit un ensemble de séquence $\mathcal{S}$.
Soit $\mathcal{L}$ un ensemble d'appariement locaux sur $\mathcal{S}$.
Soit ρ une fonction qui à un graphe d'appariement de fragment associe un score.
Trouver le graphe d'appariement de $\mathcal{L}$-fragment dont le graphe d'alignement induit étendu ne possède pas de cycle mixte et qui maximise la fonction ρ.

2.4 Relation d'ordre définie localement

De même que pour le problème GMT, nous allons définir une relation d'ordre pour formuler autrement le PAFM.

On définit la relation $\leq_f$ comme étant la relation sur les positions $\leq$ restreinte au fragment f.

On peut ainsi définir un relation $\preceq_F$ entre deux composantes connexes en posant $C_1 \preceq_F C_2$ si et seulement si il existe un fragment $f \in F$ et deux positions a_1 et a_2 de f avec $a_1 \in C_1$, $a_2 \in C_2$ et $a_1 \leq_f a_2$. On dit que $\preceq_F$ est la relation $\preceq$ restreinte à F.

En notant $\mathcal{F}(App)$ l'ensemble des fragments f utilisés dans l'appariement de fragment App, on a la proposition :

Propriété 2.1
Soit G un graphe, et App un appariement de fragments sur G les deux propositions suivantes sont équivalentes :
- *G ne possède pas de cycle mixte local*
- *La fermeture transitive de la relation $\preceq_{\mathcal{F}(App)}$ est une relation d'ordre strict sur G.*

La démonstration de cette proposition est la même que dans l'article [1], la notion de localité ne posant aucune difficulté. La propriété précédente nous permet de donner un autre formulation au PAFM :

Problème de l'alignement de fragment maximum (PAFM) 2
Soit un ensemble de séquence $\mathcal{S}$.
Soit $\mathcal{L}$ un ensemble d'appariement locaux sur $\mathcal{S}$.
Soit ρ une fonction qui à un graphe d'appariement de fragment associe un score.
Trouver le graphe d'alignement de $\mathcal{L}$-fragment tel que la fermeture transitive de la relation $\preceq$ restreinte au graphe d'alignement étendu soit une relation d'ordre total et qui maximise la fonction ρ.

3 Fonction de score et formulation en programmation linéaire

3.1 Fonction de score

Le choix d'une fonction de score n'est pas aussi trivial que pour le GMT. En effet, dans ce dernier, les blocs sont distincts, alors que nous n'interdisons pas a priori aux alignements locaux de l'être. Ainsi, si l'on dispose de scores associés aux alignements locaux, il ne suffit plus de les sommer. En effet comme on peut le voir sur la figure 1, si on considère comme score d'alignement de fragment la somme des scores des positions alignées, les deux schémas bien que représentant le même graphe d'alignement n'ont pas le même score. Dans la schéma **A** la partie verte du **B** est comptée deux fois. Il faut donc imposer la condition suivante, si deux appariements locaux L_1 et L_2 sont réalisés, alors pour tous couples (p, q) on ne peut avoir $[p, q]_{L_1}$ et $[p, q]_{L_2}$. Autrement dit, une arête ne peut appartenir à deux appariements locaux réalisés. On dit dans ce cas que L_1 et L_2 se chevauchent. En fait il faut bien comprendre que l'on ne perd pas d'information. En effet le schéma **A** peut être remplacé par un graphe vérifiant la propriété de non chevauchement, celui de la figure **B**.

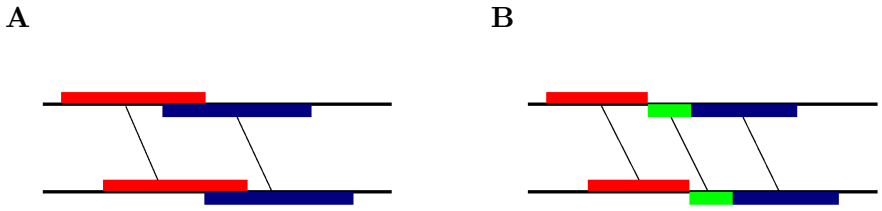


FIG. 1 – Problème de chevauchement

3.2 Le cas particulier DIALIGN

La formulation du PAFM en programmation linéaire en terme de cycle mixte, nécessite une variable par position et une variable par fragment. En se limitant aux cas où les appariements locaux sont donnés par le programme DIALIGN on peut simplifier la formulation du problème en se limitant aux variables sur les fragments. Pour plus de détail le lecteur se référera à l'annexe.

Les alignements donnés par l'algorithme DIALIGN peuvent être vus comme des alignements matriciels sans symbole gap ("-"). En effet, soit (f_1, f_2, A) un alignement de type DIALIGN, alors f_1 et f_2 ont la même taille l et $A = \{ (f_1[i], f_2[i]) , 0 \leq i \leq l-1\}$, où $f[i]$ est la i^{eme} position de f.

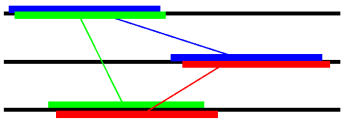
Contraintes

On définit deux nouvelles contraintes sur les fragments. On dit qu'un alignement possède des croisements stricts locaux relativement à $\mathcal{F}$ s'il existe deux positions $a \leq b$ d'un premier fragment $f_1 \in \mathcal{F}$ et $c \leq d$ deux positions d'un autre fragment $f_2 \in \mathcal{F}$ vérifiant $(a, b) \neq (c, d)$ tel que a soit aligné avec d et b avec c. On dit qu'un alignement vérifie la contrainte de non-duplication locale relativement à $\mathcal{F}$ deux positions d'un même fragment ne sont jamais alignées ensemble.

Discussion

On montre sans trop de difficulté que dans le cadre de DIALIGN avec transitivité, interdire un cycle mixte critique et local est équivalent aux contraintes locales de non-croisement et de non-duplication. Cependant ces dernières ont pour avantage d'avoir une formulation en programmation linéaire qui n'utilise que deux variables par contrainte contrairement à la formulation avec cycle mixte qui peut en contenir beaucoup plus[3]. Enfin cela permet de supprimer toutes les variables sur les positions, simplifiant ansi le problème à résoudre

Cependant, en pratique, on ne peut pas se passer des variables sur les positions. En effet on ne trouve pas de triplet de fragment (A, B, C) tel que les couples (A, B) (B, C) (C, A) soient alignables. Ce problème vient du fait que les domaines biologiques n'ont pas de bords clairement définis et sont représentés par plusieurs fragments comme on peut le voir sur la figure ci-dessous.



[3]Le lecteur trouvera toutes les inéquations utilisées pour formuler le problème en programmation linéaire sur les nombres entiers en annexe.

Conclusion et perspectives

Tout comme l'article [2] avait défini rigoureusement le problème de l'alignement matriciel en terme de graphe, nous avons cherché une définition analogue pour les alignements autorisant la duplication de fragment (deux fragments d'une même séquence alignés ensemble) et les croisements. Nous avons pour cela défini une famille de graphe vérifiant localement les contraintes de l'article [2] (les alignements au niveau local sont représentables sous forme de matrice) mais qui globalement permettent des alignements avec copie et duplication.

La formulation en programmation linéaire est relativement aisée, mais la résolution semble dans le cas général difficilement abordable. Il s'agit de résoudre un problème de programmation linéaire en nombre entier, donc un problème NP-complet. En se restreignant à une certaine classe d'appariements locaux proposés par l'algorithme de DIALIGN, on pouvait simplifier le problème de façon significative en supprimant de nombreuses variables et les inéquations faisant intervenir plus de trois variables. Cependant, empiriquement, la contrainte de fragment-transitivé empêche d'obtenir des alignements de triplets de fragments. En effet, si on a trois fragments A, B, C, avec A et B alignables ainsi que A et C, les deux autres fragments B et C ne sont généralement pas alignables. Ce problème est dû au fait que les domaines qui ont un sens biologique n'ont pas de limites clairement déterminées. Ce problème pourrait être résolu en regroupant les fragments par familles. Intuitivement deux fragments sont dans la même famille s' ils correspondent au même domaine biologique. Par exemple, pour une séquence S si on a un fragment commençant à la position (S,5) et se terminant à la position (S,50), et un second fragment qui va de la position (S,7) à la position (S,49), on comprend bien qu'ils correspondent à un même domaine biologique dont les bords ne sont pas clairement définis.

Afin de vérifier si notre formulation du problème permettait de trouver des alignements aussi intéressants que ceux de la littérature, j'ai été ammené à implémenter un programme qui à partir d'un jeu de séquence de protéine renvoit le graphe solution (sous le format PLMA). Nous avons utiliser le logiciel Cplex pour la résolution des équations linéaires en nombre entier.

Nous avons donc fait des premiers tests en abandonnant la notion d'appariement fragment-transitif (transitivité sur les fragments) pour la notion d'alignement (transitivité sur les positions). Ces premiers tests que nous avons effectués sur des exemples de la littérature ont pris beaucoup de temps, (interruption après 12h de calcul) sauf sur un jeu de protéines de petites tailles. Nous n'avons par contre pas eu le temps de faire suffisament de tests, que ce soit pour juger de la pertinence du PAFM, ou pour fixer les limites de la programmation linéaire. Nous ne savons pas encore en effet, la taille maximale des données à partir de laquelle le PAFM ne peux plus être résolue par programmation linéaire en un temps convenable.

Références

[1] Alexander Bockmayr and Knut Reinert. *Mathematische Aspekte der Bioinformatik*.

[2] John D. Kececioglu, Hans-Peter Lenhof, Kurt Mehlhorn, Petra Mutzel, Knut Reinert, and Martin Vingron. A polyhedral approach to sequence alignment problems. *DAMATH : Discrete Applied Mathematics and Combinatorial Operations Research and Computer Science*, 104, 2000.

[3] Paul Pevzner, Haixu Tang, Benjamin Raphael, and Degui Zhi. A novel method for multiple alignement of sequences with repeated and shuffled elements. *Genome Research*, 14 :2236–2346, 2004.

[4] Paul Pevzner, Haixu Tang, and Glenn Tesler. De novo repeat classification and fragment assembly. *Genome Research*, 14 :1786–1796, 2004.

A Annexes

Cette annexe donne la formulation des différentes contraintes en programmation linéaire.

A.1 Notations

On note X_{p_1,p_2} la variable qui est égale à 1 si et seulement si les positions p_1 et p_2 sont alignées. On note de même X_L la variable qui est égale à 1 si et seulement si l'appariement local L est réalisé. On note $\Phi(p,q)$ l'ensemble des alignements locaux L vérifiant $[p,q]_L$.

A.2 L'approche positions

Dans cette partie nous définissons les inéquations dans le cas où la fonction de score se calcule à partir des couples de positions alignées et où le graphe recherché est un graphe d'alignement.

Fonction à optimiser

La fonction à optimiser sera :

$$\sum_{p_1,p_2\ alignable} \rho(p_1,p_2)\cdot X_{p_1,p_2} \tag{1}$$

Contrainte de transitivité

Pour tout triplet $(p1,p2,p3)$ tel que (p_1,p_2), (p_1,p_3) et (p_2,p_3) soient alignables, on définit la contrainte de transitivité :

$$X_{p_1,p_2} + X_{p_1,p_3} - X_{p_2,p_3} \leq 1 \tag{2}$$

Si (p_1,p_2) et (p_1,p_3) sont alignables alors que (p_2,p_3) ne l'est pas alors on considère la contrainte :

$$X_{p_1,p_2} + X_{p_1,p_3} \leq 1 \tag{3}$$

Inégalité du cycle mixte

Le rôle de cette contrainte est de vérifier qu'il n'y a pas de cycle mixte critique dans le graphe solution. Pour chaque cycle mixte C, en notant $\mathcal{P}(C)$ l'ensemble des arêtes de C, on définit la contrainte :

$$\sum_{(p_1,p_2)\in\mathcal{P}(C)} X_{p_1,p_2} \leq |\mathcal{P}(C)| - 1 \tag{4}$$

Appariement de fragment

Soit $L = (\{f_1, f_2\}, A)$ un appariement local. La condition d'alignement de fragment se traduit pour tout $(p, q) \in A$ par :

$$F_L - X_{p,q} \leq 0 \tag{5}$$

$$X_{p,q} - \sum_{L \in \Phi(p,q)} F_L \leq 0 \tag{6}$$

La condition (5) traduit le fait que si deux fragments sont alignés, alors les positions correspondantes doivent être alignées. La condition (6) exprime la réciproque : «si deux positions p et q sont alignées, alors il existe une paire de fragments de $\Phi(p, q)$ alignés».

A.3 L'approche fragment

Dans cette partie nous définissons les équations dans le cadre d'un appariement fragment-transitif où l'on calcule la fonction de score à partir des alignements locaux réalisés.

Fonction à optimiser

On remplace la fonction à optimiser (1) par :

$$\sum_{L \in \mathcal{L}} \rho(L) \cdot F_L \tag{7}$$

Contrainte de fragment-transitivité

On peut aussi remplacer les contraintes de transitivité sur les positions (2 et 3) par des contraintes sur les fragments. Pour tous triplet (L_a, L_b, L_c), avec $L_a = (\{f_1, f_2\}, A)$, $L_b = (\{f_2, f_3\}, B)$ et $L_c = (\{f_1, f_3\}, C)$, on définit la contrainte de transitivité :

$$X_{L_a} + X_{L_b} - X_{L_c} \leq 1 \tag{8}$$

Si (f_1, f_2) et (f_1, f_3) sont alignables alors que (f_2, f_3) ne l'est pas alors on considère la contrainte :

$$X_{L_a} + X_{L_b} \leq 1 \tag{9}$$

Contrainte de non-chevauchement

Pour tout couple d'appariement locaux (L_a, L_b) qui se chevauchent, on rajoute la contraint

$$X_{L_a} + X_{L_b} < 1 \tag{10}$$

Pour vérifier que le graphe correspond bien aux hypothèse, il faut aussi considérer les inéquations 4, 5 et 6.

A.4 L'approche fragment avec DIALIGN

Dans le cas où les appariements locaux sont de type DIALIGN, on peut remplacer l'inéquation 4 par : Pour tout couple d'appariement locaux (L_a, L_b) qui violent les contraintes de non-duplication et de non-croisement,

$$X_{L_a} + X_{L_b} < 1 \tag{11}$$

On n'utilise plus alors les variables sur les positions. Les inéquations 5 et 6 deviennent inutiles.