

Introduction à la bioinformatique

4. L'alignement de plusieurs séquences et les profils

1

Objectifs

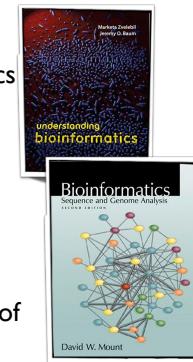
- Comprendre l'avantage d'un alignement entre plusieurs séquences
- Comment faire les alignements de plusieurs séquences
- Comprendre pourquoi la programmation dynamique n'est pas applicable
- Être capable d'expliquer les différents systèmes pour l'alignement de plusieurs séquences
- Comprendre comment on peut aligner des groupes de séquences
- Comprendre les profils (PSSM) et leur importance pour l'alignement de plusieurs séquences
- Être capable d'expliquer l'importance du *pseudocounts*
- Comprendre la différence entre l'alignement progressif et itératif
- Être capable d'expliquer les principes d'alignement progressif
- Être capable d'expliquer les principes d'alignement itératif

3

©Tom Lenaerts, 2012

Bibliographie

- Zvelebil et Baum, Understanding bioinformatics
- D.W. Mount, Bioinformatics: sequence and genome analysis
- Osamu Gotoh (1999) Multiple sequence alignment: algorithms and applications. *Adv. Biophys.* 36:159-206
- Cédric Notredame (2007) Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology* 3(8):e123
- Robert C. Edgar and Serafim Batzoglou (2006) Multiple sequence alignment. *Current opinion in structural biology* 16:368-373



2

©Tom Lenaerts, 2012

Pourquoi?

- Un alignement entre deux séquences produit une hypothèse qui est confirmée ou rejetée par le score
 - Mais ce score ne donne pas une garantie que la relation entre les deux séquences est vraiment liée à un ancêtre commun
 - De plus, il y a toujours des petits erreurs dans l'alignement
- On peut résoudre cette incertitude en ajoutant des séquences additionnelles
- Un alignement de plusieurs séquences (*multiple sequence alignment - MSA*) donne de l'information additionnelle pour chaque position:
 - similarité entre des positions ou
 - la conservation de certains acides aminés dans des positions spécifiques

4

Pourquoi? 2

Améliorer l'alignement entre deux séquences

(A) p110 α
cAMP-kinase

```
TFILGIGDRHNSNIMVKDDG-QLFHIDFGHFLDHKKKKFGYKRERVPFVLT--QDFLIVI 142
QIVLTFEYLHSLDLIYRDLKPENLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPE 179
```

(B) p110 β
p110 δ
p110 α
p110 γ
p110_dicti
cAMP-kinase

```
SYVLGIG-----DRHSDNINVKKTGQLFHIDFGHILGNFKSKFGIKRERVPFILT 136
TYVLGIG-----DRHSDNIMIRESGQLFHIDFGHFLGNFKTKFGINRERVPFILT 136
TFILGIG-----DRHNSNIMVKDDGQLFHIDFGHFLDHKKKKFGYKRERVPFVLT 135
TFVLGIG-----DRHNDNIMITETGQLFHIDFGHILGNYKSFGLINKERVPFVLT 135
TYVLGIG-----DRHNDNLMVTKGRLFHIDFGHFLGNYKKKFGFKRERAPFVFT 135
QIVLTFEYLHSLDLIYRDLKPENLIDQQGYIQVTDFGFAKRVKGRTWXLCG--TPEYLA 177
```

Les régions conservées: en vert les résidus identiques et en bleu les résidus avec les mêmes propriétés

5

Pourquoi ? 3

Les régions conservées donnent de l'information sur la fonction et la structure d'une protéine

P06241| 149-246
Q06124| 6-102
P62993| 60-152
P12931| 151-248
P41240| 82-171
P00519| 127-217
P20936| 181-272
P42224| 573-670
O60674| 401-482

```
WYFGKLGR---KDAERQLLSFGNPRGTFPLIRESETTK-GAYSLISRWDMMKGDHVKHYKI
WFHPIITGVAENLLTRG-VGDFSLARPSKS-N-PGDFTLISVRNRG---AVTHIKI
WFHGKIPRAKAEMLSKQ-R-HDGAFLIRESES-A-PGDFSLSVKFGN---DVQHFKV
WYFGKITRRESERLLNAEN-PRGTFPLIRESETT-KGAYCISVSDFDNAKGLNVKHYKI
WFHGKITRQAERLVPET---GLFLYRESTNY-PGDFTLICVSCDG---KVEHYRI
WYHGPVSRNAAEYLLSSGIN---GSFLVRESESS-PGQRSISLRYEG---RVVHYRI
WYHGLDKDTIAEERLHQAGK---SGSYLIRESDR-PGFSVVISLQSMN---VUNHFRRI
WNDCGIMCPISKERERALLKDQ-QPCTFLRPFSESSREGAITPTWVERSQNG-GE--P-
--HGPISM---DFAISKLKKAGNQITGLYVLRCSPKDF-NKYFLTFAVER--ENVIEYKCI
```

P06241| 149-246
Q06124| 6-102
P62993| 60-152
P12931| 151-248
P41240| 82-171

CLUSTAL
<http://www.clustal.org/>

IAOT.pdb

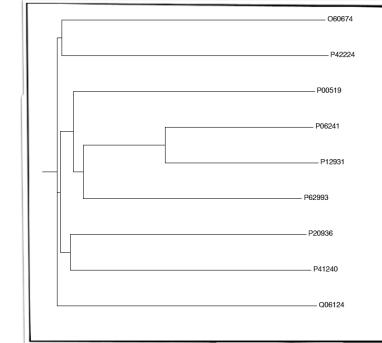
6

Pourquoi ? 5

Déterminer les relations évolutives

Un arbre phylogénétique

```
(O60674:0.14917,P42224:0.15083)
:0.00281,
(
(P00519:0.13675,
(
(P06241:0.08357,P12931:0.08643):0.04625,
P62993:0.12375)
:0.00575)
:0.00719,
(P20936:0.13375,P41240:0.13625)
:0.00531)
:0.00219,
Q06124:0.14719);
```



7

8

Pourquoi ? 4

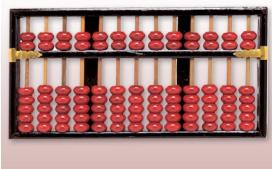
Mais il peut y avoir des différences entre les méthodes

P06241| 149-246 WYFGKLGR---KDAERQLLSFGNPRGTFPLIRESETTK-GAYSLISRWDMMKGDHVKHYKI
Q06124| 6-102 WFHPIITGVAENLLTRG-VGDFSLARPSKS-N-PGDFTLISVRNRG---AVTHIKI
P62993| 60-152 WFHGKIPRAKAEMLSKQ-R-HDGAFLIRESES-A-PGDFSLSVKFGN---DVQHFKV
P12931| 151-248 WYFGKITRRESERLLNAEN-PRGTFPLIRESETT-KGAYCISVSDFDNAKGLNVKHYKI
P41240| 82-171 WFHGKITRQAERLVPET---GLFLYRESTNY-PGDFTLICVSCDG---KVEHYRI
P00519| 127-217 WYHGPVSRNAAEYLLSSGIN---GSFLVRESESS-PGQRSISLRYEG---RVVHYRI
P20936| 181-272 WYHGLDKDTIAEERLHQAGK---SGSYLIRESDR-PGFSVVISLQSMN---VUNHFRRI
P42224| 573-670 WNDCGIMCPISKERERALLKDQ-QPCTFLRPFSESSREGAITPTWVERSQNG-GE--P-
O60674| 401-482 --HGPISM---DFAISKLKKAGNQITGLYVLRCSPKDF-NKYFLTFAVER--ENVIEYKCI

TCOFFEE
[http://www.ebi.ac.uk/
Tools/t-coffee/](http://www.ebi.ac.uk/Tools/t-coffee/)

Le problème

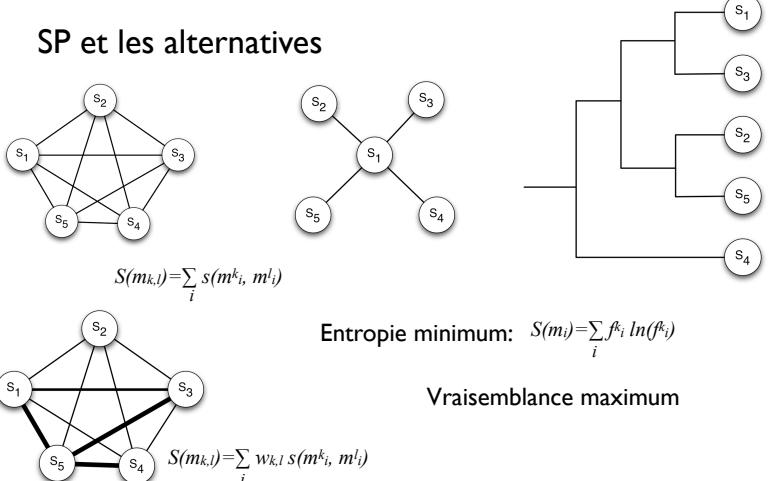
- Calculer l'alignement entre N séquences est un **problème d'optimisation combinatoire** (*combinatorial optimisation problem - COP*)
- Pour résoudre un COP, il faut fournir 2 systèmes. Un système pour
 - Assigner le score** d'alignement entre toutes les séquences
 - Trouver l'alignement avec le score optimal**
 - global ou local



9

Le score 2

SP et les alternatives



11

Le score

Comment peut-on assigner un score à un MSA?

$$\text{Le score total : } S(m) = \sum_{k,l} S(m_{k,l})$$

La somme de paires ou SP

On fait l'hypothèse que les scores de colonnes différentes sont indépendants

$$\text{Le score d'une colonne : } S(m_{k,l}) = \sum_i s(m^k_i, m^l_i)$$

m^k_i est le résidu dans la séquence k dans la colonne i
 $s(m^k_i, m^l_i)$ le score dans la matrice de substitution

10

Les Méthodes globales

Les algorithmes de Smith-Waterman et Needleman-Wunsch peuvent être utilisés pour la construction d'un APS

MAIS : l'approche n'est pas pratique car elle a besoin de beaucoup de **ressources de calcul** (taille = 200).

Nombre de séquences	$O(2^n L^n)$
2	$2^2 \times 200^2 = 0.16M$
3	$2^3 \times 200^3 = 64M$
4	$2^4 \times 200^4 = 25600M$
6	...

12

Les Méthodes globales 2

Les algorithmes de Smith-Waterman et Needleman-Wunsch peuvent être utilisés pour la construction d'un APS

MAIS : l'approche n'est pas pratique car il a besoin de beaucoup de **mémoire** (taille = 200).

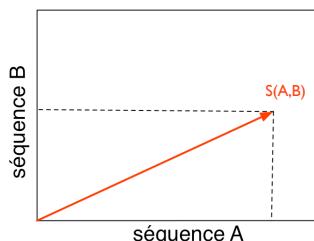
Nombre de séquences	mémoire (1 byte/élément)
2	400 bytes
3	7.63 Mbytes
4	1.5Gbytes
6	60000 Gbytes

13

Le système Lipman

Lipman et al ont proposé un logiciel qui utilise la programmation dynamique

Rappelez-vous ...



Prenez 2 séquences: A et B

Pour aligner 2 séquences nous devons calculer les scores pour chaque position jusqu'à la fin

$S(A,B)$ est le score optimal pour l'alignement de deux sous-séquences: A et B

Les Méthodes globales 3

- La programmation dynamique optimisée (le système de Lipman et al.)
- L'alignement progressif (le système CLUSTAL)
- Méthodes stochastiques (le système SAGA)

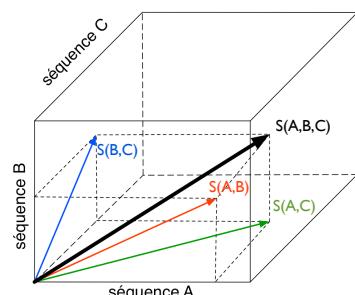
14

Le système Lipman 2

Lipman et al ont proposé un logiciel qui utilise la programmation dynamique

Prenez 3 séquences: A, B et C

Pour aligner 3 séquences nous devons calculer les scores optimaux pour chaque position dans un cube



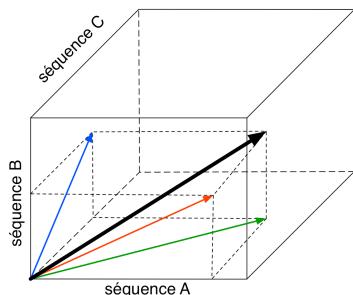
Le score $S(A,B,C)$ est lié aux scores $S(A,B)$, $S(B,C)$ et $S(A,C)$

somme de paires ou SP

15

Le système Lipman 3

Lipman et al ont proposé un logiciel qui utilise la programmation dynamique



Pour N séquences de 200 acides aminés on doit enregistrer 200^N scores

Comment peut-on réduire ceci de sorte qu'on puisse encore trouver la solution optimale ?

17

Le système Lipman 5

Carrillo et Lipman ont trouvé une méthode qui réduit le nombre de comparaison qu'on doit faire

Étapes de prétraitement :

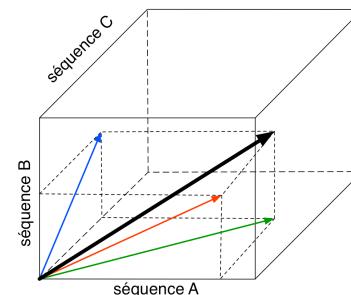
1. Calculer les scores optimaux entre chaque paire de séquences
2. Utiliser ces scores pour la construction d'un arbre phylogénétique
3. Construire le MSA en utilisant l'arbre et **une méthode heuristique**

Cet MSA temporaire donne les limites sur l'espace à l'intérieur du cube dans lequel on trouvera l'alignement optimal

19

Le système Lipman 4

Carrillo et Lipman ont trouvé une méthode qui réduit le nombre de comparaisons qu'on doit faire



La flèche **noire** (alignement pour 3 séquences) peut être projetée sur les surfaces **AB**, **AC** et **BC**, qui représentent un alignement pour chaque paire de séquences

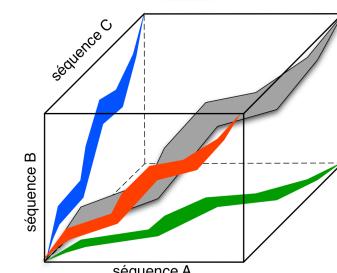
Cela veut dire aussi que les alignements pour chaque paire introduisent des limites sur les positions qui sont importantes pour l'alignement de 3 séquences !

18

Le système Lipman 6

Le méthode de Carrillo et Lipman introduit une limite sur le nombre de positions qui sera calculé en utilisant la programmation dynamique

Le nombre de séquences est limité à 10 !!!



le MSA optimal est donc l'alignement avec le plus haut SP score

Le score $S(A,B,C)$ est calculé en utilisant la méthode SP

Une pénalité constante est utilisée pour chaque taille d'espace

20

Le système Lipman 7

L'algorithme calcule aussi une valeur ϵ pour chaque paire de séquences

ϵ représente la divergence entre l'alignement par paire et l'alignement avec tous les séquences.

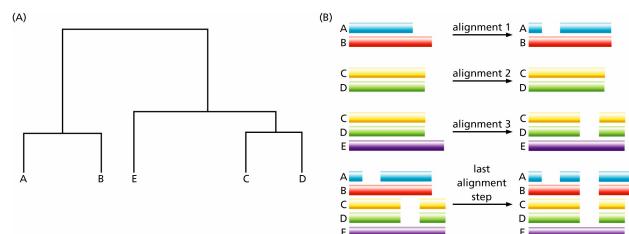
$$\epsilon = PSA(x) - MSA(x)$$

MSA essaie de diminuer la divergence, autrement l'alignement de paires ne donne pas assez d'information concernant l'alignement de toutes les séquences

21

Profils

Dans l'étape 4 de l'algorithme, on construit un MSA temporaire en utilisant un arbre. Dans cette étape, on a besoin des algorithmes qui peuvent aligner des séquences aux groupes de séquences ou des groupes de séquences aux autres groupes



Un profil est une représentation compacte d'un groupe de séquences alignées

23

Le système Lipman 8

L'algorithme complet:

1. Calculer les scores pour les alignements entre chaque paire de séquences
2. Utiliser ces scores pour la construction d'un arbre phylogénétique
3. Calculer les poids pour chaque paire de séquences en utilisant l'arbre
4. Produire l'alignement en utilisant une heuristique et l'arbre (non-optimal)
5. Calculer le ϵ maximum pour chaque paire de séquences (important pour la pondération du SP)
6. Déterminer les positions dans l'hyper-cube (dimensions N) qui seront calculées pour obtenir l'alignement optimal
7. Appliquer la programmation dynamique
8. Rapporter l'alignement optimal et le ϵ maximum

22

Profils 2

les profils enregistrent les propriétés générales d'une collection de séquences: 1) les fréquences d'acides aminés dans chaque colonne et 2) l'importance évolutive de chaque acide aminé

Prenez par exemple ces séquences:

TGVEAE**N**LL
PRAKAE**E**SLS
GRKDAE**R**QLL

$$f_{u,b} = \frac{n_{u,b}}{N_{seq}}$$

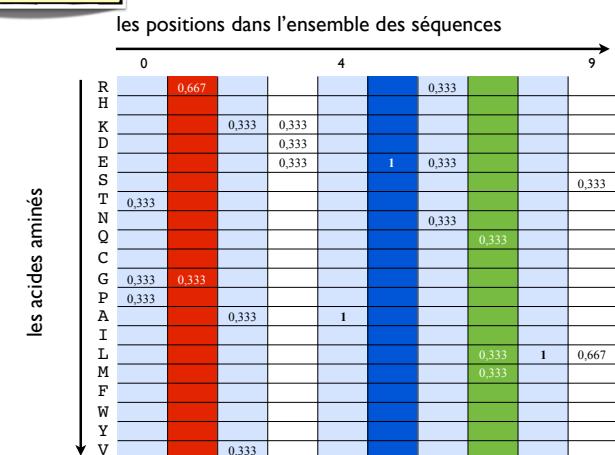
les fréquences sont:

$$f_{2,R=2/3} \quad f_{5,E=3/3} \quad f_{7,S=1/3}$$

$$f_{u,b} = \frac{\ln(1 - (n_{u,b}/(N_{seq} + 1)))}{\ln(1/(N_{seq} + 1))}$$

24

Profils 3

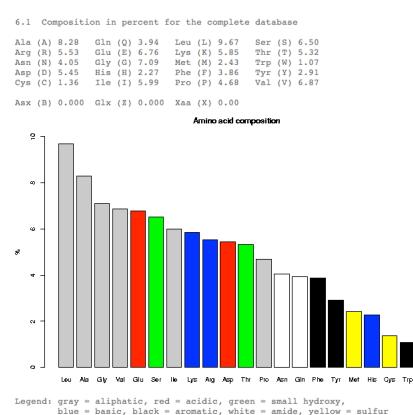


25

Profils 5

La probabilité p_a est la probabilité qu'on trouve l'acide aminé à n'importe quelle position dans les séquences

les données de swissprot



27

Profils 4

Le profil enregistre pour chaque colonne la fréquence des acides aminés multipliée par le score d'alignement (l'importance évolutive)

$$m_{u,a} = \sum_{b \in AA} f_{u,b} S_{a,b}$$

$m_{u,a}$ est un score d'alignement entre un résidu a et la colonne u

$$m_{u,a} = \log \frac{q_{u,a}}{p_a}$$

quand il y a assez de séquences et chaque acide aminés est présent au moins une fois dans chaque colonne

C	1
S	2
T	2
P	6
A	-1
G	-1
N	-1
D	-3
R	-4
Q	3
H	0
I	-1
K	-3
M	-2
L	-2
V	-2
F	0
Y	2
W	15
C	1
S	1
T	1
P	1
A	1
G	1
N	1
D	1
E	1
Q	1
H	1
R	1
K	1
M	1
I	1
L	1
V	1
F	1
Y	1
W	1

26

Profils 6

$$m_{u,a} = \sum_{b \in AA} f_{u,b} S_{a,b}$$

$$m_{0,R} = 0.333 (-1) + 0.333 (0) + 0.333 (-1) = -0.666$$

$$m_{I,R} = 0.667 (5) + 0.333 (0) = 3.335$$

$$m_{2,R} = 0.333 (4) + 0.333 (-1) + 0.333 (-3) = 0$$

0	-0.666	3.335	0	1	-1	0	1.665	-1	-3	-2.334
R	-1	0.668	-1.332	0.666	-2	0	1	-1	-3	-2.334
H										
K										
D										
E										
S										
T										
N										
Q										
C										
G										
P										
A										
I										
L										
M										
F										
Y										
W										

0	-0.666	3.335	0	1	-1	0	1.665	-1	-3	-2.334
R	-1	0.668	-1.332	0.666	-2	0	1	-1	-3	-2.334
H										
K										
D										
E										
S										
T										
N										
Q										
C										
G										
P										
A										
I										
L										
M										
F										
Y										
W										

Sans pénalité, la matrice est une PSSM (Position-specific scoring matrix)

28

Profils 7

Les scores $m_{u,a}$ représentent les scores pour aligner un résidu a à la position u

On utilise les mêmes algorithmes PD pour aligner une séquence à un profil

Le plus grand problème pour créer des profils est l'insuffisance du nombre de séquences et, par conséquence, l'absence de certains acides aminés dans certaines colonnes

$\log 0 = -\infty$

Il est impossible d'aligner un résidu à ces colonnes en utilisant le log-odd score (regardez la discussion sur PAM et BLOSUM)

→PSEUDOCOUNTS

29

Profils 9

L'équation la plus générale est exprimée en fonction de $f_{u,a}$

$$q_{u,a} = \frac{af_{u,a} + \beta p_a}{\alpha + \beta}$$

α est un facteur de cadrage pour les données observées. On utilise parfois $\alpha = N_{seq} - 1$

S'il n'y a pas de données (aucune séquence), les pseudocounts déterminent les valeurs du profil

Les pseudocounts représentent la distribution antérieure, qui est la connaissance qu'on a concernant le système avant l'introduction des données

31

Profils 8

Les pseudocounts sont des constantes qu'on ajoute aux valeurs du profile

$$q_{u,a} = \frac{n_{u,a} + 1}{N_{seq} + 20}$$

Les pseudocounts donnent une information initiale sur les acides aminés

par conséquence, $q_{u,a}$ n'est jamais 0 !

$$q_{u,a} = \frac{n_{u,a} + \beta p_a}{N_{seq} + \beta}$$

$$\beta = \sqrt{N_{seq}}$$

β est un facteur de cadrage déterminant le nombre de pseudocounts

30

Profils 10

$$m_{u,a} = \log \frac{q_{u,a}}{p_a}$$

$$q_{u,a} = \frac{n_{u,a} + \beta p_a}{N_{seq} + \beta}$$

$$\beta = 1$$

$$q_{0,R} = \frac{0.06}{4} \quad m_{0,R} = \log \frac{0.014}{0.06}$$

$$q_{1,R} = \frac{2.06}{4} \quad m_{1,R} = \log \frac{0.51}{0.06}$$

$$q_{6,R} = \frac{1.06}{4} \quad m_{6,R} = \log \frac{0.26}{0.06}$$

	0		4		9				
R	-0.602	0.968	-0.602	-0.602	-0.602	-0.602	0.679	-0.602	-0.602
H	-0.602	-0.602	-0.602	-0.602	-0.602	-0.602	-0.602	-0.602	-0.602
K									
D									
E									
S									
T									
N									
Q									
C									
G									
P									
A									
I									
L									
M									
F									
W									
Y									
V	-0.602	-0.602	0.59	-0.602	-0.602	-0.602	-0.602	-0.602	-0.602
+/-	9	9	9	9	9	9	9	9	9

Attention ! ici la matrice de substitution n'est pas considérée

32

Profils 11

On peut améliorer les pseudocounts en utilisant l'information des matrices de substitution

$$\frac{q_{a,b}}{p_a p_b} = e^{\lambda s(a,b)}$$

Chaque log-odd score dans la matrice contient de l'information sur la probabilité d'alignement de deux acides aminés

c.a.d. si une colonne u contient $f_{u,b}$ acides aminés de type b , la probabilité de rencontrer un alignement avec une acide aminé de type a est proportionnel à

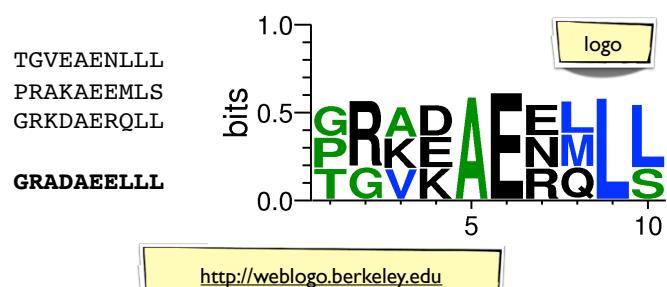
$$f_{u,b} \frac{q_{a,b}}{p_a p_b}$$

la somme de toutes ces probabilités donne la probabilité totale pour a

33

Profils 13

Quand le profil est calculé, on peut aussi calculer une **séquence consensus** qui représente pour chaque position l'acide aminée avec le plus haut score $m_{u,a}$



35

Profils 12

$$g_{u,a} = \sum_b f_{u,b} \frac{q_{a,b}}{p_b}$$

Multiplier la probabilité d'aligner un acide aminé à la colonne u avec p_a produit un meilleur pseudocount pour a

L'équation pour $q_{u,a}$ devient

$$q_{u,a} = \frac{\alpha f_{u,a} + \beta g_{u,a}}{\alpha + \beta}$$

La valeur de $g_{u,a}$ peut être obtenue à partir des matrices de substitution comme PAM et BLOSUM

34

Profils 14

Un logo est construit en calculant le contenu de l'information de chaque colonne u dans la séquence

$$I_u = \log_2 20 - H_u$$

l'information

$$H_u = -\sum f_{u,a} \log_2 f_{u,a}$$

l'incertitude

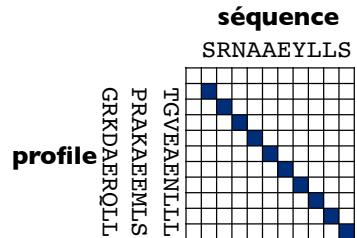
Une position avec un acide aminé conservé aura le maximum d'information

La contribution de chaque résidu est : $f_{u,a} I_u$

36

Aligner à un profil

Le Needleman-Wunsch (L3) ou Smith-Waterman (L3) peut être utilisé pour aligner une séquence à un profile.



Un profil contient des scores et des pénalités

Le plus grand problème se situe dans la manière d'assigner les pénalités

37

Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 4353-4358, July 1987
Biochemistry

Profile analysis: Detection of distantly related proteins

(amino acid/sequence comparison/protein structure/globin structure/immunoglobulin structure)

MICHAEL GRIBSKOV*, ANDREW D. McLACHLAN†, AND DAVID EISENBERG*

*Molecular Biology Institute and Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90024; and †Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England, United Kingdom

Communicated by Paul Boyer, February 17, 1987 (received for review November 19, 1986)

ABSTRACT Profile analysis is a method for detecting distantly related proteins by sequence comparison. The basis for comparison is not only the customary Dayhoff protein-alanine matrix but also the use of structural information and information implicit in the alignments of the sequences of families of similar proteins. This information is expressed in a position-specific scoring table (profile), which is created from a group of sequences previously aligned by structural or sequence criteria. Similarity of a query sequence (probe) to the group of aligned sequences (probe) can be tested by comparing the target to the profile using dynamic programming algorithms. The profile method differs in two major respects from other sequence comparison methods: (i) Any number of known sequences can be used to construct the profile, allowing more information to be used in the testing of the target than is possible with pairwise alignment methods. (ii) The profile includes the penalties for the loss of information at each position, which allows to include the three-dimensional structure in the testing scheme. Tests with globin and immunoglobulin sequences show that profile analysis can distinguish all members of these families from all other sequences in a database containing 3800 protein sequences.

Our ability to determine the three-dimensional structures of proteins has been outstripped by our capacity to determine amino acid sequences from DNA sequences. New ways of

Common methods for detection of similarity depend on pairwise alignment of sequences—for example, the dot matrix method (9, 10) or the dynamic programming methods (11–14). These methods, however, have their disadvantages, including methods (15, 16). All of these normally test every sequence in the database independently against a single probe sequence without using information implicit in the alignments of families of related sequences or implicit information available from structural or sequence criteria. We describe the family comparison dot matrix method (9), which, however, does not allow for insertion or deletion. [Profile analysis brings in both structural and family information at the expense of a modest increase in computation time.

METHODS

Construction of the profile (PROFNAKE). Profile analysis has two steps: (i) construction of a profile with the program PROFNAKE, and (ii) comparison of the profile with a database of sequences or a single sequence (program PROFANAL). The starting point for the creation of a profile is a group of sequences of similar proteins. This profile is usually a group of typical sequences of functionally related proteins that have been aligned by similarity in sequence or three-dimensional structure. Each sequence can be given a weight, which is useful when several of them are very similar. It is also possible to make a profile from a single sequence if

Lisez l'article suivant pour plus de détails

39

Aligner à un profil 2

Utilise l'information au sein du PSSM pour calculer les scores et les pénalités (*linear gap penalty*):

$$S(i,j) = \max \left\{ \begin{array}{l} S(i-1,j-1) + PSSM(seq(i),j) \\ S(i-1,j) + PSSM(" - ", j) \\ S(i,j-1) + PSSM(" - ", j-1) \\ 0 \end{array} \right. \quad \begin{array}{l} j \text{ est une colonne du PSSM} \\ PSSM(" - ", j) \text{ est la pénalité pur un trou dans la position } j \end{array}$$

seq(i) est l'acide aminé dans la position *i*

38

PSI-BLAST

Le système PSI-BLAST utilise des PSSM pour la recherche de séquences dans les bases de données

$q = \text{AQRQRQQARQ}$

Chercher les séquences d dans la base de données D

$d_1 = \text{AQAAARRQARQ}$

Construire un PSSM utilisant les séquences d avec un score E plus petit qu'un seuil E^*

$d_2 = \text{AQQRRAAQRQ}$

Raffiner le PSSM

$d_3 = \text{QQRQRRAAQA}$

$d_4 = \text{RQQAAQQARQ}$

$d' = \text{RRRQAAQAOQ}$

Utiliser le PSSM pour l'identification des séquences apparentées

40

Aligner des Profils?

On ne peut pas aligner des profils simplement parce qu'ils enregistrent des scores et des pénalités

Mais on peut faire une comparaison entre deux profils en utilisant des corrélations entre les colonnes de deux profils comme par exemple le *Pearson correlation coefficient*.

$$r(S) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

les espaces ne sont pas acceptées

41

Aligner des groupes de séquences 2

	0	1	2	3	4	5	6	7	
	S N A	A A	- A	- L	- D G	- G	- V	- K	(C)
0	0 (3)	72 (2)	108 (2)	144 (2)	162 (2)	180 (2)	198 (2)	216 (2)	
1 AS	72 (1)	-7 (3)	65 (2)	101 (2)	119 (2)	137 (2)	155 (2)	173 (2)	
2 CA	96 (1)	41 (1)	17 (3)	77 (2)	95 (2)	113 (2)	131 (2)	149 (2)	
3 DM	108 (1)	53 (1)	43 (1)	53 (3)	77 (2)	95 (2)	113 (2)	131 (2)	
4 G-	132 (1)	77 (1)	77 (1)	54 (3)	77 (3)	107 (2)	125 (2)	143 (2)	
5 F-	144 (1)	89 (1)	89 (1)	78 (1)	79 (3)	106 (3)	126 (3)	139 (2)	
6 V-	156 (1)	101 (1)	101 (1)	90 (1)	97 (1)	98 (3)	120 (3)	144 (2)	
7 H-	180 (2)	137 (1)	133 (3)	127 (3)	123 (3)	143 (3)	137 (3)	152 (3)	

O. Gotoh (1993) Optimal alignment between groups of sequences and its application to multiple sequence alignment. CABIOS 9(3):361-370

43

Aligner des groupes de séquences

Gotoh a proposé 4 algorithmes pour trouver l'alignement optimal qui utilisent une variation de Needleman et Wunsch en utilisant la pénalité affine pour les espaces

Une évaluation des coûts d'espaces plus précise

- Algorithme A
- Algorithme B
- Algorithme C
- Algorithme D

$A = \text{ACDFVH}$ $B = \text{NALGVV}$
 SAM--- AA-G--K

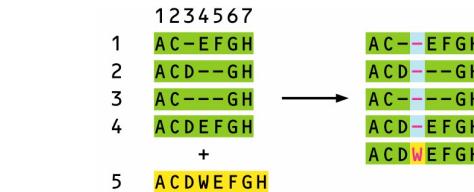
$C = \text{AC-DGFVH}$
 SA-M---
 S---G
 NALDG-V-
 AA-G--K

O. Gotoh (1993) Optimal alignment between groups of sequences and its application to multiple sequence alignment. CABIOS 9(3):361-370

42

Aligner des groupes de séquences 3

La partie la plus difficile est le calcul correct du coût d'espaces (le coût d'ouverture et le coût d'extension)



Le coût d'espace pour la séquence 4 est égal au coût d'ouverture et pour les autres il est égal au coût d'extension

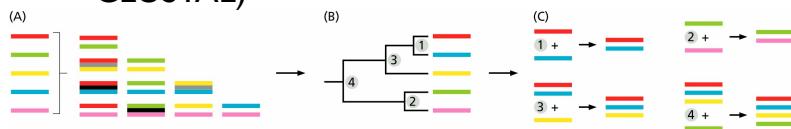
44

Les Méthodes globales 4

- La programmation dynamique optimisée (le système MSA)

Pour l'alignement de beaucoup de séquences on a besoin d'heuristiques

- L'alignement progressif (le système CLUSTAL)



45

L'alignement progressif 2

Comment calculer la matrice de distances?

Faites un alignement entre chaque paire de séquences (programmation dynamique ou une autre méthode)

Calculez la distance entre chaque alignement :

$$d_{ij} = \frac{s_{ij}}{L_{ij}}$$

s_{ij} nombre de substitutions
 L_{ij} taille de l'alignement

Les espaces ne sont pas pris en considération

La matrice est symétrique

les éléments sur la diagonale sont 0

47

L'alignement progressif

L'alignement progressif est une approche **heuristique** pour aligner plusieurs séquences

Aucune garantie qu'on trouve l'alignement optimal

3 étapes:

Calculer une matrice de distances entre les paires de séquences

Construire un arbre phylogénétique en utilisant cette matrice

Utiliser cet arbre pour aligner chacune des séquences

(cfr les étapes 1-4 de MSA)

46

L'alignement progressif 3

Comment construire l'arbre ?

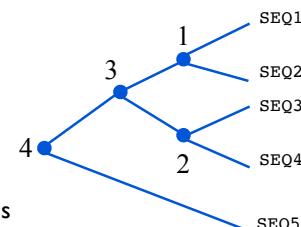
Regrouper d'abord les deux séquences les plus proches (c.-à.-d. 1)

Ensuite, regrouper :

A. les deux séquences suivantes les plus proches (c.-à.-d. 2)

B. Les deux groupes 1 et 2 (c.-à.-d. 3)

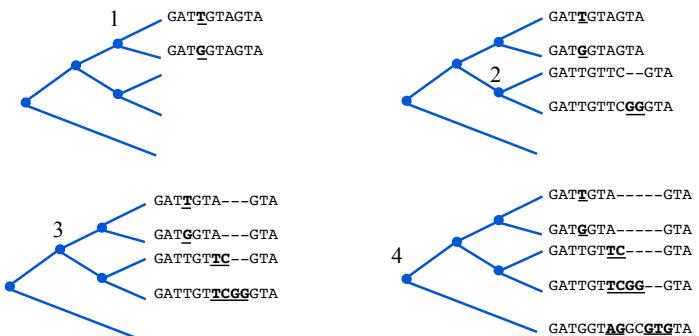
C. une séquence avec le groupe qui était construit précédemment (c.-à.-d. 4)



48

L'alignement progressif 4

L'arbre détermine l'ordre dans lequel on ajoute chaque séquence au APS (voyez algorithmes de Gotoh)



49

L'alignement progressif 6

SCORE

Alignement par paire et le calcul des scores de différences

$$D_{ij} = -\ln \frac{S_{ij} - S_{rand}}{S_{iden} - S_{rand}} \times 100 \quad S_{iden} = \frac{S_{ii} + S_{jj}}{2}$$

S_{ij} Le score d'alignement (en utilisant p.e. PAM250)

$$S_{rand} = (1/L) \sum \sum S(a,b) N_i(a) N_j(b) - N(g) g_{penalty}$$

Le score d'alignement de deux séquences aléatoires avec la même composition et la même taille

51

L'alignement progressif 5

Le système de Feng et Doolittle:

Les systèmes APS ont un souci :ils enlèvent ou changent trop les espaces qui étaient présent auparavant. Ce qui est peut-être plausible d'une perspective d'optimisation, mais pas d'une perspective biologique

“une fois un espace, toujours un espace”

le système est composé de 6 fonctions. Ici, seulement les fonctions les plus importantes sont expliquées

SCORE

BORD

DFAAlign

D.F. Feng and R.F. Doolittle (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees J Mol Evol 25:351-360

50

L'alignement progressif 7

SCORE

4 segments obtenus de 4 protéines qui font parties de la famille I-immunoglobulin

x_1 ILDDMDVVEGSAARFDCKVEGYDPPEVMWFKDDNPVKESRHFQIDYDEEGN
 x_2 RDPVKTHEGWGVMLPCNPAAHYPGLSYRWLLNEFPNFIPTDGRHFVQSQT
 x_3 ISDTEADIGSNLWRGCAAAGKPRPMVRWLNRGEPLASQNRVEVLA
 x_4 RRLIPAARGGEISILCQPRAAPKATILWSKGTEILGNSTRVTVTSD

La matrice de substitution PAM250

$$g_{penalty} = 8$$

Un alignement entre chaque paire de séquences et entre les séquences elles-mêmes est produit (Needleman et Wunsch algorithme)

52

L'alignement progressif 8

SCORE

$$S_{11}=262, S_{22}=287 \dots$$

Alignment 1 : $S_{12}=31$

x_1 ILDMVVVEGSAARFDCKVEG-YPDPEVMWFKDDNPVKESRHFQIDYDEEGN
 x_2 RDPVKTHEGWGVMLPCNPPAHPGLSYRWLLNEFPNFIPTD-GRHFVSQLT

Alignment 2 : $S_{13}=44$

x_1 ILMDVVVEGSAARFDCKVEGYPDPEVMWFKDDNPVKESRHFQIDYDEEGN
 x_3 ISDTEADIGSNLRWGCAAAGKPRPMVRWLNGEPL-ASQN-RV--EVLA-

Alignment 3 : $S_{14}=13$

x_1 ILMDVVVEGSAARFDCKVEGYPDPEVMWFKDDNPVKESRHFQIDYDEEGN
 x_4 RRLIPAARGGEISILCQPRAAPKATILWSKGTE-ILGNST-RV--TVTSD

...

53

L'alignement progressif 10

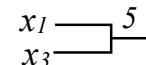
BORD

Construit un arbre préliminaire en utilisant l'algorithme proposé par Fitch et Margoliash

À chaque étape joignez les séquences ou groupes de séquences avec la plus petite distance et recalculez la distance entre ce nouveau groupe et les séquences (ou groupes) restantes.

D_{ij}	x_1	x_2	x_3	x_4
x_1	0	1.25	0.95	1.31
x_2		0	1.24	1.30
x_3			0	1.13
x_4				0

Les séquences x_1 et x_3 sont les plus proches



W.M. Fitch and E. Margoliash (1967) Construction of phylogenetic trees, Science 155(3760):279-284

55

L'alignement progressif 9

SCORE

S_{ij}	x_1	x_2	x_3	x_4
x_1	262	31	44	13
x_2		287	15	16
x_3			222	45
x_4				215

S_{rand}	x_1	x_2	x_3	x_4
x_1		-66.94	-80.28	-70.48
x_2			-82.86	-72.52
x_3				-37.85
x_4				

D_{ij}	x_1	x_2	x_3	x_4
x_1	0	1.25	0.95	1.31
x_2		0	1.24	1.30
x_3			0	1.13
x_4				0

$$D_{ij} = -\ln \frac{S_{ij} - S_{rand}}{S_{iden} - S_{rand}}$$

54

L'alignement progressif 11

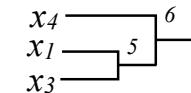
BORD

Construit un arbre préliminaire en utilisant l'algorithme proposé par Fitch et Margoliash

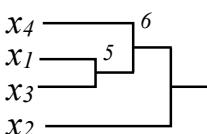
$$D_{32} = \frac{D_{12} + D_{32}}{2} = 1.245$$

$$D_{54} = \frac{D_{14} + D_{34}}{2} = 1.22$$

D_{ij}	5	x_2	x_4
5	0	1,245	1,22
x_2		0	1,3
x_4			0



D_{ij}	6	x_2
6	0	1,263
x_2		0



56

L'alignement progressif 12

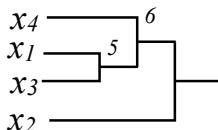
DFAAlign Utilisez l'arbre pour la construction du APS

première étape

x_1 ILDMVVVEGSAARFDCKVEGYPDPEVMWFKDDNPVKESRHFQIDYDEEGN
 x_3 ISDTEADIGSNLRWGCAGAKPRPMVRWLNGEPL-ASQN-RV--EVLA-

deuxième étape

x_1 ILDMVVVEGSAARFDCKVEGYPDPEVMWFKDDNPVKESRHFQIDYDEEGN
 x_3 ISDTEADIGSNLRWGCAGAKPRPMVRWLNGEPL-ASQN-RV--EVLA-
 x_4 RRLIPAARGGEISILCQPRAPKATILWSKGTEIL-GNST-RV--TVTSD



troisième étape

x_1 ILDMVVVEGSAARFDCKVEGYPDPEVMWFKDDNPVKESRHFQIDYDEEGN
 x_3 ISDTEADIGSNLRWGCAGAKPRPMVRWLNGEPL-ASQN-RV--EVLA-
 x_4 RRLIPAARGGEISILCQPRAPKATILWSKGTEIL-GNST-RV--TVTSD
 x_2 RDPVKTHEGWGVMLCPNPAHYPLSYRWLNEFPNPIPTD-GRHFVSQTT

Les résultats dépendent de la matrice de substitution et de la pénalité g

57

L'alignement progressif 14

Réglages de paramètres introduits par CLUSTAL W:

Des pénalités dynamiques qui changent selon le type d'acide aminé ou selon la position dans la séquence

Information concernant la probabilité de trouver un espace à côté d'un des 20 acides aminés est utilisée pour changer localement la pénalité d'ouverture

Des régions courtes de résidus hydrophiles indiquent la présence d'une boucle, exigeant la réduction de la pénalité d'ouverture

...

J.D Thompson, D.G. Higgins and T.J. Gibson (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acid Research 22:4673-4680

59

L'alignement progressif 13

Cette méthode a deux soucis :

Le problème du maximum local

Le séquences sont ajoutées sur des alignements existants

Par conséquent, chaque erreur dans l'alignement introduit des erreurs supplémentaires dans les alignements qui sont construits plus tard

Comment choisir les paramètres

Il faut choisir au moins une matrice de substitution, une pénalité d'ouverture et une pénalité d'extension

CLUSTAL W a essayé de résoudre ce problème

Ceux-ci fonctionnent bien en cas de séquences homologues, pourtant ils commencent à échouer sérieusement dès que les séquences divergent

58

L'alignement progressif 15

Réglages de paramètres introduits par CLUSTAL W:

Des matrices de substitution sont utilisées dynamiquement selon la divergence des séquences à aligner à chaque étape

Les séquences sont pesées pour corriger l'échantillonnage inégal à travers toutes les distances évolutives dans les données

Des séquences similaires sont pesées vers le bas

J.D Thompson, D.G. Higgins and T.J. Gibson (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acid Research 22:4673-4680

60

L'alignement progressif 13

Ce méthode a deux soucis :

Le problème du maximum local

Le séquences sont ajoutées sur des alignements existants

Par conséquent, chaque erreur dans l'alignement introduisent des erreurs supplémentaires dans les alignements qui sont construit plus tard

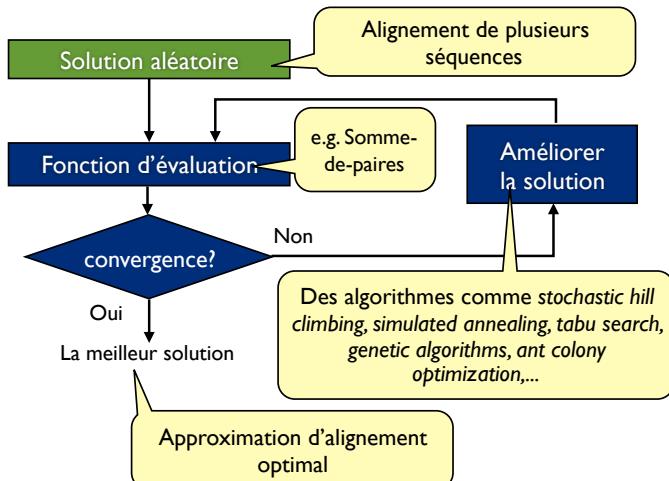
Cor

Des algorithmes stochastiques peuvent résoudre ce problème puisqu'ils peuvent s'échapper de solutions localement optimales

Ceux-ci fonctionnent bien en cas de séquences homologues, pourtant ils commencent à échouer sérieusement dès que les séquences divergeront

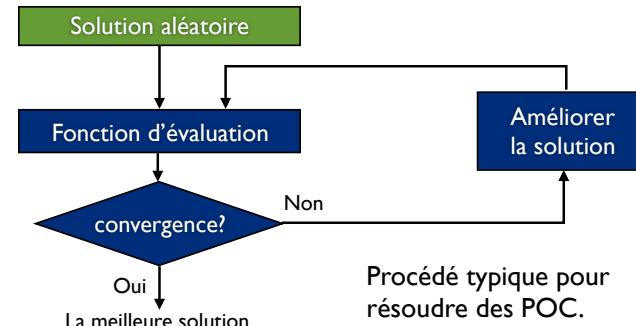
61

Amélioration itérative 2



63

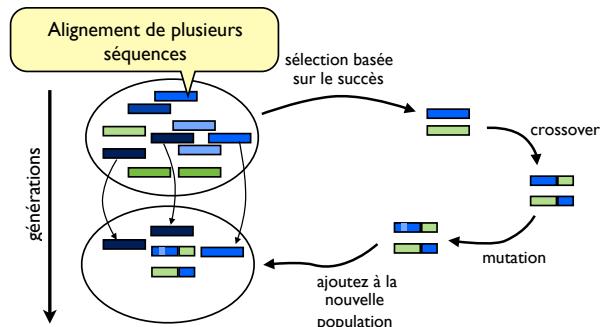
Amélioration itérative



62

Algorithmes Stochastiques

SAGA = sequence alignment by genetic algorithm

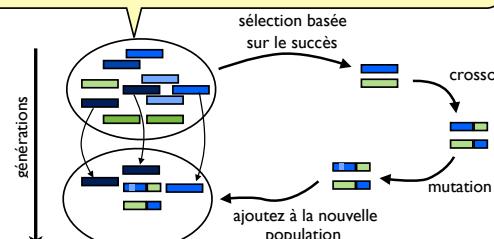


C. Notredame and D.G. Higgins (1996) SAGA: sequence alignment by genetic algorithm. Nucleic Acid Research 24:1515-1524

64

Algorithmes Stochastiques 2

Au début une population d'alignements de N séquences sans espace internes est créée (~ 100) (On ajoute des espaces à la fin des séquences pour créer des alignements de taille L) = génération 0

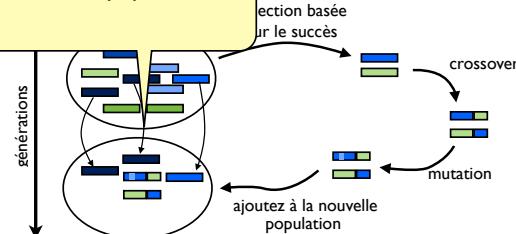


Les séquences dans l'APS peuvent être décalées vers la droite, en remplaçant les positions au début avec des espaces

65

Algorithmes Stochastiques 4

Chaque génération, 50% des meilleurs APS sont copiés dans la population suivante

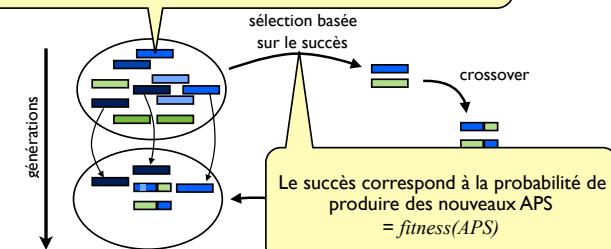


Ainsi, 50% de la population suivante est produite par les opérateurs

67

Algorithmes Stochastiques 3

La qualité d'un APS est évaluée en utilisant des fonctions: ici deux fonctions ressemblant à la somme de paires pondérées avec une pénalité d'espaces affine

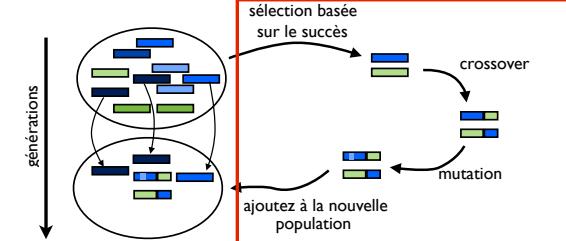


Les solutions avec un succès élevé pourraient produire entre 0 et 2 nouveaux APS

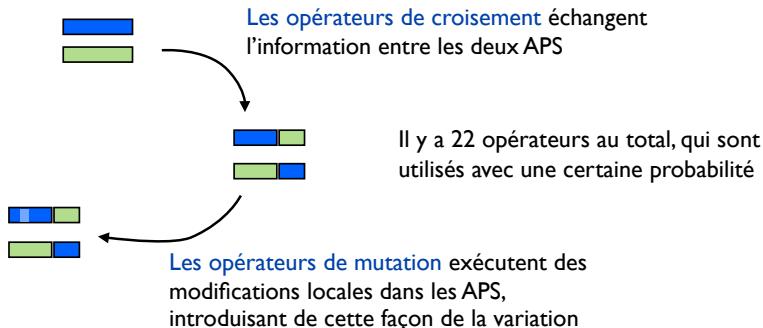
66

Algorithmes Stochastiques 5

Pendant cette étape, les meilleurs APS sont sélectionnés et des nouvelles solutions sont produites à partir d'eux



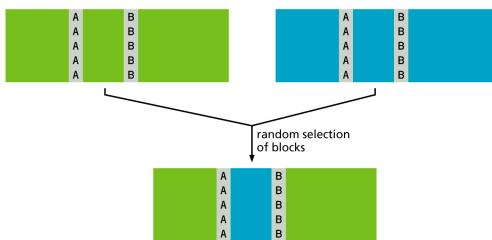
Algorithmes Stochastiques 6



69

Algorithmes Stochastiques 8

Le croisement uniforme recherche d'abord des colonnes contenant les mêmes acides aminés aux mêmes positions (colonnes consistantes)

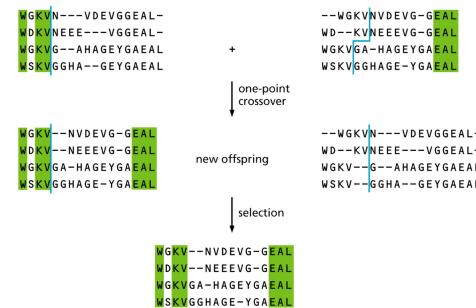


Dans le nouvel APS, ces colonnes consistantes sont préservées et les régions dans l'intervalle sont remplies avec les alignements d'un des deux APS

71

Algorithmes Stochastiques 7

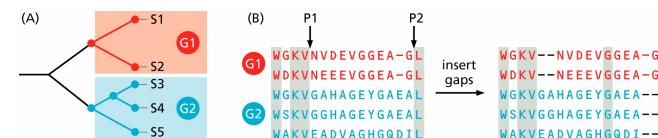
Le croisement à un point prend deux APS takes two MSA, les coupe à une certaine position, échange les deux parties et les colle ensemble



70

Algorithmes Stochastiques 9

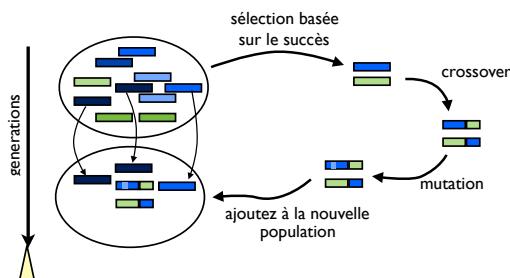
Gap-insertion est un opérateur de mutation. 1) Les séquences dans un MSA sont divisées en deux groupes (utilisant un arbre grossier) 2) Un espace avec une taille aléatoire est inséré dans le groupe G1



3) Un espace avec la même taille est inséré dans G2 dans une position à une distance limitée par la position de l'espace dans le groupe G1

72

Algorithmes Stochastiques 10

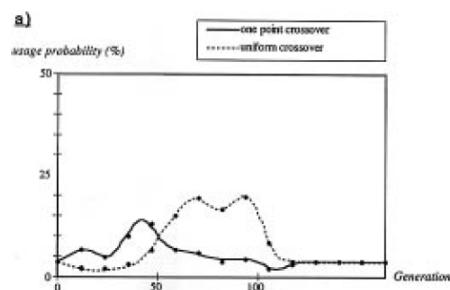


L'algorithme se termine quand les APS cessent de s'améliorer, c-à-d le succès n'augmente plus loin

73

Algorithmes Stochastiques 12

Planification dynamiques des opérateurs de croisement



75

Algorithmes Stochastiques 11

Planification dynamique des opérateurs

Au début la probabilité d'utiliser un opérateur est 1/22 (on garantit que chaque opérateur n'obtient jamais une probabilité de zéro)

Les probabilités sont adaptées en utilisant la performance de ces opérateurs durant les 10 générations précédentes

Attribution de crédit correcte

Tous les opérateur sont crédités pour la création d'un meilleur APS

Le dernier obtient 50% du crédit, l'avant-dernier obtient 50% du crédit restant (25% de l'original), etc

74

Algorithmes Stochastiques 13

SAGA comparé au système MSA (pour des petits groupes) et CLUSTAL W (pour des alignements grands)

```

1ton DMLLCAGEME-GGKDTCTCDDSGGP-LICAGG-----VLGIGTSGAT-----
2pkA ESNLTCAGTLP-GGKDTCTCDDSGGP-LICAGG-----MWGIGTSGMHT-----
2ptn DMMFCAGLRL-GGKDTCTCDDSGGP-LICAGG-----WQGIGTSGS-----
2trm DMVCUGHR-GGKDTCTCDDSGGP-LICAGG-----BLGGIGTSGY-----
4cha DMICAGO--aa-SGSVSSDGSGP-LICAGG-----GANTLGIGTSGV-----
3est NSMVCAO--gd-CVKSCQDGSGP-LICAGG-----SQUAVGIGTSGVSR-----
1hne RSMVCTLVRG-RQAQGVCFDGSGP-LICAGG-----LIMGIGTSGVRS-----
3rp2 KFOUCVQGSP-1LRAAPNDGSGP-LICAGG-----VANGIGTSGH-----
1agt NEELICAGVPDtgppV7TQGDGSGP-MERKDNADHWIOKGIGTSGM-----
2aga ssgivygmiq-tmVCAQDDSGGS-LFAGs-----TALGLTGCGS-----
3agb sgdvvygmiq-tmVCAERPDSGGP-LVAGt-----RAZGLTGCGS-----
2alp egaqv-rgltq-gnACMGRRDSGGSwitSag-----QAGGVMSGGNTVQSGN

```

SAGA fonctionne aussi bien que MSA sur les petits groupes de séquences et surpassé CLUSTAL W sur les grands groupes de séquences

76