Mini-projet 4 ; Prédiction de la structure secondaire

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 de ce projet est d'implémenter l'algorithme GOR III et de faire quelques testes avec votre propre implémentation. Les détails expliquant l'implémentation de cet algorithme se trouvent dans l'article de Jean Garnier et al (voir le PDF sur le site). L'ensemble de cet article et les slides donneront l'information nécessaire pour implémenter votre propre version de GOR III; Les équations 8 et 9 dans l'article expliquent comment l'implémenter.

Après avoir implémenté GOR III en utilisant les données d'entrainement (voire en bas), vous devez tester votre algorithme. Dans le fichier compressé `datasets.zip` nous avons aussi fournis un fichier `CATH_info_test.txt`. Dans ce fichier vous pouvez retrouver cinq noms des séquences pour tester votre prédicteur. Utilisez ces cinq exemples pour tester la qualité de vos prédictions.

1. Comparez vos prédictions avec les résultats attendus. Où sont les similarités et les différences ?
2. Quels sont les scores Q3 et MCC pour vos prédictions ?

Expliquez sur votre wiki comment vous avez construit le parser (regardez en bas) et le prédicteur. Montrez et expliquez en détail vos résultats.

Les données d'entrainement

L'algorithme GOR utilise des informations concernant la probabilité de trouver les acides aminés dans une hélice-α, un brin-β, une β-boucle et des bobines (coils) pour prédire la structure secondaire d'une protéine. Pour déterminer ces probabilités vous avez besoin des données, qui sont fournies dans le fichier `datasets.zip`.

Le répertoire `dssp` contient une grande collection des structures des protéines (tirées de l'ensemble de WHATIF) et les informations structurelles secondaires. Ces informations étaient produites par l'outil DSSP.

Au-dessous une petite partie des données pour `1A58.dssp` comme il apparaît dans le répertoire `dssp`.

```
  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    LADDERS PER SHEET             .
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
    1    1 A M              0   0  193      0, 0.0     2,-0.1     0, 0.0    29,-0.0   0.000 360.0 360.0 360.0 -46.9   62.1   21.2   10.1
    2    2 A S    >   -     0   0   50      1,-0.1     3,-1.3    27,-0.0    28,-0.2  -0.451 360.0-113.9 -94.2 167.5   61.2   20.4    6.5
    3    3 A K  G >  S+     0   0  146      1,-0.3     3,-1.0     2,-0.2    26,-0.1   0.717 118.1  64.3 -70.6 -20.0   58.3   21.9    4.4
    4    4 A K  G 3  S+     0   0  184      1,-0.2    -1,-0.3    26,-0.0     0, 0.0   0.476  86.4  73.9 -80.9  -2.4   56.9   18.4    4.4
    5    5 A D  G <  +      0   0   87     -3,-1.3     2,-0.2     2,-0.1    -1,-0.2   0.544  69.2 119.3 -84.7 -11.7   56.5   18.7    8.2
    6    6 A R    <   -     0   0   53     -3,-1.0    22,-0.2    -4,-0.1     2,-0.1  -0.411  52.5-150.8 -65.0 122.8   53.5   21.1    7.8
    7    7 A R  E     -A   27   0A 108     20,-0.7    20,-3.1    -2,-0.2     2,-0.4  -0.468   3.7-134.6 -91.9 160.5   50.3   19.8    9.3
    8    8 A R  E     -AB  26 176A  97    168,-0.7   168,-2.9    18,-0.2     2,-0.3  -0.963  20.4-175.7-119.9 139.2   46.7   20.5    8.2
    9    9 A V  E     -AB  25 175A   0     16,-2.5    16,-2.5    -2,-0.4     2,-0.3  -0.914   8.7-146.9-132.4 153.7   43.8   21.4   10.5
   10   10 A F  E     -AB  24 174A  29    164,-2.8   164,-1.5    -2,-0.3     2,-0.4  -0.917  10.6-167.2-128.0 155.4   40.1   22.0   10.0
   11   11 A L  E     -AB  23 173A   0     12,-1.8    12,-2.9    -2,-0.3     2,-0.7  -0.976   9.4-158.6-137.7 116.7   37.2   24.1   11.4
```

```
  12   12 A D  E     -AB  22 172A  16    160,-3.0   159,-1.9    -2,-0.4   160,-1.1  -0.897  23.6-161.8 -95.0 118.7   33.6   23.3   10.5
  13   13 A V  E     -AB  21 170A   0      8,-2.7     7,-2.8    -2,-0.7     8,-1.3  -0.843  14.0-166.0-112.1 140.4   31.6   26.5   11.1
```

La troisième, la quatrième et cinquième colonne contient les données pertinentes, c'est-à-dire respectivement l'identifiant de la chaîne, l'acide aminé et la structure secondaire à laquelle l'acide aminé appartient. Donc par exemple le résidu 9 est un Valine (V) qui est situé sur la chaîne A et appartient à un brin (E) dans la structure de protéine. Si il n'y a pas de l'information dans cette colonne, il n'y pas une structure secondaire pour cet élément. La classe de cet élément est bobine (ou coil).

Il y a huit symboles pour les structures secondaires dans ce fichier DSSP qui peuvent être réduites à quatre catégories/classes :

1. Les symboles H, G et I correspondent une classe d'hélice (H)
2. Le symbole E correspond à la classe de ß-reliure (E)
3. Le symbole T correspond à la classe de ß-tour (T)
4. Les symboles C, S et B(=espace) correspondent à la classe de bobine aléatoire (C)

Donc la première étape du projet sera d'implémenter un parser qui peut lire ces fichiers et qui peut collecter l'information concernant les probabilités qu'une certaine acide aminé appartient à une certaine classe (H, E, T ou C). Les noms de tous les fichiers DSSP sont enregistrés dans le fichier `CATH_info.txt`. Le plus simple est de donner ce fichier comme input à votre parser pour collectionner les données dans le répertoire `dssp`.

ATTENTION ; Il peut y exister plusieurs chaînes (copies de la même séquence) dans le même fichier DSSP. Le nome de la chaine est indiquer par les symboles dans la troisième colonne du fichier dssp (voire l'exemple `1A58.dssp` plus haute). On n'utilise pas tous les chaines. Dans le fichier `CATH_info.txt` on n'a pas seulement mis le nom du fichier qu'on peut retrouver dans le répertoire `dssp`. Pour chaque fichier on a aussi ajouté quelle chaines vous devez utiliser. Au-dessous vous voyez de certaines des entrées de ce fichier:

3NIRA	3A38A	2VB1A	1US0A	1R6JA
2DSXA	1UCSA	1P9GA	2WFIA	1GCIA
2H5CA	3MFJA	2JFRA	1PQ7A	…

Chaque entrée contient un identifiant PDB (les quatre premiers caractères) suivi par un identifiant de chaîne (le cinquième caractère). Par exemple, une des structures de protéine a été utilisée dans l'analyse est la chaîne `A` de `3NIR`. Cela signifie que vous devez seulement utiliser la chaîne A dans le fichier `dssp/3NIR.dssp`.

Donc, pour chaque entrée dans le fichier `CATH_info.txt` vous devez obtenir la séquence protéique et pour chaque position dans cette séquence l'élément secondaire. Cela vous donne un fichier avec le format suivant ;

```
> identifier|protein name|organism
```

```
MTAEPSIVARSNFNVCRLPGTPEAICATYTGSIIIPGATSPGDYAN
CCEECCCCHHHHHHHHHCCCCCHHHHHHHHCCEECCCCCCCHHHCC
> …
```

Ce fichier sera utilisé pour calculer les probabilités qui seront à leur tour utiliser pour l'implémentation de l'algorithme GOR III.

Les données de teste

Voyez le fichier `CATH_info_test.txt`. Les ordres d'acide aminé et la correspondance d'annotations de structure secondaires attendues peuvent être déterminés de la même manière comme avant. Notez que ces données de test ne font pas partie des données d'entrainement.