Practical course 6
Synthesis of the practical courses

A] Monellin and brazzein are sweet-tasting proteins (see http://pdb101.rcsb.org/motm/199). Their PDB code are 3MON and 2BRZ, respectively.

1) Analyze the quality of the experimental structure of monellin. The PDB file of monellin contains four dimers (chains AB, CD, EF and GH). Select one of the dimer for the task A.2, on the basis of their experimental quality. Justify your selection.

On peut voir via le lien suivant (https://www.rcsb.org/structure/3MON) que la méthode expérimentale qui a permis d'obtenir la structure de la monelline est la cristallographie aux rayons X. La résolution est de 2,8 angströms associée à une R-Value Work de 0.193. On dit en général que la résolution est bonne en dessous de 2,5 angströms, on peut donc la qualifier de moyenne ici. Quant à la R-Value, elle décrit la différence entre les observations expérimentales et les valeurs idéales calculées à partir du modèle cristallographique. Le minimum de la R-value est 0 correspondant à un accord parfait entre les valeurs prédites à partir d'un modèle et les valeurs observées expérimentalement. Pour les grosses molécules, la R-value se situe entre 0,2 et 0,6 et elle est inférieure à 0,2 pour les plus petites molécules. La R-value est donc correcte ici. Toutefois, bien qu'elle ait une signification, celle-ci peut être améliorée en utilisant les résultats expérimentaux et le modèle construit afin d'améliorer celui-ci. C'est pour cette raison qu'il est préférable d'y associer la R-Value Free qui est calculée avec la même formule mais sur un petit échantillon stochastique de données mis de côté dans ce but et jamais inclus dans le raffinement que subit la R-value work. Mais on n'a pas accès à cette valeur dans ce cas.
On peut ensuite creuser les investigation grâce au « Full Validation Report » (https://files.rcsb.org/pub/pdb/validation_reports/mo/3mon/3mon_full_validation.pdf) afin de choisir le meilleur dimère en terme de qualité. On peut par exemple comparer la qualité des chaînes entre elles :

Mol	Chain	Length	Quality of chain
1	A	44	61% 32% 5%
1	C	44	61% 32% 7%
1	E	44	61% 30% 7%
1	G	44	61% 27% 7% 5%
2	B	50	44% 40% 14%
2	D	50	62% 30% 8%
2	F	50	78% 16%
2	H	50	62% 26% 12%

On voit que les chaînes C et D ne contiennent pas de valeurs aberrantes (en rouge). Concernant les angles de torsion du squelette protéique, on peut également noter - selon le rapport - qu'un pourcentage de valeurs aberrantes dans le diagramme de Ramachandran apparaît pour les chaînes A, B et H. Pour ces raisons, j'ai choisi de superposer le dimère CD avec la brazzéine dans la question suivante.

2) Superimpose the 3D structure of brazzein to the two chains of monellin selected in section A.1. Analyze the results in detail (statistical significance of the superimpositions, value of the rmsd's, analyse of the superimpositions, analyze of the superimposed regions, ...). Remark: keep in mind that it is possible to superimpose a structure to all the chains of another structure individually or as a whole.

Afin de superposer la brazzéine avec les chaines C et D de la monelline, on utilise PDBeFold (http://www.ebi.ac.uk/msd-srv/ssm/ssmstart.html). On obtient ceci :

##	Scoring			RMSD	N_{align}	N_g	$\%_{seq}$	Query					
	Q	P	Z					Ch	N_{res}	$\%_{sse}$	Match	$\%_{sse}$	N_{res}
1	0.14	0.9	3.0	1.06	19	1	11	C	44	67	2brz.pdb:A	50	54
2	0.096	2.3	4.2	1.50	18	1	6	D	50	33	2brz.pdb:A	25	54

On a donc pour la chaîne C : Z-score = 3 Q-score = 0,14 P-score = 0,9 et RMSD = 1,06
Pour la chaîne D, on a : Z-score = 4,2 Q-score = 0,096 P-score = 2,3 et RMSD = 1,50

Le Q-score représente la qualité de l'alignement, c'est à dire que la RMSD et la longueur de l'alignement sont pris en compte. Sa valeur atteint 1 seulement dans le cas de structures identiques. On voit ici que les Q-scores sont faibles du fait surtout de la longeur d'alignement faible.

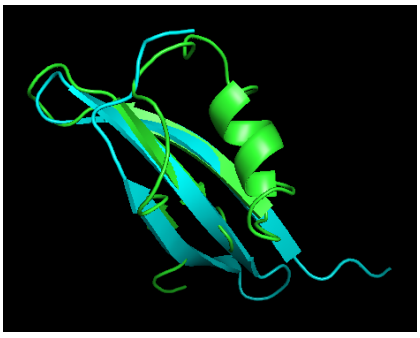
P-score = -log(P-value) sachant que la P-value mesure la probabilité d'arriver à la même qualité de match par chance. Plus grand est le P-score, plus le match est significatif. Les P-scores inférieurs à 3 indiquent des matchs insignifiants. On est donc dans ce cas là pour nos deux chaînes, les matchs sont pas vraiment significatifs.

Le Z-score mesure la signification statistique d'un match en terme de statistiques Gaussiennes. Plus le Z-score est haut, plus le match est significatif.

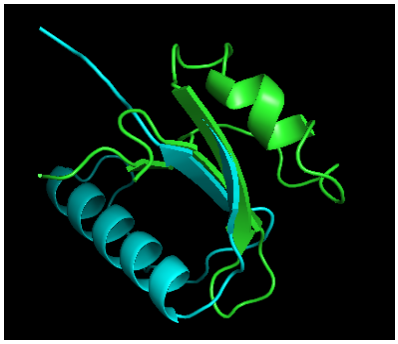
On peut également noter qu'on a un pourcentage d'identité plus grand entre C et la brazzéine qu'entre D et la Brazzéine (11 contre 6).

La RMSD (root-mean-square deviation) mesure la distance en angströms entre les atomes de carbone alpha qui correspondent entre eux/se superposent. Donc plus la RMSD est grande, plus les structures correspondantes sont distantes. Selon ce critère, on a donc une correspondance plus proche en terme de distance entre la chaîne C et la brazzéine qu'entre la chaîne D et la brazzéine.

on peut visualiser les superpositions via Pymol :



monelline_C vs brazzéine



monelline_D vs brazzéine

On remarque que dans les deux cas, ce sont les feuillets bêta qui se superposent. On ne peut évidement pas faire d'alignement global puisqu'une des deux protéines n'est alignée structurellement qu'a une partie de l'autre protéine. Si on regarde dans le détail le rapport PDBeFold, on voit qu'il y a deux régions de superposition entre la chaîne C et la brazzéine qui s'étendent sur 9 et 10 acides aminés respectivement, une avec des distances de superposition entre 0,55Å et 2,33Å et une autre avec des distances de superposition entre 0,15Å et 1,56Å. Cette seconde région est fusionnée, c'est d'ailleurs visible sur la visualisation avec Pymol. En ce qui concerne la chaîne D, on observe deux régions de superposition avec la brazzéine qui s'étendent sur 9 acides aminés toutes les deux, une avec des distances de superposition entre 0,41Å et 1,82Å et une autre avec des distances de superposition plus importantes entre 0,55Å et 2,84Å.

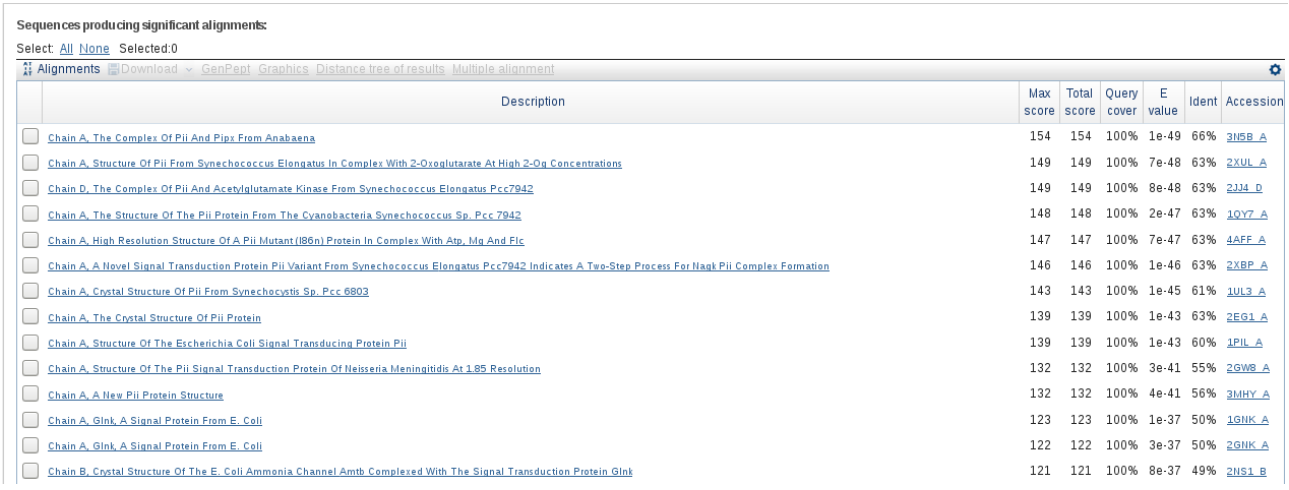
B] You will work in this section on the nitrogen regulatory protein P-II from Porphyra purpurea (Uniprot code P51254; http://www.uniprot.org).

1) Model the 3D structure of this protein by comparative modelling and analyze the quality of your model. Which tool and template did you use?

A partir d'uniprot, on obtient le fichier fasta de P51254 :

```
>sp|P51254|GLNB_PORPU Nitrogen regulatory protein P-II OS=Porphyra purpurea OX=2787 GN=glnB PE=3 SV=1
MKKIEAIIRPFKLNEVKLALVKGGIGGMTVVKVSGFGRQKGQTERYKGSEYSIDIIDKIK
IEIIVSDDKVNSITEIIIKTAKTGEIGDGKIFISDVEQVIRIRTNDLNSAAL
```

On réalise un BLASTp contre la PDB. On obtient ceci :



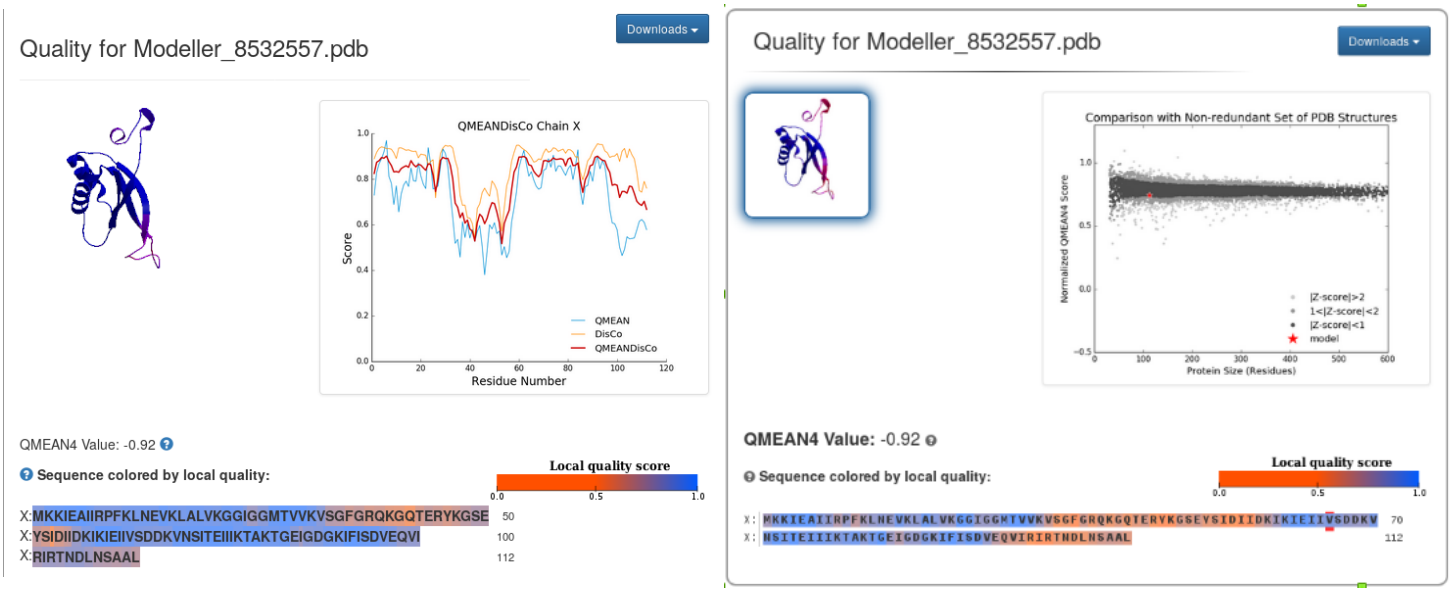
Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

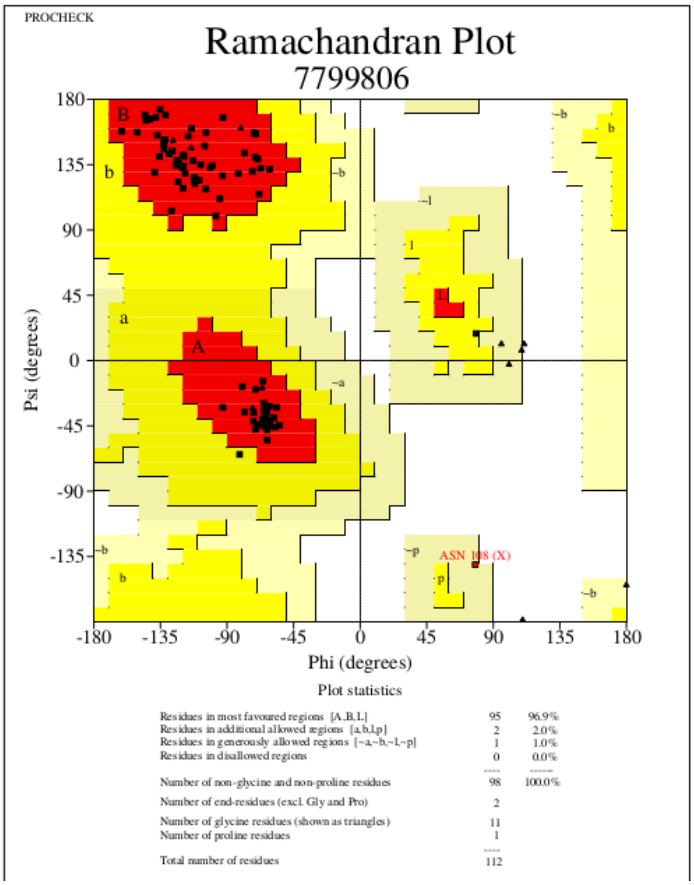
Description	Max score	Total score	Query cover	E value	Ident	Accession
Chain A, The Complex Of Pii And Pipx From Anabaena	154	154	100%	1e-49	66%	3N5B_A
Chain A, Structure Of Pii From Synechococcus Elongatus In Complex With 2-Oxoglutarate At High 2-Og Concentrations	149	149	100%	7e-48	63%	2XUL_A
Chain D, The Complex Of Pii And Acetylglutamate Kinase From Synechococcus Elongatus Pcc7942	149	149	100%	8e-48	63%	2JJ4_D
Chain A, The Structure Of The Pii Protein From The Cyanobacteria Synechococcus Sp. Pcc 7942	148	148	100%	2e-47	63%	1QY7_A
Chain A, High Resolution Structure Of A Pii Mutant (I86n) Protein In Complex With Atp, Mg And Flc	147	147	100%	7e-47	63%	4AFF_A
Chain A, A Novel Signal Transduction Protein Pii Variant From Synechococcus Elongatus Pcc7942 Indicates A Two-Step Process For Nagk Pii Complex Formation	146	146	100%	1e-46	63%	2XBP_A
Chain A, Crystal Structure Of Pii From Synechocystis Sp. Pcc 6803	143	143	100%	1e-45	61%	1UL3_A
Chain A, The Crystal Structure Of Pii Protein	139	139	100%	1e-43	63%	2EG1_A
Chain A, Structure Of The Escherichia Coli Signal Transducing Protein Pii	139	139	100%	1e-43	60%	1PIL_A
Chain A, Structure Of The Pii Signal Transduction Protein Of Neisseria Meningitidis At 1.85 Resolution	132	132	100%	3e-41	55%	2GW8_A
Chain A, A New Pii Protein Structure	132	132	100%	4e-41	56%	3MHY_A
Chain A, Glnk, A Signal Protein From E. Coli	123	123	100%	1e-37	50%	1GNK_A
Chain A, Glnk, A Signal Protein From E. Coli	122	122	100%	3e-37	50%	2GNK_A
Chain B, Crystal Structure Of The E. Coli Ammonia Channel Amtb Complexed With The Signal Transduction Protein Glnk	121	121	100%	8e-37	49%	2NS1_B

On sélectionne le premier modèle 3N5B_A car il a le meilleur pourcentage d'identité. La cover est similaire donc non déterminante dans le choix. On utilise ensuite le programme d'alignement global Stretcher (https://www.ebi.ac.uk/Tools/psa/emboss_stretcher/) afin d'aligner 3N5B_A avec P51254, on convertit le format Markx3 en format PIR et on soumet ce fichier au serveur disposant du programme Modeller (https://toolkit.tuebingen.mpg.de/#/tools/modeller). On obtient ainsi le modèle suivant dont on analyse la qualité grâce à deux outils : QMEAN (Qualitative Model Energy Analysis) et Procheck.



QMEAN est une fonction de notation globale qui analyse la qualité d'un modèle, elle reflète la fiabilité d'un modèle et sa valeur est entre 0 et 1, 1 étant le meilleur score pour un acide aminé considéré. Mais lorsqu'on calcule le QMEAN global pour toute la séquence, il est transformé en Z-score et ce Z-score doit être le plus petit possible car cela signifie qu'on obtient un score pour notre protéine proche du score obtenu pour des protéines de haute qualité, QMEAN est en quelques sortes comme un écart type entre le score de notre protéine et le score de protéines obtenues via la database qui sont de haute qualité et qui ont une longueur similaire à celle de notre protéine.
Ici Qmean= -0.92, ce n'est pas mauvais toutefois, pour les résidus 30 à 60, le QMEAN descend jusqu'à 0,4 ce qui n'est pas très bon puisque le score idéal est de 1.

Via PROCHECK, on obtient le plot du diagramme de Ramachandran contenant les résidus (points noirs).



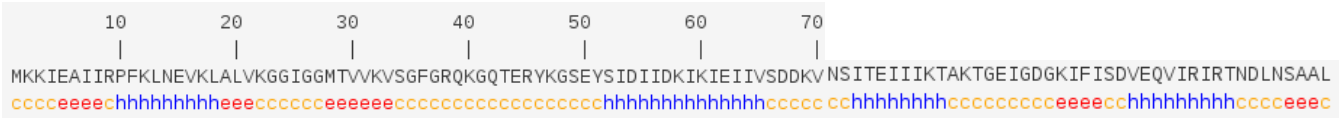
On constate que les acides aminés sont pour 96,9 % d'entre eux présents dans les régions favorisées, cela signifie que le modèle est assez bon concernant les angles de torsion.

2) Predict the secondary structure of this protein. Which tool(s) did you use and why?
What are your results: give the limits of the predicted alpha helices and beta strands?
Discuss the results.

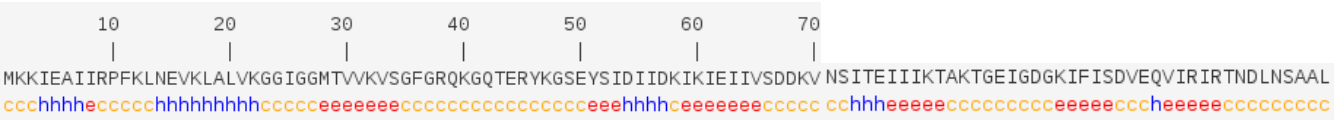
Afin de prédire la structure secondaire de la protéine, nous allons utiliser 3 outils différents :
- GOR IV : c'est une méthode statistique (Bayésienne) de prédiction sur une fenêtre de 17 acides aminés suivant une base de données de protéines résolue par cristallographie aux rayons X ;
- HNN : méthode qui combine deux réseaux neuronaux : Séquence/Structure et Structure/structure ;
- SOPMA : méthode de prédiction de la structure secondaire par homologie associée à un apprentissage automatisé.

```
Alpha helix      (Hh)
3₁₀ helix        (Gg)
Pi helix         (Ii)
Beta bridge      (Bb)
Extended strand  (Ee)
Beta turn        (Tt)
Bend region      (Ss)
Random coil      (Cc)
Ambiguous states (?)
Other states
```

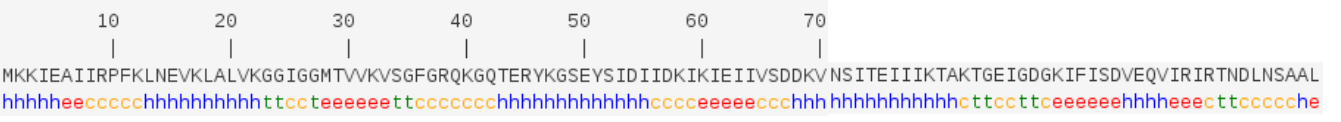
GOR IV donne ceci :



HNN donne cela :



SOPMA donne ceci :



On peut remarquer que chaque modèle donne une prédiction de structure secondaire différente. Cependant, dans certains cas, on peut voir que les feuillets bêta suivent directement les hélice alpha, ce qui n'est pas concevable. De même, des hélice ou feuillets bêta de 2 ou 3 acides aminés ne sont également pas possible. De plus, ces modèles de prédiction prennent en compte les informations locales c'est à dire l'influence des acides aminés voisins et non les interactions/influences plus éloignées. Dans certains cas, les interactions locales sont contraintes par les interactions 3D mais ce n'est pas pris en compte. Les modèles de prédictions de structures secondaires ne donneront donc jamais plus de 65% de performance.

Je propose une sorte de consensus des 3 prédictions précédentes :

```
          10        20        30        40        50        60        70        80        90       100       110
          |         |         |         |         |         |         |         |         |         |         |
MKKIEAIIRPFKLNEVKLALVKGGIGGMTVVKVSGFGRQKGQTERYKGSEYSIDIIDKIKIEIIVSDDKVNSITEIIIKTAKTGEIGDGKIFISDVEQVIRIRTNDLNSAAL
hhhhh     hhhhhhhhhh      eeeeee          hhhhhhhhhhhhh     eeeeee      hhhhhhhh      eeeeee    eeee
1-5       13-22           28-33           43-55             59-65       73-80           90-95     100-103
```

3) The STRIDE server (http://webclu.bio.wzw.tum.de/cgi-bin/stride/stridecgi.py) has been developed to assign (not predict) secondary structures from a PDB file (experimental or modelled structure). Use this tool to assign the secondary structure of the models obtained in section B.1. Compare the limits of the secondary structure of your 3D model to the secondary structure prediction of section B.2 and discuss your result.

Via le serveur STRIDE, nous obtenons le résultat suivant auquel j'ai associé en dessous la séquence consensus des 3 prédictions:

```
MKKIEAIIRPFKLNEVKLALVKGGIGGMTVVKVSGFGRQKGQTERYKGSE YSIDIIDKIKIEIIVSDDKVNSITEIIIKTAKTGEIGDGKIFISDVEQVI RIRTNDLNSAAL
 EEEEEEE GGGHHHHHHHHHH      EEEEEEE       EEEETTTTE EEETTEEEEEEEEEE GGGHHHHHHHHHHH    TTTT EEEEEE     B TTTTT BGGGG

          10        20        30        40        50        60        70        80        90       100       110
          |         |         |         |         |         |         |         |         |         |         |
MKKIEAIIRPFKLNEVKLALVKGGIGGMTVVKVSGFGRQKGQTERYKGSEYSIDIIDKIKIEIIVSDDKVNSITEIIIKTAKTGEIGDGKIFISDVEQVIRIRTNDLNSAAL
hhhhh     hhhhhhhhhh      eeeeee          hhhhhhhhhhhhh     eeeeee      hhhhhhhh      eeeeee    eeee
1-5       13-22           28-33           43-55             59-65       73-80           90-95     100-103
error     ok              ok              error             ok          ok              ok        error
```

On constate qu'il y a certaines correspondances (signalées par ok), des erreurs (error) et évidemment, les limites des hélices ainsi que des feuillets bêta ne sont pas les mêmes sauf pour quelques rares cas. Chaque méthode a des biais, à la fois notre modèle mais aussi les méthodes de prédiction de structures secondaires, c'est ainsi que l'on peut expliquer de telles différences.