PLaSMA, un nouvel algorithme progressif pour l'alignement multiple de séquences

V. Derrien J.M. Richer J.K. Hao

LERIA – Université d'Angers

Plan

- Introduction
- L'alignement de séquences,
  - Alignement par paire,
  - Alignement multiple,
- Principe de PLaSMA,
- Résultats,
- Conclusion perspectives.

Introduction

L'alignement de séquences est un outil fondamental pour de nombreuses applications de biologie :

- Découverte de gènes,
- Analyse phylogénétique,
- Découverte de motifs,
- …

Alignement par paire

- Similarité entre 2 séquences (matrices),
- Comparaison d'une séquence avec un ensemble de séquences,
- Algorithmes efficaces basés sur la programmation dynamique,
- Complexité en $O(m.n)$.
- Evaluation : somme des valeurs associées à chaque paire de résidus.

Alignement multiple

- Similarité d'un ensemble de K séquences (K>2),
- Problème NP-complet [Wang & Jiang 94],
- Exemple :

```
AG-CTAGC-ACTGA
---CTAGC--ATG-
AGCCTAGC---TGC
A-TC-AGCAAATGC
```

Alignement multiple

- Evaluation :
  - Somme des paires,
  - Coffee,
  - T_Coffee,
  - NorMD,
  - …
- Deux grandes approches :
  - Progressive : alignement de sous groupes de séquences (*Clustal W* [Thompson et al, 1994])
  - Itérative : alignement de toutes les séquences simultanément (*SAGA* [Notredame et al, 1996])

Principe de *Clustal W*

- Construction d'une matrice de distances à partir des alignements par paire,
- Construction d'un arbre (*guide-tree*) par la méthode du *Neighbour-Joining*, chaque nœud interne représentant un *profile*,
- Alignement progressif des séquences, basé sur la programmation dynamique : *profile*.
- Problème : perte de qualité.

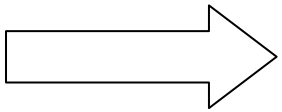
PLaSMA

- Pour réduire ce problème, PLaSMA aligne directement deux groupes de séquences alignées : notion de blocs.
- Extension de la méthode basée sur la programmation dynamique pour aligner :
  - Une séquence avec un bloc,
  - Deux blocs.
- La principale différence est dans la méthode d'alignement.

Principe de *PLaSMA*

```
ATCCAGCT
A-CGAGCT
ATCC-GC-
```

```
ACGTCGT-T
AC-TCGTC-
```



```
A--TCCAG-CT
A---CGAG-CT
A--TCC-G-C-
ACGTC--GT-T
AC-TC--GTC-
```

Algorithme de PLaSMA

Entrée : *E*, un ensemble de *n* séquences.

Sortie : *L* un alignement de *E.*

1. Construction du guide-tree de *E*
2. Construction de *n* blocs composés d'une seule séquence
3. $L=\{B_1,...,B_{|E|}\}$
4. **Tant que** $|L| > 1$ **Faire**
   - Prendre B_i et B_j les deux blocs les plus similaires (guide tree)
   - $B = \text{align}(B_i, B_j)$
   - $L = L \cup \{B\} \setminus \{B_i, B_j\}$

Balibase

- http://www-igbmc.u-strasbg.fr/Bioinfo/Balibase,
- Ensemble de 139 jeux d'essai regroupés en 5 catégories (Références),
- Mise à disposition pour chaque jeux d'essai d'un meilleur alignement,
- 2 fonctions d'évaluation sont proposées,
- Résultats pour 10 algorithmes différents progressifs et itératifs.

→ Comparaison.

Fonctions d'évaluation

Comparaison avec l'alignement de référence

– Par paires de résidus,

```
GAAWQGQIVG
EPQNDDELPM
-A-SGDNTLS
KKEREEDIDL
```

Résultat

```
GAAWQGQIVG
EPQNDDELPM
-ASGDNTLSM
KKEREEDIDL
```

Référence

Fonctions d'évaluation

Comparaison avec l'alignement de référence

- Par paires de résidus,
- Par colonnes.

```
GAAWQGQIVG
EPQNDDELPM
-A-SGDNTLS
KKEREEDIDL
```

Résultat

```
GAAWQGQIVG
EPQNDDELPM
-ASGDNTLSM
KKEREEDIDL
```

Référence

Résultats

- Nous avons testé notre algorithme sur les 139 jeux de Balibase :

	PRRP	ClustalW	SAGA	Dialign	SB_Pima	ML_Pima	Multalign	Pilup8	Multal	HMMT	**Plasma**
	Première fonction de comparaison										
Moy.	0,592	0,838	0,841	0,631	0,550	0,562	0,511	0,583			**0,799**
	Deuxième fonction de comparaison										
Ref. 1	0,877	0,864	0,841	0,788	0,821	0,810	0,834	0,832	0,673	0,487	**0,750**
Ref. 2	0,541	0,583	0,586	0,384	0,401	0,379	0,371	0,517	0,429		**0,391**
Ref. 3	0,532	0,446	0,506	0,314	0,175	0,267	0,372	0,303	0,323		**0,463**
Ref. 4	0,323	0,361	0,289	0,853	0,794	0,705	0,292	0,710			**0,442**
Ref. 5	0,700	0,705	0,642	0,836	0,508	0,572	0,627	0,639			**0,643**

- Temps de calcul très faibles.

Résultats

- PLaSMA obtient le meilleur résultat sur 7 jeux d'essais :

	PRRP	Clustal W	SAGA	Dialign	SB_Pima	ML_Pima	Multalign	Pilup8	Multal	HMMT	**Plasma**
1ubi	0,498	0,415	0,452	0,000	0,370	0,493	0,488	0,428	0,428	0,140	**0,525**
1dox	0,887	0,919	0,879	0,859	0,868	0,868	0,799	0,812	0,480	0,806	**0,963**
2mhr	0,980	0,985	0,952	0,951	0,975	0,908	0,962	0,965	0,961	0,803	**0,991**
1ezm	0,941	0,948	0,932	0,911	0,956	0,956	0,958	0,948	0,610	0,851	**0,975**
1pysA	0,931	0,922	0,906	0,936	0,935	0,935	0,926	0,931	0,917	0,616	**0,970**
1uky	0,139	0,130	0,269	0,139	0,037	0,083	0,148	0,241	0,083		**0,327**
1ajsA	0,128	0,163	0,186	0,000	0,006	0,000	0,000	0,000	0,110		**0,203**

- PLaSMA obtient un meilleur résultat que Clustal W et SAGA sur 23 jeux d'essais.

Conclusion et perspectives

- Résultats intéressants,
- Algorithme en cours de développement :
  - Redistribution des brèches dans les blocs, avec une méthode de descente,
  - Autre fonction d'évaluation (Coffee, T-Coffee)
- Post optimisation du résultat final, pour améliorer les régions de faible similarité,
- Utilisation en ligne au moyen d'une interface Web

Multiple Sequence Alignment by PLASMA

Multiple sequence alignment by PLASMA

Paste protein sequences:

```
>NPRE_BACCL
GVLGDQKYINTTYSSYYGYYYLQDNTRGSGIFTYDGRNRTVLPGSLWADGDNQFFASYDAAAVDAHYYAGVVYDYYKNVHGRL
>NPRE_BACCE
TGKGVLGDTKSLNTTLSGSSYYLQDNTRGATIFTYDAKNRSTLPGTLWADADNVFNAAYDAAAVDAHYYAGKTYDYYKATFNR
>NPRE_BACAM
TGTGTTLKGKTVSLNISSESGKYVLRDLSKPTGTQIITYDLQNREYNLPGTLVSSTTNQFTTSSQRAAVDAHYNLGKVYDYFY
>NPRE_BACPO
ATGKGVLGDSKSFTTTASGSSYQLKDTTRGNGIVTYTASNRQSIPGTLLTDADNVWNDPAGVDAHAYAAKTYDYYKSKFGRNS
>NPRE_BACBR
ATGKGVLGDTKQFETTKQGSTYMLKDTTRGKGIETYTANNRTSLPGTLMTDSDNYWTDGAAVDAHAHAQKTYDYFRNVHNRNS
```

or upload a file: Parcourir...

Parameters:

Scoring matrix: BLOSUM30

Gap opening penalty (K): -10

Gap extension penalty (H): -1

Execute Multiple Alignment | Reset

PLASMA Result

matrix = blosum45
Gap opening penalty = -10
Gap extension penalty = -1

```
PileUp


  MSF:  322  Type: P    Check:  1234   ..

Name: NPRE_BACAM oo  Len:  322  Check:  1234  Weight:  1.00
Name: NPRE_BACPO oo  Len:  322  Check:  1234  Weight:  1.00
Name: NPRE_BACBR oo  Len:  322  Check:  1234  Weight:  1.00
Name: NPRE_BACCE oo  Len:  322  Check:  1234  Weight:  1.00
Name: NPRE_BACCL oo  Len:  322  Check:  1234  Weight:  1.00

//


NPRE_BACAM      TGTGTTLKGK TVSLNISSES --GKYVLRDL SKPTGTQIIT YDLQNREYNL
NPRE_BACPO      -ATGKGVLGD SKSFTTTASG --SSYQLKDT TR--GNGIVT YTASNRQ-SI
NPRE_BACBR      -ATGKGVLGD TKQFETTKQG --STYMLKDT TR--GKGIET YTANNRT-SL
NPRE_BACCE      --TGKGVLGD TKSLNTTLSG --SSYYLQDN TR--GATIFT YDAKNRS-TL
NPRE_BACCL      -----GVLGD QKYINTTYSS YYGYYYLQDN TR--GSGIFT YDGRNRT-VL


NPRE_BACAM      PGTLVSSTTN QFTTSSQRAA VDAHYNLGKV YDYFYQKFNR NSYDNKGGKI
NPRE_BACPO      PGTLLTDADN VWN---DPAG VDAHAYAAKT YDYYKSKFGR NSIDGRGLQL
NPRE_BACBR      PGTLMTDSDN YWT---DGAA VDAHAHAQKT YDYFRNVHNR NSYDGNGAVI
NPRE_BACCE      PGTLWADADN VFNAAYDAAA VDAHYYAGKT YDYYKATFNR NSINDAGAPL
NPRE_BACCL      PGSLWADGDN QFFASYDAAA VDAHYYAGVV YDYYKNVHGR LSYDGSNAAI
```

```
NPRE_BACAM      VSSVHYGSRY NNAAWIGDQM IY--GDGDGS FFSPLSGSMD VTAHEMTHGV
NPRE_BACPO      RSTVHYGSRY NNAFWNGSQM TYGDGDGDGS TFIAFSGDPD VVGHELTHGV
NPRE_BACBR      RSTVHYSTRY NNAFWNGSQM VY--GDGDGT TFLPLSGGLD VVAHELTHAV
NPRE_BACCE      KSTVHYGSNY NNAFWNGSQM VY--GDGDGV TFTSLSGGID VIGHELTHAV
NPRE_BACCL      RSTVHYGRGY NNAFWNGSQM VY--GDGDGQ TFLPFSGGID VVGHELTHAV

NPRE_BACAM      TQETANLNYE NQPGALNESF SDVFGYFND- ----TEDWDI GEDIT---VS
NPRE_BACPO      TEYTSNLEYY GESGALNEAF SDVIGNDIQ- ----RKNWLV GDDIYTPNIC
NPRE_BACBR      TERTAGLVYQ NESGALNESM SDIFGAMVD- ----NDDWLM GEDIYTPGRS
NPRE_BACCE      TENSSNLIYQ NESGALNEAI SDIFGTLVEF YDNRNPDWEI GEDIYTPGKA
NPRE_BACCL      TDYTAGLVYQ NESGAINEAM SDIFGTLVEF YANRNPDWEI GEDIYTPGVA

NPRE_BACAM      QPALRSLSNP TKYGQPDNF- KNYKNLPNTD AGDYGGVHTN SGIPNKAAY-
NPRE_BACPO      GDALRSMSNP TLYDQPHHYS NLY-----KG SSDNGGVHTN SGIINKAYYL
NPRE_BACBR      GDALRSLQDP AAYGDPDHYS KRY-----TG SQDNGGVHTN SGINNKAAYL
NPRE_BACCE      GDALRSMSDP TKYGDPDHYS KRY-----TG SSDNGGVHTN SGIINKQAYL
NPRE_BACCL      GDALRSMSDP AKYGDPDHYS KRY-----TG TQDNGGVHTN SGIINKAAYL

NPRE_BACAM      ---------N TITKIGVNKA EQIYYRALTV YLTPSSTFKD AKAALIQSAR
NPRE_BACPO      LAQGGTFHGV TVNGIGRDAA VQIYYSAFTN YLTSSSDFSN ARAAVIQAAK
NPRE_BACBR      LAEGGTHYGV RVNGIGRTDT AKIYYHALTH YLTPYSNFSA MRRAAVLSAT
NPRE_BACCE      LANGGTHYGV TVTGIGKDKL GAIYYRANTQ YFTQSTTFSQ ARAGAVQAAA
NPRE_BACCL      LSQGGVHYGV SVTGIGRDKM GKIFYRALVY YLTPTSNFSQ LRAACVQAAA

NPRE_BACAM      DLYG--SQDA ASVEAAWNAV GL
NPRE_BACPO      DLYGANSAEA TAAAKSFDAV G-
NPRE_BACBR      DLFGANSRQV QAVNAAYDAV GV
NPRE_BACCE      DLYGANSAEV AAVKQSFSAV GV
NPRE_BACCL      DLYGSTSQEV NSVKQAFNAV GV
```