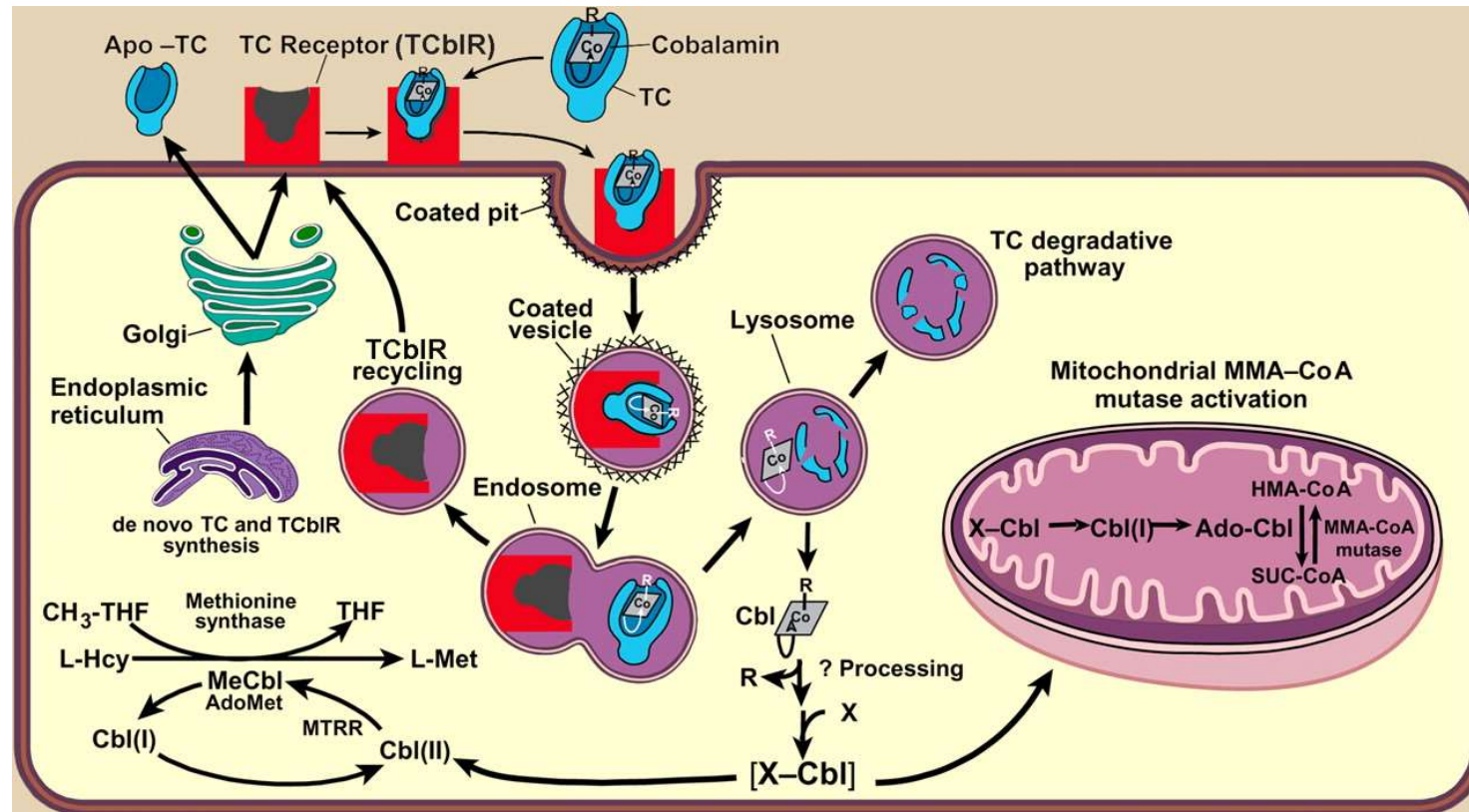


# Annotation de protéines

Nous allons étudier les protéines suivantes :

## La TransCobalamine-2 et son Recepteur



```
>sp|Q9NPF0.1| transcobalamin-2 receptor (TCbIR) CD320 gene [Homo sapiens]
MSGGWMAQVGAWRTGALGLALLLLGLGLGLEAAASPLSTPTSAQAAGPSSGSCPTKFQCRTSGLCVPL
TWRCDRDLDCSDGSDDEECRIEPTQKGQCPPLGLPCPTGVSDCSGGTDKKLRNCSRLACLAGELRCT
LSDDCIPLTWRCDGHPDCPDSSDELGCGTNEILPEGDATTMGPPVTLESVTLRNATTMGPPVTLESVPS
VGNATSSSAGDQSGSPTAYGVIAAAVLSASLVATLLLLSWLRAQERLRPLGLLVAMKESLLLSEQKTS
LP
```

```
>NP_000346.2 transcobalamin-2 isoform 1 precursor [Homo sapiens]
MRHLGAFLLGLVGLALTEMCEIPEMDSLVEKLGQHLLPWMDRLSLEHLNPSIYVGLRLSSLQAGTKED
```

```
LYLHSLKLGYYQCLLGSAFSEDDGDCQKGPSMGQLALYLLALRANCEFVRGHKGDRLVSQKWFLEDEKR  
AIGHDHKGHPHTSYQYGLGILALCLHQKRVHDSVVDKLLYAVEPFHQGHHSVDTAAMAGLAFTCLKRSN  
FNPGRQRITMAIRTVREEILKAQTPEGHFGNVYSTPLALQFLMTSPMRGAELGTACKARVALLASLQD  
GAFQNALMISQLLPVLNHKTYIDLIFPDCLAPRVMLEPAAETIPQTQEIIISVTLQVLSLLPPYRQSIISVL  
AGSTVEDVLKKAHELGGFTYETQASLSGPYLTSVMGKAAGEREFWQLLRDPNTPLLQGIADYRPKDGETI  
ELRLVSW
```

```
>NP_987376 imidazole glycerol phosphate synthase subunit HisH [Methanococcus maripaludis]
```

```
MIAIIDYNAGNLRISIEKALELYTKNIVVTSDPETILSADKLVLPGVGNFGDSMKNISQKTGDCSLNEIIN  
KCVQKVPFLGICLGMQLLLEKSEECPETPGLGVIKGDVIKFKHSEKIPHMGWNTVNQVQDIPLFEGIANN  
EYFYFVHSYHVNPSKDVISGTTNYGYEFPICILNKKNVYATQFHPEKSGKNGLKMIENFVEIL
```

## Etude de la fonction.

### Par recherche de motifs ou domaines protéiques.

Nous allons aller plus loin à l'aide de la prédiction de domaines.

La banque [InterPro](#) regroupe les données de nombreuses banques de motifs, domaines et familles protéiques. Une entrée concernant, par exemple, un motif contient une description de la fonction biologique de ce motif ; ainsi que des liens vers les entrées des autres banques qui contiennent ce motif. Lorsque l'on soumet une séquence protéique à InterProScan, un logiciel recherche tous les domaines (et autres) connus présents sur cette protéine.

#### Question 1

Est-ce que toutes les entrées trouvées par InterPro vous semblent pertinentes ?

Qu'elle semble être la fonction de la protéine (notez le numéro EC s'il est donné) ?

Est-ce que vous avez des informations sur la localisation des domaines protéiques ?

Est-ce que les commentaires des entrées InterPro vous donnent une idée de la localisation cellulaire de la protéine ?

Les entrées InterPro contiennent des liens vers les termes GO associés aux fonctions décrites.

Consultez ces termes GO pour en savoir plus sur la protéine que nous étudions.

## Localisation cellulaire.

En combinant plusieurs sources d'informations, il est possible d'avoir une idée de la localisation cellulaire d'une protéine, ou, au moins, de prédire les endroits où une protéine ne se trouve pas.

### Recherche de domaines transmembranaires.

Utilisez le logiciel [DAS](#), [TMPRED](#), ainsi que [TM-HMM](#) pour voir si la protéine contient des domaines transmembranaires.

[résultat TM-HMM pour sp|Q9NPF0.1|, résultat TM-HMM pour NP\_000346.2, résultat TM-HMM pour cette protéine]

#### Question 2

Est-ce que la protéine peut être une protéine membranaire ?

Quelle est la fiabilité des résultats ?

Le cas échéant, notez les bornes des domaines transmembranaires prédits.

#### Recherche de signaux d'adressage.

SignalP permet la recherche de peptides d'adressage vers la membrane externe et pour la sécrétion.

[résultat SignalP pour sp|Q9NPF0.1|, résultat SignalP pour NP\_000346.2]

#### Question 3

Est-ce que la protéine est sécrétée ?

Quelle est la position du site de coupure, s'il y en a un ?

Est-ce que les résultats sont cohérents avec la fonction prédite de la protéine ?

Confrontez les résultats obtenus avec TopPred et ceux obtenus avec SignalP. Quelle est votre conclusion ?

#### Prédiction de la localisation cellulaire

Le logiciel TargetP et le site [psort.org](http://psort.org) donnent la possibilité de réaliser une prédiction de localisation cellulaire, en précisant souvent le type d'organisme. Choisissez le ou les logiciels, ainsi que les paramètres adaptés à la séquence (Wolf Psort est aussi [accessible ici](#)).

[résultat TargetP pour sp|Q9NPF0.1|, résultat TargetP pour NP\_000346.2]

#### Question 4

Qu'elle semble être la localisation cellulaire de la protéine ?

Est-ce cohérent avec leur fonction et les résultats précédents ?

### Etude des structures 2D et 3D

#### PDB : la banque de structures 3D déterminées expérimentalement.

Les structures 3D déterminées expérimentalement ne représentent qu'une petite partie des protéines dont la séquence est connue. Lorsque l'on étudie une protéine, on peut rechercher si une protéine de la même famille a sa structure déterminée expérimentalement. Pour cela, on peut interroger directement la banque PDB avec le terme "*Transcobalamin-2*" ou "*imidazole glycerol phosphate*"

### Question 5

Combien de structures sont trouvées ?

Consultez la liste de ces structures, est-ce qu'elles correspondent toutes à la fonction que l'on avait identifiée ?

Consulter l'entrée qui porte le numéro d'accension PDB : *2BB5 (transcobalamin receptor + Cobalamin B12)*, *1K9V (imidazole glycerol phosphate)*.

Il y a un onglet en haut de la page qui porte la mention "sequence details". Il permet d'accéder à la structure 2D de la protéine.

### Prédiction de la structure 2D.

Pour prédire la structure 2D de votre séquence, vous pouvez utiliser le [calcul de consensus](#) proposé au PBIL ou le logiciel [PSIPRED](#).

### Question 6

Est-ce que la structure 2D prédite diffère beaucoup de la structure 2D déterminée expérimentalement ?

## Etude des réseaux cellulaires.

La banque [KEGG](#) contient de nombreux schémas de réseaux cellulaires (voies métaboliques, réseaux de régulation, ...). KEGG contient aussi d'autres banques, comme celle qui est centrée sur les génomes. Cliquez sur le lien "KEGG Organisms" pour obtenir la liste des organismes dont les génomes et autres données sont présents dans KEGG. Pour interroger les données d'un organisme donné, il faut cliquer sur son nom. Dans le formulaire, vous pouvez saisir un numéro EC, si vous l'avez noté, ou une fonction protéique.

[\[exemple\]](#)

Vous obtenez alors une liste de protéines provenant de l'organisme interrogé et correspondant à votre requête. Consultez la page de description de la protéine qui vous semble correspondre à celle que nous avons étudiée.

Sur la page de description des protéines, il y a des liens potentiels vers les schémas des voies auxquelles elles participent (ligne "Pathway" : présent pour "transcobalamin-2"). Vous pouvez les visualiser.

[\[exemple\]](#)

A partir du schéma de voie, il est possible d'accéder à la table d'orthologues (lien "Pathway entry" en haut de la page puis bouton "Ortholog table"). Cette table contient une ligne par organisme dont on connaît les gènes qui codent les protéines de la voie. Les colonnes correspondent aux fonctions protéiques de la voie. De plus, les gènes qui sont organisés en opérons sont représentés avec la même

couleur de fond. Ce tableau met en évidence la relation entre l'enchainement des gènes sur le génome et les fonctions enzymatiques.

[\[exemple\]](#)

### Autres outils.

Il existe d'autres outils pour étudier les protéines. Je vous indique ici l'adresse de deux sites qui répertorient des liens vers de nombreux logiciels d'étude de protéines. Je vous laisse les explorer selon vos besoins :

[Proteomics tools](#) d'ExPASy et [NPSA](#) du PBIL (Pole BioInformatique Lyonnais).

Page réalisée par les membres de l'équipe [bonsai](#), mise à jour en novembre 2013