le cnam

Atelier de BioInformatique

RAPPORT DE STAGE

présentée par Sandrine PERRIN

Dans le cadre de la Licence professionnelle
de biotechnologie, option bioinformatique

Calcul de score d'alignements multiples de séquences

Encadrants:

Joël Pothier,	Maître de conférence
Sophie Brouillet	Ingénieur d'étude

Stage réalisé du 1er juin au 15 septembre 2010 à

Atelier de BioInformatique (ABI)
Université Paris VI
Maison de la Pédagogie, Aile C - RdC
4 place Jussieu, 75005 Paris
site web : http://wwwabi.snv.jussieu.fr/

Remerciements

Je tiens vivement à remercier Joël Pothier pour m'avoir accueillie au sein de son équipe à l'ABI et pour toutes les démarches entreprises qui ont rendu ce stage possible. Je le remercie pour son encadrement, sa patience (dont j'ai souvent abusé), sa gentillesse, sa disponibilité et sa grande pédagogique.
Cela a été réellement très enrichissant !

Je remercie Sophie Brouil et pour son accompagnement et ses conseils tant techniques qu'humains.

Je remercie également Isabelle Gonçalves pour la relecture éclairée de ce rapport.

Citation de Hubbard – 1996

« l'alignement par paire chuchote... l'alignement multiple crie »

Table des matières

1. Introduction

1.1. Présentation du laboratoire d'accueil

L'Atelier de BioInformatique (ABI), créé en 1992, est une "structure ouverte" rassemblant biologistes, biophysiciens, informaticiens et mathématiciens rattachés à des laboratoires variés et désirant travailler à l'interface Biologie/Informatique.

L'essentiel de l'activité s'articule autour de l'analyse de séquences et de structures biologiques basée sur l'utilisation de l'alignement multiple de séquences, le criblage de banques de séquences et de structures, la recherche de motifs lexicaux ou structuraux récurrents.

1.2. Présentation de l'alignement multiple

L'alignement multiple de séquences est un outil fondamental pour de nombreuses analyses en biologie. Il permet de comparer un groupe de protéines ou de gènes apparentés, afin d'établir des relations évolutives. Si deux séquences ont une similarité significative, il est fait l'hypothèse qu'elles partagent un ancêtre commun, elles sont donc homologues. Si deux séquences ont des motifs communs, il est fait l'hypothèse qu'elles sont soumises à une pression de sélection qui empêchent les mutations de se fixer, probablement parce que le motif est important pour assurer une fonction.

L'alignement multiple est principalement utilisé pour :

- trouver des caractéristiques communes à une famille de protéines soit des régions conservées (des motifs), soit des acides aminés strictement conservés permettant de relier une séquence à une structure et à une fonction ;
- construire l'arbre phylogénétique des séquences homologues considérées;
- déduire des contraintes de structures pour les ARN.

a) Principaux algorithmes d'alignement multiple

Les algorithmes d'alignement global développés depuis 1972 reposent sur la programmation dynamique. Elle s'appuie sur le principe qu'une solution optimale s'appuie elle-même sur des sous-problèmes résolus de façon optimale. Les premiers algorithmes d'alignement multiple sont issus de l'adaptation d'algorithmes l'alignement deux à deux.

Les trois principaux algorithmes d'alignement multiple présentés ici permettent de couvrir les principales méthodes utilisées.

Alignement global par programmation dynamique

CLUSTAL [1] : programme le plus couramment utilisé. Il se base sur l'algorithme d'alignement global deux à deux développé par Needleman et Wunsch (1970). Il utilise une approche progressive en construisant un alignement à partir de séquences ou groupes de séquences alignées deux à deux selon un ordre de branchement donné par un arbre de distance. On part des séquences les plus proches en intégrant progressivement les séquences plus éloignées.

MUSCLE [2,3] : programme basé sur une méthode similaire à Clustal d'alignement progressif, avec une optimisation de l'alignement obtenu par maximisation du score. L'arbre guide obtenu est parcouru par valeur de distance décroissante. Dans la phase d'optimisation, chaque branche est coupée pour obtenir deux sous-arbres qui sont ré-alignés. Si le score augmente, le nouvel alignement est retenu. L'itération s'arrête quand toutes les branches de l'arbre ont été visitées sans qu'un changement soit retenu.

Alignement local

DIALIGN [4] : algorithme basé sur une méthode itérative pour repérer des similarités locales fortes entre les séquences (ex : diagonales du dot plot) pour construire un alignement. La sélection des meilleures diagonales se base sur la somme des paires en choisissant le meilleur score.

b) Score & somme des paires

Le score d'un alignement multiple doit rendre compte de la qualité de l'alignement. Les algorithmes utilisés cherchent à maximiser ce score, qui est une indication de l'alignement optimal.

Quelle que soit la méthode d'alignement multiple, le problème de la méthode de calcul du score se pose. La plus utilisée est le score somme des paires (SP) " sum of pairs " : somme sur chaque colonne de tous les scores entre acides aminés pris deux à deux (selon une matrice de substitution). En faisant la moyenne par paires ou la somme sur l'ensemble des colonnes, on obtient un score pour l'alignement.

En outre, chaque algorithme implémente son propre calcul de score selon plusieurs critères, notamment :

- les modalités de prise en compte des pénalités de gap : ouverture, extension, fermeture ;
- la prise en compte de la région concernée : défavoriser les gaps dans les régions hydrophobes et les favoriser dans les régions hydrophiles.

c) Problématique

Face aux nombreux algorithmes d'alignements multiples disponibles, possédant chacun ses propres avantages et inconvénients, et son propre calcul de score, il est difficile de juger quel est le meilleur, ou le plus adapté à son problème.

Il a semblé utile de disposer d'un outil de comparaison d'alignements multiples en élaborant une méthode de calcul de score d'un alignement multiple de séquences protéiques donné. Cette méthode est basée sur la somme des scores des paires des acides aminés par colonne.

En soumettant au programme différents alignements d'un même ensemble de séquences obtenus avec plusieurs algorithmes, le programme fournira des éléments de comparaison indépendants de l'algorithme utilisé. Ainsi le bioanalyste pourra choisir celui répondant le mieux à ses besoins.

2. Matériel : les moyens informatique

Le travail a été réalisé sous Mac OS X. L'essentiel du programme a été écrit en langage C et j'ai fait en sorte qu'il s'intègre dans la panoplie des outils développés à l'Atelier de BioInformatique. Par exemple, j'ai utilisé et amélioré la routine de lecture des fichiers d'alignements multiples lisant différents formats (fasta, clustal).

L'interface web comprend un formulaire de recherche écrit en html associé à un script cgi en python qui traite la requête. Ce script crée et exécute un script shell qui contient la ligne de commande du programme avec les options correspondant aux paramètres saisis. L'exécution du programme principal va générer un fichier au format texte contenant les résultats. Le script en python utilise ce fichier pour afficher une page de résultats formatée. Les tests ont été réalisés avec MAMP[1], environnement local de serveur web gratuit disponible sous Mac.

1 MAMP : http://www.mamp.info/en/index.html (acronyme de Macintosh, Apache, Mysql et PHP qui propose un environnement local de serveur PHP/mySQL sur un poste Mac OS X pour, par exemple, tester un site internet avant la mise en ligne).

3. Méthode : le programme

3.1. Calcul du score somme des paires

La comparaison entre alignements multiples se base fréquemment sur la valeur du score de la somme des paires de l'alignement. Il repose sur le poids des paires d'acides aminées donné par la matrice de substitution choisie et sur les pénalités de gaps définies.

a) Matrice de substitution

Pour pouvoir comparer plusieurs scores, il faut les ramener à un intervalle commun, ici entre 0 et 1. A chaque étape du calcul, les valeurs intermédiaires du score ont dues être ramenées dans ce même intervalle.

Redéfinition des poids de la matrice de substitution :

$$poids\,normalisé(x) = \frac{poids(x) - Vmin}{Vmax - Vmin}$$

avec - poids(x) : poids initial de la matrice de substitution;
- Vmin : valeur minimale de la matrice de substitution;
- Vmax : valeur maximale de la matrice de substitution;

ex avec Blosum62 : Vmin = -4 et Vmax = 11.

b) Traitement des gaps

La prise en compte des gaps est un point critique du calcul de score. Nous avons choisi de distinguer les différentes positions d'un gap et de permettre à l'utilisateur de modifier les valeurs de pénalité définies par défaut.

On discerne les différentes positions d'un gap par des lettres (O pour ouverture, J pour extension et U pour fermeture), qui sont ajoutées à la matrice de substitution ainsi que les pénalités de gaps face à un acide aminé (AA). Le poids d'une paire gap contre acide aminé est indépendant de ce dernier. Par défaut, les pénalités, modifiables par l'utilisateur, sont définies par rapport à la valeur minimale de la matrice choisie (annexe 1) :

- (O,AA) : ouverture contre acide aminé vaut 3 fois la valeur minimale ;
- (J, AA) : extension contre acide aminé vaut la moitié de la valeur minimale ;
- (U, AA) : fermeture contre acide aminé vaut la moitié de la valeur minimale;
- (O,O) ou (J,J) ou (U,U) : vaut 0 ;
- (O,J) ou (J,U) ou (O,U) : vaut la valeur minimale de la matrice.

Pratiquement, ces valeurs sont entrées dans la matrice de substitution, ce qui permet de traiter les symboles correspondant aux gaps de la même manière que les symboles

correspondant aux acides aminés. Cette opération permet ainsi d'obtenir directement le poids d'une paire comprenant au moins un gap comme on obtient le poids d'une paire d'acides aminés.

c) Score de l'alignement

Le calcul du score se base sur la somme du score des paires par colonne, en considérant les colonnes comme indépendantes. Dans le cas de figure où l'alignement soumis contient plusieurs séquences très similaires face à une ou plusieurs séquences plus éloignées, il est apparu intéressant d'implémenter un calcul de score avec une pondération des séquences selon leur similarité. En effet, on sait que ce ne sont pas les séquences très similaires qui posent problème dans les alignements. Elles seront bien alignées quelle que soit l'algorithme. Donc les séquences très similaires compteront moins dans notre score pour éviter de donner trop d'importance à cette information redondante.

Pour obtenir le score, plusieurs ensembles de données intermédiaires doivent être calculés.

Matrice de similarité entre séquences :

Pour chaque paire de séquence dans l'alignement, le calcul de la somme des scores de substitution ou gap entre elles permet de déterminer leur similarité. Ces similarités sont ramenées à une valeur comprise entre 0 et 1.

$$similarité(i,j) = \frac{\sum_{k=1}^{longueur} w(i_k, j_k)}{longueur\ de\ l'\ alignement}$$

avec - w(x,y) : poids donné par la matrice de substitution modifiée avec l'ajout des pénalités de gaps ;
- longueur : nombre de colonnes de l'alignement ;
- i, j : deux séquences de l'alignement.

Pondération des séquences :

Si la pondération des séquences est à prendre en compte pour le calcul du score, la valeur de pondération pour chaque séquence est obtenue en fonction de la similarité de chaque séquence par rapport aux autres. Le calcul utilise la matrice de similarité précédente.

$$similarité(i) = 1 - \frac{\sum_{j \neq i}^{N} similarité(i,j)}{(N-1) * similarité(i,i)}$$

avec - N : nombre de séquences dans l'alignement;
- i, j : deux séquences de l'alignement.

La somme des poids doit être ramenée à 1 pour utiliser la même routine de calcul de score indépendamment de la pondération ou non des séquences.

$$pondération(i) = \frac{similarité(i)}{\sum_{j=1}^{N} similarité(j)}$$

Le calcul du score de l'alignement est donc défini formellement ainsi :

$$score\,alignement = \frac{\sum_{k=1}^{longueur} \sum_{i=1}^{N} \sum_{j \neq i}^{N} (w(i_k, j_k) * pondération(i) * pondération(j))}{longueur}$$

3.2. Entrées / Sorties

a) Entrées

L'utilisateur doit fournir au programme un ou plusieurs fichiers d'alignement multiple et choisir une matrice de substitution. Une routine, développée par l'équipe, sert à lire les fichiers d'alignements dans différents formats (fasta, clustal). Il faut également préciser l'option de pondération des séquences pour le calcul du score.

Le but du programme étant de pouvoir comparer plusieurs alignements, il est possible de transmettre plusieurs fichiers d'alignements qui seront traités séparément, tous avec les mêmes paramètres (matrice et calcul de score).

b) Sorties

En plus du score global, le programme fournit d'autres données sur l'alignement et une représentation graphique établie à partir du score de chaque colonne.

Fichier de résultats

Le programme génère un fichier contenant le détail de informations calculées à partir de l'alignement multiple :

- des informations générales sur l'alignement : le nombre de séquences, la longueur ;
- le score somme des paires pour l'alignement avec ou sans pondération ;
- des données optionnelles : la pondération de chaque séquence, le score de chaque colonne de l'alignement multiple.

Représentation graphique

Le score le long de l'alignement est représenté graphiquement pour permettre une analyse de la ou des régions à étudier par le bioanalyste. Le score par colonne est ramené à un entier compris entre 0 et 9. Une courbe est tracée au dessus de l'alignement multiple. (annexe 2).

3.3. Utilisation du programme

Une page web permet la soumission au programme d'un ou plusieurs alignements (annexe 3) :

- coller un alignement multiple ou télécharger de un à dix fichiers d'alignements ;
- choisir la matrice de substitution parmi BLOSUM ou PAM ;
- choisir l'option du calcul de score : avec ou sans pondération des séquences ;
- modifier les valeurs de pénalité de gap par défaut.

Une page de réponse est affichée avec un lien vers la représentation graphique de chaque alignement soumis.
Le programme est accessible sur le site de l'ABI : http://wwwabi.snv.jussieu.fr/public/Calimul/

4. Résultats & discussion

4.1. Astuces de codage pour optimiser le temps de calcul

Les alignements multiples peuvent porter sur plusieurs centaines de séquences et donc entraîner des temps de calcul longs. J'ai cherché à optimiser l'opération la plus répétitive du programme : la recherche du poids dans la matrice de substitution pour chaque paire. Le comptage de chaque type d'acide aminé par colonne est réalisé avec un dictionnaire (tableau à une dimension contenant 26 cases, correspondant à chaque lettre; pratiquement, il y a un dictionnaire par colonne de l'alignement).

Le recours au dictionnaire permet, pour chaque colonne, de consulter la matrice de substitution pour chaque type de paires et non plus pour chaque paire. En divisant par le nombre de paires, le score est ramené à une valeur comprise entre 0 et 1.

Si le calcul du score utilise la pondération des séquences, chaque lettre est comptée pour la valeur de pondération calculée pour la séquence considérée.

$$score\,d'une\,colonne = \frac{\sum_{a}^{26}\sum_{b=a+1}^{26} w(a,b)*n_a*n_b + \sum_{a}^{26} w(a,a)*n_a*(n_a-1)/2}{N*(N-1)/2}$$

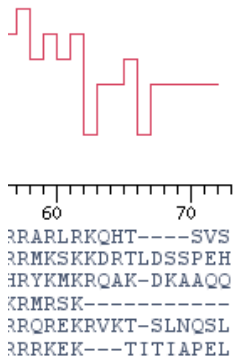
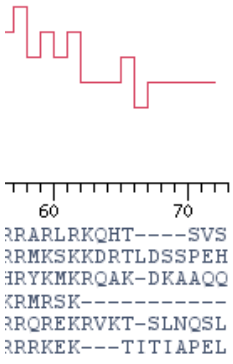
avec - w(a,b) : poids donné par la matrice de substitution modifiée pour traiter les gaps ;
- n_a : nombre de a dans la colonne considérée obtenue à partir du dictionnaire ;
- N : nombre de séquences de l'alignement.

En utilisant cette astuce, le gain obtenu devient d'autant plus important que le nombre de séquences (N) est supérieur à 26 : N+26*25/2 opérations au lieu de N(N-1)/2. L'algorithme reste linéaire avec le nombre de séquence.

Nota Bene : le dictionnaire et la matrice comprennent les 26 lettres dans l'ordre alphabétique, même si certaines ne sont jamais rencontrées dans un alignement. Car la recherche dans la matrice se fait en convertissant directement le code du caractère x en indice dans la matrice de 0 à 25 : (int) x – (int) 'A'.

4.2. Modification du score par colonne pour la représentation graphique

La forte pénalité d'ouverture de gap pèse plus lourdement sur le score de sa colonne que les extensions de gap à côté ne pèsent sur le score de leur colonne. Pour corriger cet artefact dans la représentation graphique, le coût total d'un gap est réparti sur toute la longueur du gap. Le score final est donc inchangé mais les valeurs par colonne s'en trouvent harmonisées.

Représentation sans répartition	Représentation avec répartition
	

L'ensemble des séquences est parcouru pour définir le coût d'un gap en fonction de la longueur (n) de la zone de gaps :

$$coût\,d'un\,gap = \frac{pénalité\ ouverture + pénalité\ extension*(n-2) + pénalité\ \ fermeture}{n}$$

avec - n : nombre de colonne de la zone de gap considérée sur une séquence.

La valeur obtenue servira de poids pour l'occurrence d'un gap contre un acide aminé de la colonne. Elle sera donc propre à chaque colonne.

$$poids\ paires(gap, AA) = \frac{coût\ gap\ pour\ la\ colonne}{n_O + n_J + n_U}$$

avec - n_O, n_J, n_U : nombre de gap ouverture (O), extension (J), fermeture (U) dans la colonne donnée par le dictionnaire.

Pour harmoniser le calcul détaillé au chapitre 4.1, cette valeur est entrée dans la matrice à la place des valeurs définies par défaut pour les trois lettres marquant les gaps face aux lettres symbolisant les acides aminés. Les valeurs pour les paires gap contre gap ne sont pas modifiées.

4.3. Ajout de l'option de maximisation du score

Dans la pratique, un alignement peut être manipulé par un bioanalyste, qui modifie à l'oeil le résultat obtenu d'après ses connaissances sur les séquences étudiées (utilisation d'outil spécifique, tel que Seaview [2]). Les modifications sont apportées au niveau des gaps en déplaçant les acides aminés d'une extrémité à l'autre.

Mon programme va chercher à optimiser l'alignement selon le même principe. Tous les déplacements d'acides aminés bordant une zone de gaps sont testés. La variation du score est calculée, s'il y a un gain, le déplacement est mémorisé dans une pile. Après parcours de tous les gaps de l'alignement, la pile est triée selon le gain de score. Les déplacements sont reportés dans la séquence à partir du meilleur gain si les gaps concernés ne sont pas chevauchant avec les déplacements déjà appliqués. Effectivement, il faut éviter les effets de bord entre gaps pour s'assurer d'une augmentation du score. Après traitement de la pile, le score global est recalculé. L'opération peut être répétée tant qu'un gain du score est observé.

Cette étape n'a pu être complètement finalisée pendant la durée du stage, elle est en cours de finition et sera prochainement disponible.

2 Seaview : outil de visualisation d'alignement multiple et de construction d'arbre phylogénétique http://pbil.univ-lyon1.fr/software/seaview.html

5. Conclusion et perspectives

Le programme développé permet la comparaison des alignements multiples selon plusieurs critères et indépendamment des algorithmes utilisés pour les générer. Il donne un score global pour l'alignement mais également un score par colonne consultable grâce à la représentation graphique (format plus riche que les étoiles proposées dans Clustal). Il offre également à l'utilisateur la possibilité de fixer les poids de pénalités de gap en fonction de ses besoins.

D'autres éléments de comparaison ont été envisagées tel que l'exploitation de la matrice de similarité entre séquences pour réaliser une représentation graphique d'un arbre de distance entre les séquences de l'alignement.

Le programme est mis à disposition de tous et sera utilisé dans des formations d'écoles doctorales et des formations permanentes en analyse de séquence sur Paris 6.

6. Bibliographie

[1] CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice Thompson, J.D. and Higgins, D.G. and Gibson, T.J. Nucleic Acids Res. 1994 November 11; 22(22): 4673–4680.

[2] MUSCLE: multiple sequence alignment with high accuracy and high throughput.
Edgar RC Nucleic Acids Res 2004 Mar 19;32(5):1792-7. Print 2004.

[3] MUSCLE: a multiple sequence alignment method with reduced time and space complexity
Edgar, R.C. 2004 BMC Bioinformatics, (5) 113

[4] DIALIGN 2 : improvement of the segment-to-segment approach to multiple sequence alignment
Bukhard Morgenstern Bioinformatics 1999 Mar;15(3):211-8.

Performance evaluation of amino acid substitution matrices. S. Henikoff and J.G. Henikoff. Proteins: Structure, Function, and Genetics, 17:49-61, 1993.

Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance reliability of predictions. Briffeuil P, Baudoux G, Lambert C, De Bolle X, Vinals C, Feytmans E, Depiereux E. Bioinformatics. 1998;14(4):357-66.

Comparative analysis of multiple protein-sequence alignment methods. McClure MA, Vasi TK, Fitch WM. Mol Biol Evol. 1994 Jul;11(4):571-92.

Cours sur l'alignement de séquences de l'Atelier de BioInformatique :
http://abiens.snv.jussieu.fr/OBI/OBI3/cours_seq.pdf

Sites Internet consultés

Plateforme d'outils bioinformatique : http://mobyle.pasteur.fr/cgi-bin/portal.py?form=dialign

Site de l'EBI, plateforme : http://www.ebi.ac.uk

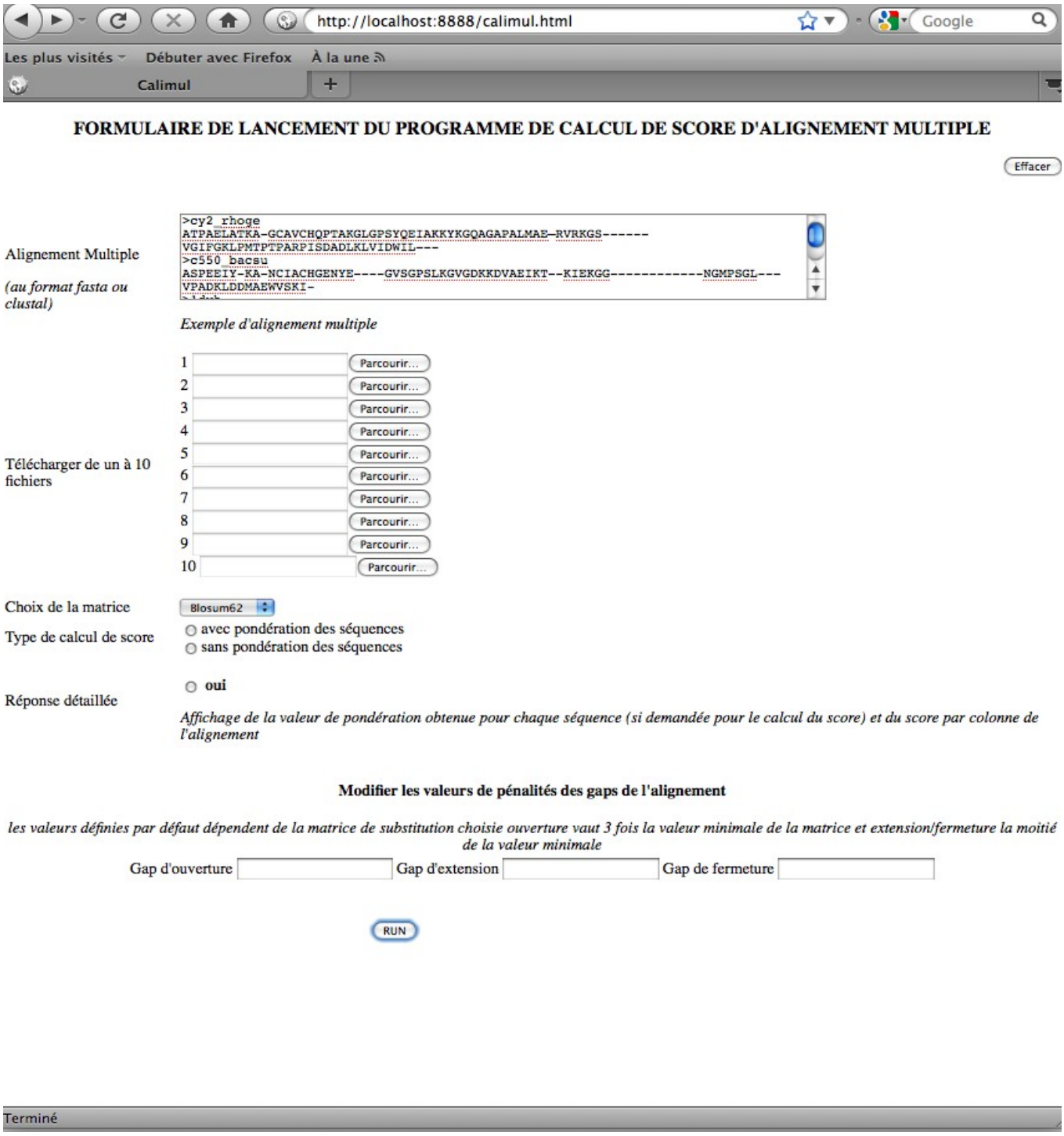
7. Annexes

ANNEXE 1 : Matrice de substitution et pénalité de gap :

Exemple avec la matrice BLOSUM62 adaptée pour le traitement des gaps

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =  -0.5209
#  Prog Calimul : utilisation des lettres O-J-U pour les pénalités de gap
#  O (ouverture) vaut 3 * valeur minimum(-4) du gap
#  J (extension) U (fermeture) valent la moitié de la valeur minimum
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  O  J  U
   4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0-12 -2 -2
  -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1-12 -2 -2
  -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1-12 -2 -2
  -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1-12 -2 -2
   0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2-12 -2 -2
  -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1-12 -2 -2
  -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -1-12 -2 -2
   0 -2  0 -1 -3 -2 -2  6 -2 -4 -4 -2 -3 -3 -2  0 -2 -2 -3 -3 -1 -2 -1-12 -2 -2
  -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1-12 -2 -2
  -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1-12 -2 -2
  -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1-12 -2 -2
  -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1-12 -2 -2
  -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1-12 -2 -2
  -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1-12 -2 -2
  -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2-12 -2 -2
   1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0-12 -2 -2
   0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0-12 -2 -2
  -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2-12 -2 -2
  -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1-12 -2 -2
   0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1-12 -2 -2
  -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1-12 -2 -2
  -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1-12 -2 -2
   0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1-12 -2 -2
-12-12-12-12-12-12-12-12-12-12-12-12-12-12-12-12-12-12-12-12-12-12-12  0 -4 -4
 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -4  0 -4
 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -2 -4 -4  0
```

ANNEXE 2 : Capture écran de l'interface web développée pour le programme



ANNEXE 3 : Représentation graphique du score par colonne

Exemple réalisé à partir d'un alignement multiple réalisé avec cinq séquences de la cytochrome C obtenu avec Muscle. Le calcul du score a été lancé avec la matrice BLOSUM62, les valeurs de pénalités de gap par défaut et sans pondération des séquences.