

Mini projet 2 : La construction des matrices de substitution

Professeur Tom Lenaerts
Assistant : Catharina Olsen

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 SH3 et PDZ.

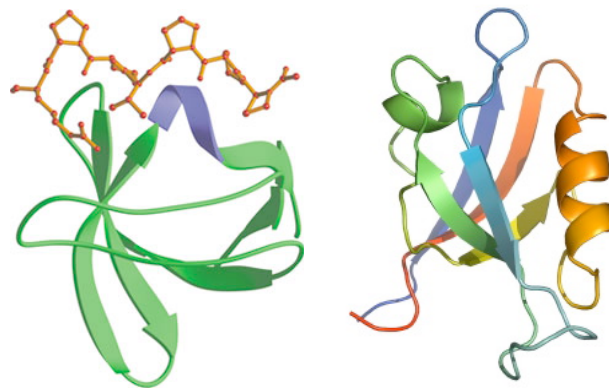


Figure 1 : La structure d'un membre de chaque famille. La première montre un domaine SH3 et la dernière un membre de la famille PDZ.

Pour leur construction, vous utiliserez l'approche BLOSUM comme expliqué pendant le cours (diapositives de L4 : pages 33-48).

Faites attention que pour chaque famille il y a plusieurs BLOCK (4 pour la famille SH3 par exemple). Les valeurs $f_{a,b}$ sont calculées sur les 4 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 deux familles comme exemple. Donnez la possibilité de télécharger les matrices de votre wiki. Examinez aussi la similarité de vos matrices avec la matrice BLOSUM62. Est-ce que les valeurs sur le diagonal sont différent ? Est-ce que certaines substitutions sont maintenant accepté qui n'étaient pas accepté en BLOSUM62 (ou vice versa) ?

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 deux familles peuvent être trouvés sur le site de BLOCKS.

Pour la famille SH3 : <http://blocks.fhcrc.org/blocks-bin/getblock.pl?PR00452>

Pour la famille PDZ : <http://blocks.fhcrc.org/blocks-bin/getblock.pl?PR00401>

Pour la famille SH3 vous obtenez la page suivante, qui commence avec une petite table de contenu ou menu sur l'information disponible sur ce page :



PR00452: SH3DOMAIN

SH3 domain signature

- [Introduction](#)
- [Block number PR00452A](#)
- [Block number PR00452B](#)
- [Block number PR00452C](#)
- [Block number PR00452D](#)
- PRINTS Entry [PR00452](#) (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

Cette page montre qu'il y a 4 blocks conservés dans les séquences de la famille SH3: les blocks A-D. L'information dans chaque BLOCK est montrée après

ce menu. Par exemple pour le premier BLOCK on voit (seulement les premières lignes) :

Block PR00452A

```
ID  SH3DOMAIN; BLOCK
AC  PR00452A; distance from previous block=(1,6613)
DE  SH3 domain signature
BL  adapted; width=11; segs=257; 99.5%=716; strength=1083
FYN_HUMAN|P06241 ( 84) TLFVALYDYEA 3
Q62844 ( 85) TLFVALYDYEA 3
Q16248 ( 85) TLFVALYDYEA 3
FYN_MOUSE|P39688 ( 84) TLFVALYDYEA 3
FYN_XENLA|P13406 ( 84) TLFVALYDYEA 3
YES_CHICK|P09324 ( 92) TVFVALYDYEA 3
SRC1_XENLA|P13115 ( 82) TTFVALYDYES 4
FYN_CHICK|Q05876 ( 84) TLFEALYDYEA 6
Q85466 ( 368) TVFVALYDYEA 3
YES_AVISY|P00527 ( 84) TVFVALYDYEA 3
FYN_XIPHE|P27446 ( 84) TLFVALYDYEA 3
YES_XIPHE|P27447 ( 95) TFFVALYDYEA 5
YRK_CHICK|Q02977 ( 83) TLFIALYDYEA 4
YES_HUMAN|P07947 ( 94) TIFVALYDYEA 3
YES_XENLA|P10936 ( 88) TVFVALYDYEA 3
SRC2_XENLA|P13116 ( 82) TTFVALYDYES 4
YES_MOUSE|Q04736 ( 92) TIFVALYDYEA 3
SRC_RSVF|P00526 ( 84) TTFVALYDYES 4
SRC_RSVH1|P25020 ( 84) TTFVALYDYES 4
SRC_HUMAN|P12931 ( 86) TTFVALYDYES 4
SRC_CHICK|P00523 ( 83) TTFVALYDYES 4
SRC_AVIST|P14085 ( 84) TTFVALYDYES 4
SRC_AVISS|P14084 ( 84) TTFVALYDYES 4
SRC_AVISR|P00525 ( 84) TTFVALYDYES 4
SRC_AVIS2|P15054 ( 84) TTFVALYDYES 4
SRCN_MOUSE|P05480 ( 85) TTFVALYDYES 4
Q98915 ( 84) TTFVALYDYES 4
Q90993 ( 84) TTFVALYDYES 4
Q90992 ( 84) TTFVALYDYES 4
Q64817 ( 84) TTFVALYDYES 4
```

Il y a donc 257 séquences dans ce BLOCK et chaque séquence a une taille de 11 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 SH3DOMAIN; BLOCK
AC PR00452A; distance from previous block=(1,6613)
DE SH3 domain signature
BL adapted; width=11; seqs=257; 99.5%=716; strength=1083
FYN_HUMAN|P06241 ( 85) TLFVALYDYEA 3
Q62844 ( 85) TLFVALYDYEA 3
Q16248 ( 85) TLFVALYDYEA 3
FYN_MOUSE|P39688 ( 84) TLFVALYDYEA 3
FYN_XENLA|P13406 ( 84) TLFVALYDYEA 3
YES_CHICK|P09324 ( 92) TVFVALYDYEA 3
SRC1_XENLA|P13115 ( 82) TTFVALYDYES 4
FYN_CHICK|Q05876 ( 84) TLFVALYDYEA 6
Q85466 (368) TVFVALYDYEA 3
YES_AVISY|P00527 ( 84) TVFVALYDYEA 3
FYN_XIPHE|P27446 ( 84) TLFVALYDYEA 3
YES_XIPHE|P27447 ( 95) TFFVALYDYEA 5
YRK_CHICK|Q02977 ( 83) TLFVALYDYEA 4
YES_HUMAN|P07947 ( 94) TIFVALYDYEA 3
YES_XENLA|P10936 ( 88) TVFVALYDYEA 3
SRC2_XENLA|P13116 ( 82) TTFVALYDYES 4
```

Select an output alignment format:

Fasta

Re-format

Reset

Le plus simple est de reformater les données en format FASTA. Donc dans l'option indiquée avec le carré rouge vous sélectionnez l'option « *Fasta* » et vous appuyez le 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

```
>ABP1_SACEX|P38479|560 from PR00452 blocks
PWATAEYDYEA
NELTFAENDKIINIEF
DDDWLGELE
QKGLFPSNYVVLG
>ABP1_YEAST|P15891|535 from PR00452 blocks
PWATAEYDYDA
NELTFVENDKIINIEF
DDDWLGELE
SKGLFPSNYVSLG
>AMPH_CHICK|P50478|612 from PR00452 blocks
FKVEVLHDFEA
DELNLKRGDIVLVIPS
EAGWLTGIKE
YKGLFPENFTRHL
>AMPH_HUMAN|P49418|625 from PR00452 blocks
YKVETLHDFEA
DELTQLRGDVVLVVPS
DAGWLGVKE
YKGLFPENFTRRL
>ANM2_HUMAN|P55345|33 from PR00452 blocks
EEFVAIADYAA
TQLSFLRGEKILILRQ
TADWWWGERA
CCGYIPANHVKGK
>BLK_HUMAN|P51451|60 from PR00452 blocks
HFVVALYDYTA
RDLQMLKGEKLQVLKG
TGDWWLARSL
REGYVPSNFVARV
>BLK_MOUSE|P16277|54 from PR00452 blocks
RFVVALFDYAA
RDLQVLKGEKLQVLRS
TGDWWLARSL
REGYVPSNFVAPV
>BTK_HUMAN|Q06187|217 from PR00452 blocks
KKVVALYDYMP
NDLQLRKGDYFIEE
NLPWWRARDK
```

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.