Mini projet 2 ; La construction des matrices de substitution

Professeur Tom Lenaerts
Assistant Catharina Olsen et Elisa Cilia

Information additionnelle sur :
http://www.ulb.ac.be/di/map/tlenaert/Home_Tom_Lenaerts/INFO-F-208.html

Le but du mini projet est de créer des matrices de substitution spécifiquement construites pour des familles de protéines en utilisant l'information dans la base de données BLOCKS (http://blocks.fhcrc.org/). Les familles qu'on utilisera sont les familles des domaines SH2 et les tyrosine kinases.

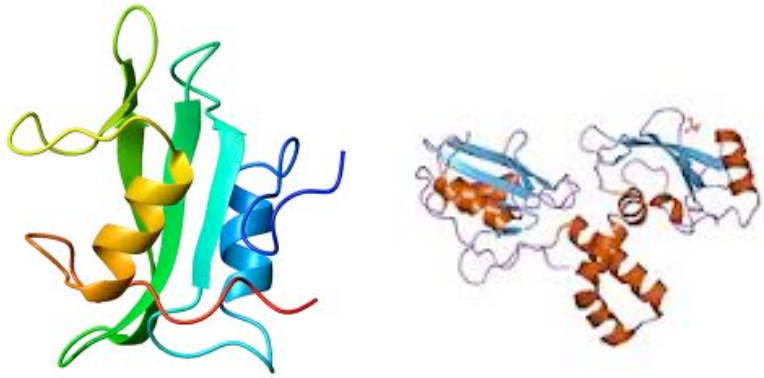


Figure 1 : La structure d'un membre de chaque famille. La première montre un domaine SH2 et la dernière est le kinase.

Pour leur construction, vous utiliserez l'approche BLOSUM comme expliqué dans le cours (diapositives de L4 : pages 32-47).

Faites attention que pour chaque famille il y a plusieurs BLOCK (5 pour la famille SH2 par exemple). Pour les valeurs $f_{a,b}$ il faut calculer d'abord les fréquence pour chaque BLOCK indépendamment. Après le $f_{a,b}$ total pour tous les BLOCK ensemble est obtenu en faisant la somme normalisée des ces $f_{a,b}$ par BLOCK.

Pour chaque famille, vous créerez 2 matrices qui sont générées en utilisant des groupements différents : c.-à-d. 70% et 40% d'identité entre les séquences qui font partie du même groupe.

Quand les matrices sont créées, vous expliquez une fois chaque étape de la méthode BLOSUM en utilisant une de ces trois familles comme exemple. Donnez la possibilité de télécharger les matrices de votre wiki. Examiner aussi la similarité de vos matrices avec la matrice BLOSUM62.

Montrez aussi quelques exemples d'alignement pour des séquences de la même famille (en utilisant le logiciel que vous avez implémenté dans le premier mini

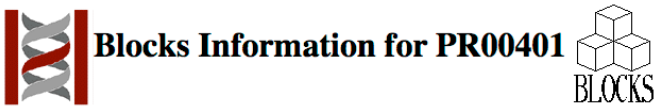
projet). Est-ce qu'il y aura des différences entre les alignements quand vous utiliserez des matrices de 70% ou 40% ?

Comparez aussi vos résultats avec les alignements pour les mêmes séquences en utilisant par exemple BLOSUM62. Est-ce que les alignements obtenus en utilisant les matrices que vous avez construites sont meilleurs?

Les données

Les BLOCKS pour les trois familles peuvent être trouvés sur le site de BLOCKS.
SH2: http://blocks.fhcrc.org/blocks-bin/getblock.pl?PR00401
Tyrosine kinases: http://blocks.fhcrc.org/blocks-bin/getblock.pl?PR00109

Pour la famille SH2 vous obtenez la page suivante, qui commence avec une petite table de contenu ou menu sur l'information disponible sur ce page :



PR00401: SH2DOMAIN

SH2 domain signature

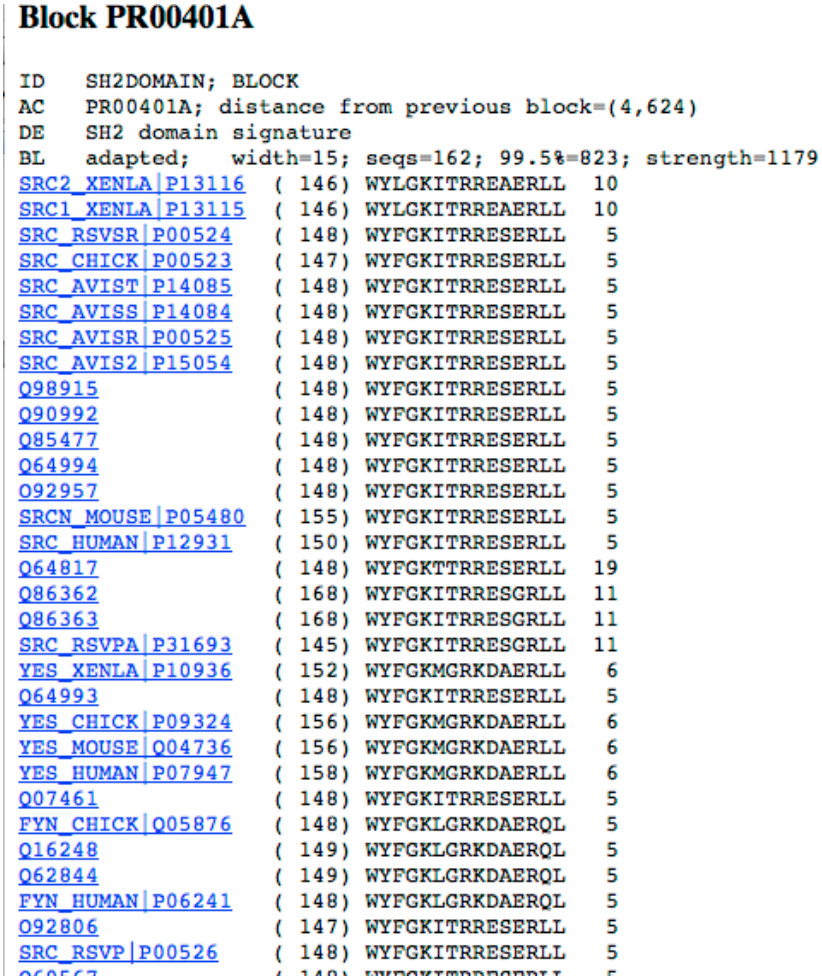
- Introduction
- Block number PR00401A
- Block number PR00401B
- Block number PR00401C
- Block number PR00401D
- Block number PR00401E

- PRINTS Entry PR00401 (source of blocks)
- Protein Sequences Used to Make Blocks.[Sequences in fasta format]
- Block Maps.[Graphical Map] [Text Map] [Map Positions] [About Maps]
- Logos.[About Logos]
  Select display format: [GIF] [PDF] [Postscript]
- Tree from blocks alignment. [About Trees] [About ProWeb TreeViewer]
  [Data] [TreeViewer] [XBitmap] [GIF] [PDF] [Postscript]
- Structures
- Search blocks vs other databases:
  - COBBLER sequence and BLAST searches [About COBBLER]
  - MAST Search of all blocks vs a sequence database [About MAST]
  - LAMA search of all blocks vs a blocks database [About LAMA]
- CODEHOP to design PCR primers from blocks [About CODEHOP]
- SIFT to predict amino acid substitutions in blocks [About SIFT]
- Re-format blocks as a multiple alignment

```
        Prints Database 35 in Blocks Format, July 2002
     Made available by the Fred Hutchinson Cancer Research Center
   1100 Fairview AV N, A1-162, PO Box 19024, Seattle, WA 98109-1024

Based on PRINTS Database as described by TK Attwood, et al (1994),
NAR 22(17):3590-3596. ID is from PRINTS gc line, AC is from
PRINTS gx line, DE is from PRINTS gt line, BL is BLOCK information.
```

Cette page montre qu'il y a 5 blocks conservés dans les séquences de la famille SH2: les blocks A-e. L'information dans chaque BLOCK est montrée après ce menu. Par exemple pour le premier BLOCK on voit (seulement les premières lignes) :



Block PR00401A

```
ID   SH2DOMAIN; BLOCK
AC   PR00401A; distance from previous block=(4,624)
DE   SH2 domain signature
BL   adapted;   width=15; seqs=162; 99.5%=823; strength=1179
SRC2_XENLA|P13116  ( 146) WYLGKITRREAERLL  10
SRC1_XENLA|P13115  ( 146) WYLGKITRREAERLL  10
SRC_RSVSR|P00524   ( 148) WYFGKITRRESERLL   5
SRC_CHICK|P00523   ( 147) WYFGKITRRESERLL   5
SRC_AVIST|P14085   ( 148) WYFGKITRRESERLL   5
SRC_AVISS|P14084   ( 148) WYFGKITRRESERLL   5
SRC_AVISR|P00525   ( 148) WYFGKITRRESERLL   5
SRC_AVIS2|P15054   ( 148) WYFGKITRRESERLL   5
Q98915             ( 148) WYFGKITRRESERLL   5
Q90992             ( 148) WYFGKITRRESERLL   5
Q85477             ( 148) WYFGKITRRESERLL   5
Q64994             ( 148) WYFGKITRRESERLL   5
O92957             ( 148) WYFGKITRRESERLL   5
SRCN_MOUSE|P05480  ( 155) WYFGKITRRESERLL   5
SRC_HUMAN|P12931   ( 150) WYFGKITRRESERLL   5
Q64817             ( 148) WYFGKTTRRESERLL  19
Q86362             ( 168) WYFGKITRRESGRLL  11
Q86363             ( 168) WYFGKITRRESGRLL  11
SRC_RSVPA|P31693   ( 145) WYFGKITRRESGRLL  11
YES_XENLA|P10936   ( 152) WYFGKMGRKDAERLL   6
Q64993             ( 148) WYFGKITRRESERLL   5
YES_CHICK|P09324   ( 156) WYFGKMGRKDAERLL   6
YES_MOUSE|Q04736   ( 156) WYFGKMGRKDAERLL   6
YES_HUMAN|P07947   ( 158) WYFGKMGRKDAERLL   6
Q07461             ( 148) WYFGKITRRESERLL   5
FYN_CHICK|Q05876   ( 148) WYFGKLGRKDAERQL   5
Q16248             ( 149) WYFGKLGRKDAERQL   5
Q62844             ( 149) WYFGKLGRKDAERQL   5
FYN_HUMAN|P06241   ( 148) WYFGKLGRKDAERQL   5
O92806             ( 147) WYFGKITRRESERLL   5
SRC_RSVP|P00526    ( 148) WYFGKITRRESERLL   5
```

Il y a donc 162 séquences dans ce BLOCK et chaque séquence a une taille de 15 acides aminés.

Pour obtenir chaque BLOCK vous devez télécharger les séquences du site. Dans le menu il y a une ligne avec le texte « *Re-format blocks as a multiple alignment* ». Appuyez « *Re-format* » et vous arrivez au page suivant :

Re-format Blocks as an Alignment

You can make blocks from unaligned protein sequences with Block Maker.

Enter your Blocks in BLOCKS format:

```
ID   SH2DOMAIN; BLOCK
AC   PR00401A; distance from previous block=(4,624)
DE   SH2 domain signature
BL   adapted;   width=15; seqs=162; 99.5%=823; strength=1179
SRC2_XENLA|P13116  ( 146) WYLGKITRREAERLL  10
SRC1_XENLA|P13115  ( 146) WYLGKITRREAERLL  10
SRC_RSVSR|P00524   ( 148) WYFGKITRRESERLL  5
SRC_CHICK|P00523   ( 147) WYFGKITRRESERLL  5
SRC_AVIST|P14085   ( 148) WYFGKITRRESERLL  5
SRC_AVISS|P14084   ( 148) WYFGKITRRESERLL  5
SRC_AVISR|P00525   ( 148) WYFGKITRRESERLL  5
SRC_AVIS2|P15054   ( 148) WYFGKITRRESERLL  5
Q98915             ( 148) WYFGKITRRESERLL  5
Q90992             ( 148) WYFGKITRRESERLL  5
Q85477             ( 148) WYFGKITRRESERLL  5
Q64994             ( 148) WYFGKITRRESERLL  5
O92957             ( 148) WYFGKITRRESERLL  5
SRCN_MOUSE|P05480  ( 155) WYFGKITRRESERLL  5
SRC_HUMAN|P12931   ( 150) WYFGKITRRESERLL  5
Q64817             ( 148) WYFGKTTRRESERLL  19
```

Select an output alignment format [Fasta]

Re-format Reset

Blocks home

Contact us

Page last modified on August 2003

Le plus simple est de reformater les données en format FASTA. Donc sélectionnez l'option « *Fasta* » dans « *select and output alignment format* » et appuyez laprès e bouton « *Re-format* ». Cela vous donne la page suivante sur laquelle on peut voir pour chaque protéine les quatre BLOCK.

Re-format Blocks as Alignment

```
>ABL1_CAEEL|P03949|179              from    PR00401 blocks
WYHGKISRSDSEAIL
TGSFLVRESET
IGQYTISVRHDG
RVFHYRINVDN
KFRTLGELVHHHSVH
>ABL1_HUMAN|P00519|127             from    PR00401 blocks
WYHGPVSRNAAEYLL
NGSFLVRESES
PGQRSISLRYEG
RVYHYRINTAS
RFNTLAELVHHHSTV
>ABL2_HUMAN|P42684|173             from    PR00401 blocks
WYHGPVSRSAAEYLL
NGSFLVRESES
PGQLSISLRYEG
RVYHYRINTTA
RFSTLAELVHHHSTV
>ABL_DROME|P00522|271              from    PR00401 blocks
WYHGPISRNAAEYLL
NGSFLVRESES
PGQRSISLRYEG
RVYHYRISEDP
KFNTLAELVHHHSVP
>ABL_FSVHY|P10447|76              from    PR00401 blocks
WYHGPVSRNAAEYLL
NGSFLVRESES
PGQRSISLRYEG
RVYHYRINTAS
RFNTLAELVHHHSTV
>ABL_MLVAB|P00521|13              from    PR00401 blocks
WYHGPVSRNAAEYLL
NGSFLVRESES
PGQRSISLRYEG
RVYHYRINTAS
RFNTLAELVHHHSTV
>ABL_MOUSE|P00520|127              from    PR00401 blocks
WYHGPVSRNAAEYLL
NGSFLVRESES
PGQRSISLRYEG
RVYHYRINTAS
RFNTLAELVHHHSTV
>BLK_HUMAN|P51451|123              from    PR00401 blocks
```

Copiez-collez ou sauvegardez les données (sans le titre) vers un fichier texte qui pourrait être utilisé dans votre logiciel.

La seule chose que vous devez faire avant de démarrer avec la construction des matrices est de regrouper chaque BLOCK dans un fichier indépendant.