Introduction à la bioinformatique (UE SSV3U15)
TP7. Réseaux et systèmes biologiques
Diaporama d'accompagnement du TP

Conception: Andreas Zanzoni
Révision : Jacques van Helden, Emese Meglecz

Objectifs et exemples traités

Objectifs

- Mettre en pratique les concepts sous-jacents à l'analyse des réseaux biologiques
- Apprendre à utiliser la base de données STRING-DB et les outils d'analyse associés pour explorer des réseaux d'interactions autour d'une protéine d'intérêt

Exemples traités durant le TP

1. Analyse fonctionnelle du sous-réseau d'interaction de la protéine CDC15 dans l'interactome de Saccharomyces cerevisiae
2. Impact du réseau d'interaction de la protéine humaine PAX6 sur la santé
3. Identifications des variants de PAX6 associés aux maladies des yeux.

Analyse fonctionnelle du sous-réseau d'interaction de la protéine CDC15 dans l'interactome de *Saccharomyces cerevisiae*

La levure du boulanger, *Saccharomyces cerevisiae*, a servi d'organisme modèle pour l'étude des mécanismes du cycle cellulaire, et a par ailleurs servi pour établir les approches de génomique fonctionnelle (transcriptome, localisation cellulaire des protéines) et d'interactome (notamment avec la méthode des doubles hybrides). Pour se familiariser avec STRING-DB, nous nous intéresserons à la protéine CDC15 qui joue un rôle important à la sortie de métaphase.

Impact du réseau d'interaction de la protéine humaine PAX6 sur la santé

Comme nous l'avons vu précédemment, le gène PAX6 code pour un facteur transcriptionnel dont l'expression, précisément contrôlée durant le développement embryonnaire, détermine la formation de l'oeil chez les animaux. Nous collecterons les protéines interagissant directement avec PAX6 et analyserons leur lien avec les pathologies humaines.

Connaissances et compétences acquises au cours de ce TP

Notions mises en pratique

- Graphes mathématiques et réseaux biologiques
- Propriétés topologiques des graphes / réseaux: degré, chemin, distance, centralité, coefficient de regroupement, sous-réseaux, voisinage… (définis ci-après)
- Enrichissement fonctionnel (défini ci-après)

N'oubliez pas que vous pouvez à tout moment consulter le **glossaire du cours** pour obtenir une définition sommaire des principaux termes utilisés.

Compétences

A l'issue de ce TP, vous devriez avoir acquis les compétences suivantes.

- Consulter une base de données d'interactions biomoléculaires pour extraire des informations sur le sous-réseau d'interactions d'une protéine d'intérêt
- Adapter les paramètres d'affichage pour faire ressortir différents types d'information
- Interpréter des indicateurs de propriétés topologique des réseaux.
- Analyser l'enrichissement fonctionnel d'un sous-réseau d'interactions.

Déroulé du TP

Etapes

- Définition des concepts (rappels et compléments du CM)
- Tutoriel : Prise en main de la base de données STRING
- Exercice 1 : voisinage de CDC15 dans l'interactome de *Saccharomyces cerevisiae*
- Exercice 2 : voisinage de la protéine PAX6 dans l'interactome humain
- Exercice 3 : Identification des variants de PAX6 associés aux maladies des yeux

Complétion

- Tous les exercices doivent être réalisés par chaque étudiant.
- En principe, l'ensemble des exercices devraient être complétés en séance (avec explications par les enseignants). Si nécessaire, ils pourront être terminés ultérieurement.

Ressources bioinformatiques utilisées

Nom	URL	Description
STRING-DB	https://string-db.org/	Base de données d'interactions physiques ou fonctionnelles entre protéines, ou entre les gènes codant pour ces protéines.
UniProtKB	https://www.uniprot.org/	Principale base de données mondiale de séquences protéiques et d'informations fonctionnelles

Définition des concepts (rappels et compléments du CM)

Pour répondre à la question que vous allez vous / nous poser

- Les **notions** complémentaires présentées pendant les TP **font partie de la matière d'examen**.
- Les **formules mathématiques ne font pas partie de la matière d'examen**. Nous les fournissons par souci de précision, et pour que vous puissiez y revenir si vous êtes un jour amenés à approfondir ces matières, mais nous ne vous demandons ni de les retenir ni de pouvoir en expliquer les détails.
- Nous vous demandons de bien comprendre les concepts, et de pouvoir les appliquer sur des exemples précis, mais pas de connaître leur formulation mathématique.

Graphes mathématiques et réseaux biologiques

En mathématique, le terme **graphe** désigne une représentation formelle d'un ensemble d'entités et de relations entre elles.

- Les entités sont dénommées **noeuds** du graphe.
- Les relations sont dénommées **arêtes** si elles sont non-orientées, et **arcs** si elles sont orientée

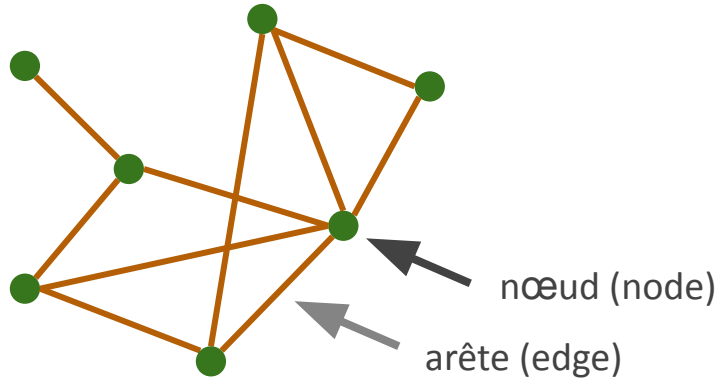
Les mathématiciens ont développé une **théorie des graphes**, qui traite de leurs propriétés en tant qu'objets mathématiques, et permet d'effectuer des opérations :

- Calcul de propriétés topologiques,
- Recherche de chemins
- Extraction de sous-graphes
- …

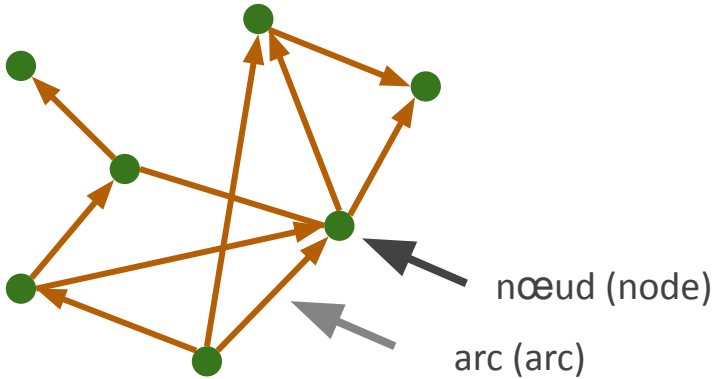
Depuis quelques décennies, on a utilisé des graphes mathématiques pour représenter des **réseaux d'interactions** entre entités biologiques et plus particulièrement biomoléculaires :

- Réseaux métaboliques (substrats ⟶ réactions ⟶ produits)
- Réseaux de régulation (facteurs transcriptionnels ⟶ gènes)
- Interactions protéine – protéine
- Co-expression de gènes à partir de données transcriptomiques

Graphe non orienté



Graphe orienté



Propriétés topologiques

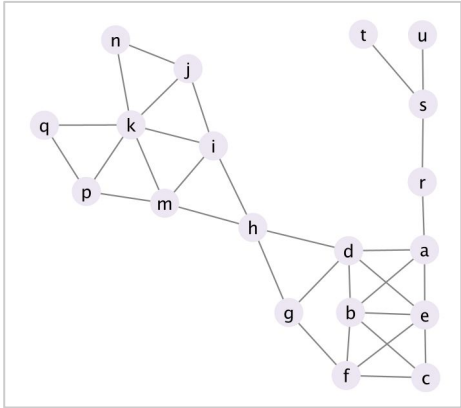
Les propriétés topologiques d'un graphe décrivent sa structure et son organisation :

Degré : le degré (*degree*) *k* d'un nœud est le nombre d'arêtes qui lui sont adjacentes (connectées). Il correspond donc au nombre de connexions ou nombre de premiers voisins.

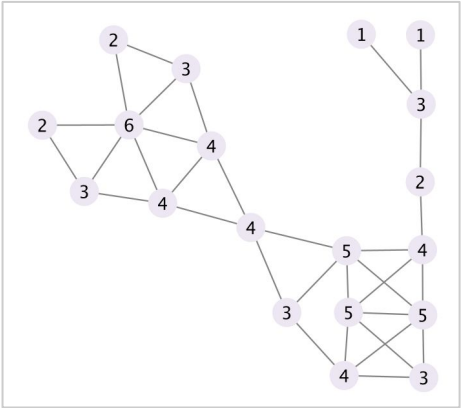
Chemin : Un chemin (*path*) dans un graphe est une suite de nœuds connectés par des arêtes, qui permet de passer d'un nœud de départ à un nœud d'arrivée en suivant les connexions existantes. Dans un chemin, chaque nœud ne peut être parcouru qu'une seule fois.

Le plus court chemin : Le plus court chemin (*shortest path*) entre deux noeuds d'un graphe est le chemin reliant ces deux nœuds en passant par le moins d'arêtes possibles. La *distance* entre deux nœuds est le nombre d'arêtes comprises dans le plus court chemin.

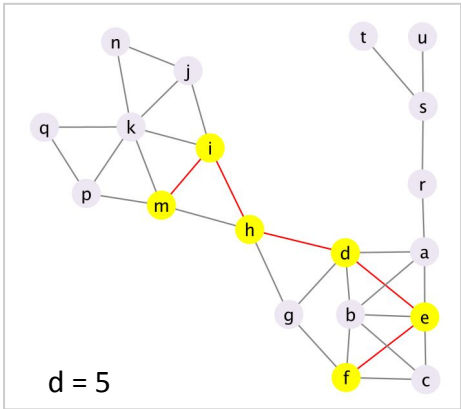
Exemple de réseau



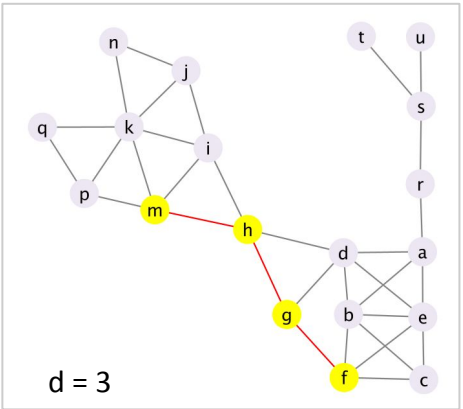
Degré de chaque noeud



Un des chemins possibles de m à f



Chemin le plus court de m à f



Propriétés topologiques

Centralité (*centrality*) **:** la centralité mesure l'importance d'un nœud au sein d'un réseau. Il existe plusieurs types de centralité, dont :

- **Centralité de distance** (*distance centrality*) : distance moyenne entre chaque noeud et tous les autres noeuds. Il s'agit d'une mesure globale de la centralité (dépend de l'ensemble du réseau).
- **Centralité de degré** (*degree centrality*) : la centralité de degré C_D d'un nœud *v* est égale à son degré. Il s'agit d'une centralité *locale* (ne dépend que du voisinage immédiat de chaque noeud).

$$C_D(v) = \deg(v)$$

- **Centralité d'intermédiarité** *(betweenness centrality)* : "fréquence de passage" ; fréquence à laquelle un nœud *v* se trouve sur le plus court chemin entre deux autres noeuds (indices *s* et *t)*. Il s'agit d'une mesure globale (dépend de l'ensemble du réseau)

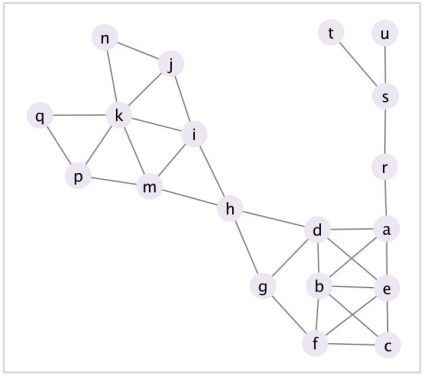
$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

v	noeud dont on veut calculer la centralité
s, t	indices qui énumèrent toutes les paires de noeuds distincts de *v*
σ_{st}	nombre total de plus courts chemins entre les nœuds *s* et *t*
$\sigma_{st}(v)$	nombre de ces chemins qui passent par le noeud *v*

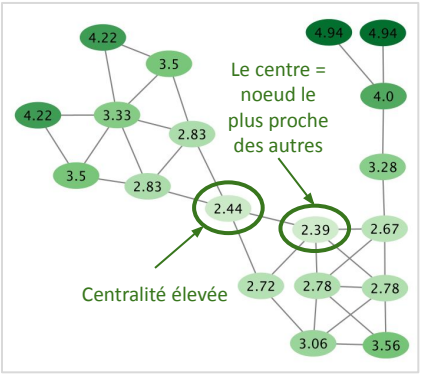
Note : dans certains cas il peut y avoir plusieurs chemins de même taille entre deux noeuds *s* et *t*, dont certains passent par *v* et d'autres pas

Le centre d'un réseau dépend du choix de la mesure de centralité (exemple ci-contre).

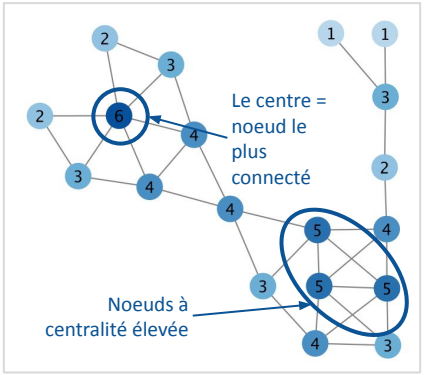
Exemple de réseau



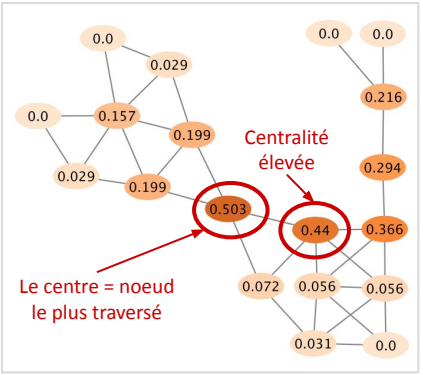
Centralité de distance



Centralité de degré



Centralité d'intermédiarité



Propriétés topologiques

Coefficient de regroupement (*clustering coefficient*) **:** pour un nœud *i*, le coefficient de regroupement C_i indique la proportion des paires de voisins qui sont également connectées entre elles :

$$C_i = \frac{2E_i}{k_i(k_i-1)}$$

E_i = nombre de connexions entre les voisin de *i*

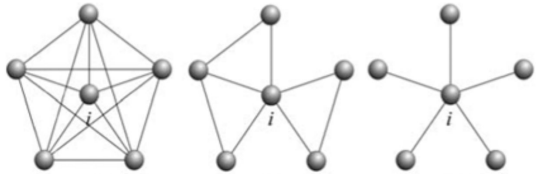
k_i = degré de i (i.e., nombre de voisins de *i*)

$k_i(k_i\text{-}1)/2$ = *nombre d'interactions a priori possibles entre* k_i *noeuds*

Cette mesure permet d'évaluer la densité locale (i.e., le voisinage du nœud *i*) des connexions dans un réseau et d'identifier des groupes de nœuds interconnectés.

Il est possible d'estimer la densité globale d'un réseau grâce à la moyenne des coefficients de regroupement de tous les nœuds (*average clustering coefficient*) :

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$$



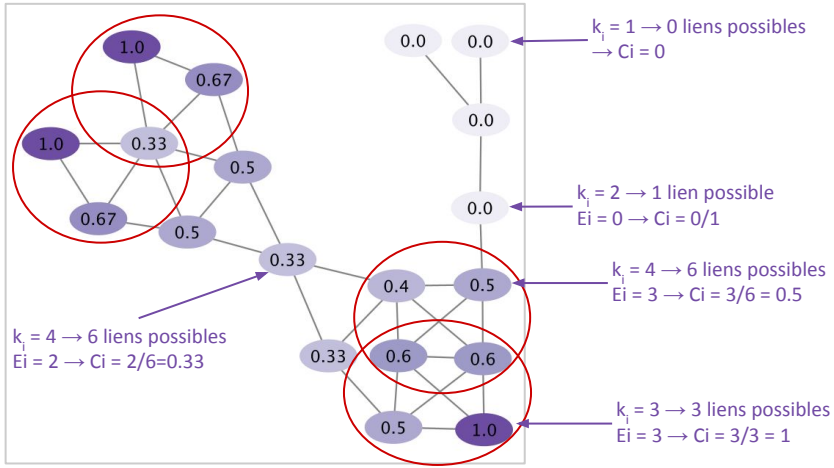
C_i = 1 **C_i = 0.3** **C_i = 0**

On appelle **clique**, ou **sous-graphe complet**, un sous-graphe dont les noeuds sont tous directement interconnectés.

Quand C_i = 1, le nœud *i* et tous ses voisins forment une **clique**.

Note : l'inverse n'est pas vrai – une clique peut comporter des noeuds de coefficient de clustering inférieur à 1 (exemple ci-dessous).

Coefficients de clustering et quelques cliques



Annotation par association (métaphore "coupable par association")

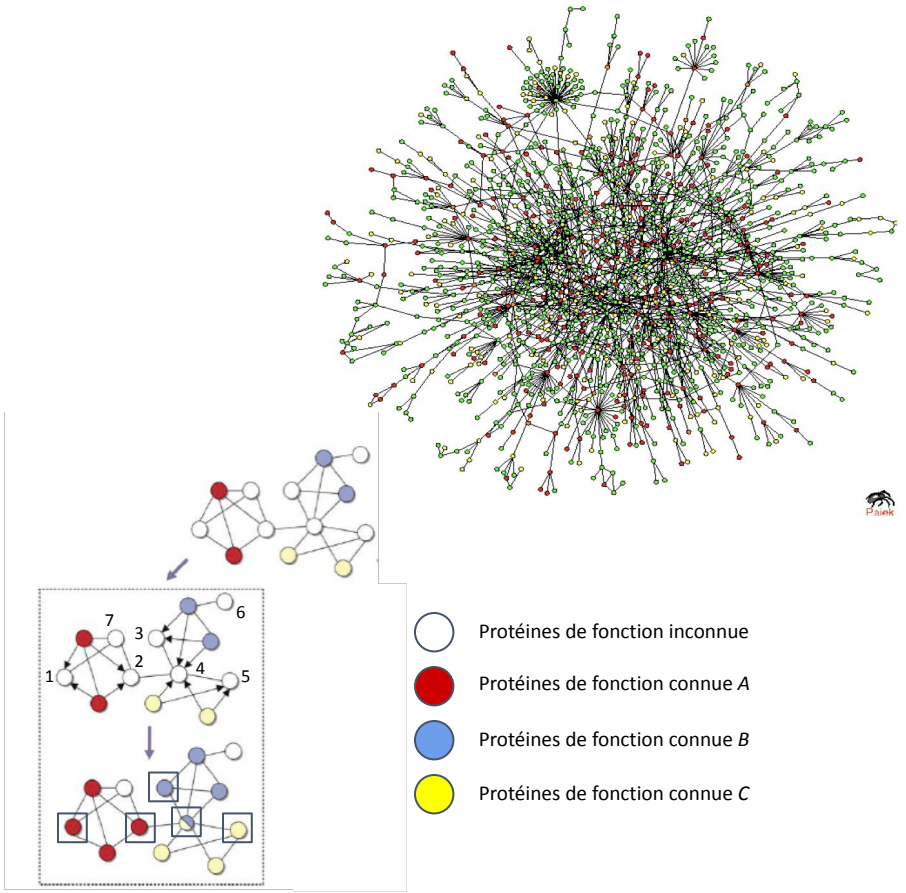
L'étude des propriétés topologiques de l'interactome nous permet d'aborder de nombreuses questions biologiques et de formuler des hypothèses concernant les fonctions possibles de protéines.

Par exemple, si une protéine de fonction inconnue est observée dans un réseau avec des protéines bien caractérisées (ayant des rôles spécifiques dans le métabolisme, la signalisation, etc.), il est probable qu'elle partage une fonction similaire ou liée.

Ce principe peut être utilisé pour identifier des nouvelles protéines potentiellement impliquées dans une maladie : si une protéine d'intérêt interagit avec plusieurs protéines connues pour leur implication dans une maladie spécifique (ex., le cancer du côlon ou maladie de Parkinson), cette protéine pourrait également jouer un rôle dans cette pathologie.

Exemple: Dans le réseau montré ici à droite on pourrait inférer la fonctions des protéines 1-5 grâce aux interactions avec des protéines de fonction connue, en considérant par exemple un seuil de 2 interactions pour assigner une fonction. Cette approche fournirait présente bien entendu un certain risque d'erreur, comme toute inférence.

Dans le cas de la protéine 4, ceci nous amènerait à lui associer une double fonction (elle interagit avec 2 protéines de fonction *B*, et 2 protéines de fonction *C*).



Sous-graphes

Un **sous-graphe** est une partie d'un graphe plus vaste, composée d'un sous-ensemble de nœuds et d'arêtes de ce graphe. Il conserve la structure et les connexions présentes dans le graphe d'origine, mais se limite aux éléments sélectionnés.

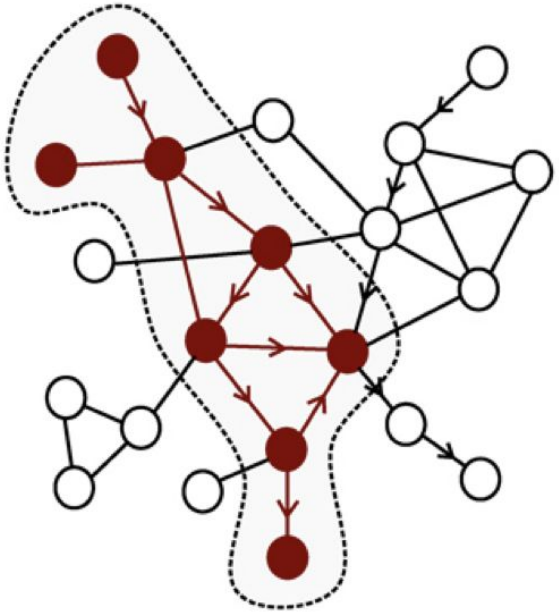
Dans les réseaux biologiques, les sous-graphes sont généralement nommés **sous-réseaux**.

Les critères pour identifier ou extraire des sous-graphes dans un réseau peuvent inclure :

- sélection de nœuds selon des **propriétés topologiques** spécifiques, tels que le coefficient de regroupement ;
- **module** ou c**ommunauté** (ensemble de noeuds fortement interconnectés) ;
- ensemble de noeuds qui partagent des **propriétés biologiques** (facteurs transcriptionnels, enzymes, enzymes, protéines participant à un processus biologique particulier, …), et les interactions entre ces noeuds.

Par exemple, on pourrait extraire un sous-réseau de l'interactome humain en sélectionnant

- comme noeuds, les protéines impliquées dans une pathologie donnée ;
- comme arêtes, toutes les interactions entre ce sous-ensemble de protéines.

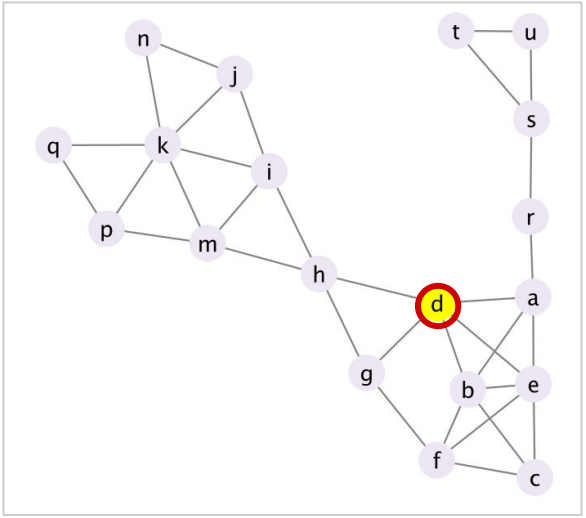


Voisinage d'un noeud

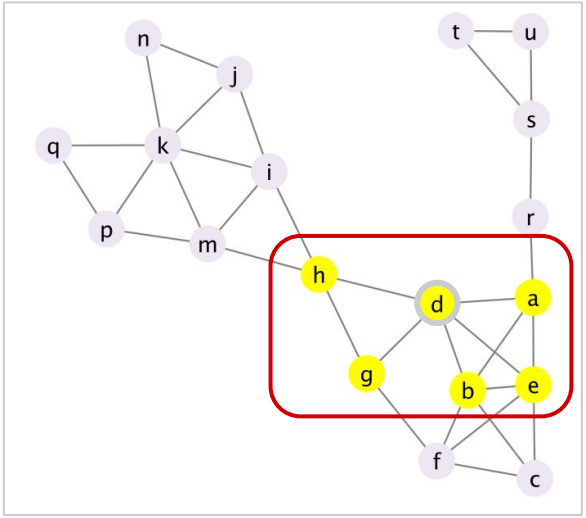
Le voisinage d'un noeud d'intérêt est l'ensemble des noeuds qui y sont connectés, en précisant une distance minimale

- Voisinage de premier ordre : ensemble des premiers voisins, c'est-à-dire les noeuds immédiatement connectés au noeud d'intérêt.
- Voisinage de deuxième ordre : ensemble des noeuds connectés directement au noeud d'intérêt, ou connectés à ses premiers voisins. Le voisinage d'ordre 2 inclut le voisinage de premier ordre.
- Voisinage d'ordre *n* : ensemble des noeuds connectés au noeud d'intérêt via un chemin de maximum *n* arêtes. Le voisinage d'ordre *n* inclut tous les voisinages d'ordre inférieur à *n*.

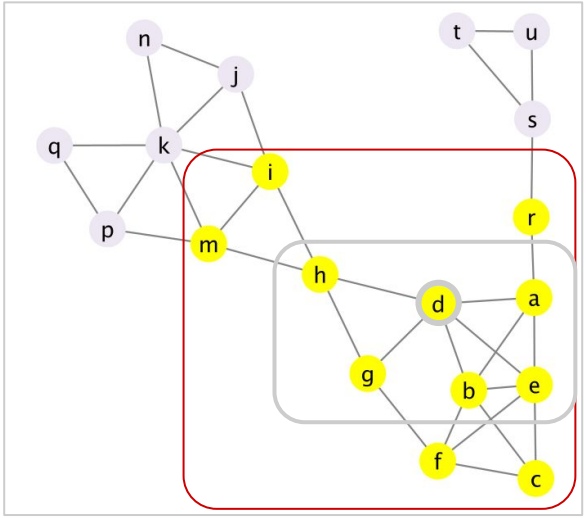
Noeud d'intérêt



Voisinage de premier ordre



Voisinage de deuxième ordre



La base de données STRING

La base de données **STRING** (*Search Tool for the Retrieval of Interacting Genes/Proteins,* https://string-db.org/) est une ressource biologique qui centralise et intègre les informations sur les interactions entre protéines.

Elle comprend à la fois des interactions **physiques** (contacts physiques, directs ou indirects, entre protéines) et **fonctionnelles** (protéines qui participent à la même voie biologique sans nécessairement interagir directement), issues de plusieurs sources : des données expérimentales, des bases de données de référence, ainsi que des prédictions bioinformatiques.

Les interactions sont collectées à partir de données expérimentales, de bases de données publiques, de données de co-expression génique, ainsi que par "fouille de texte" dans la littérature scientifique. Chaque interaction est pondérée pour indiquer sa fiabilité, avec des **scores de confiance** (***confidence*** en anglais) basés sur la quantité et la qualité des indications disponibles.

STRING offre des **outils de visualisation et d'analyse** des réseaux d'interaction, permettant d'explorer les relations entre protéines et de formuler des hypothèses concernant leurs fonctions biologiques ou leur lien avec des pathologie.

Il est possible de rechercher des interactions pour des protéines ou gènes en saisissant leur nom, leur identifiant, ou en entrant une liste de protéines d'intérêt. L'interface permet également d'interroger la base de données par processus biologique, pathologie ou organisme.

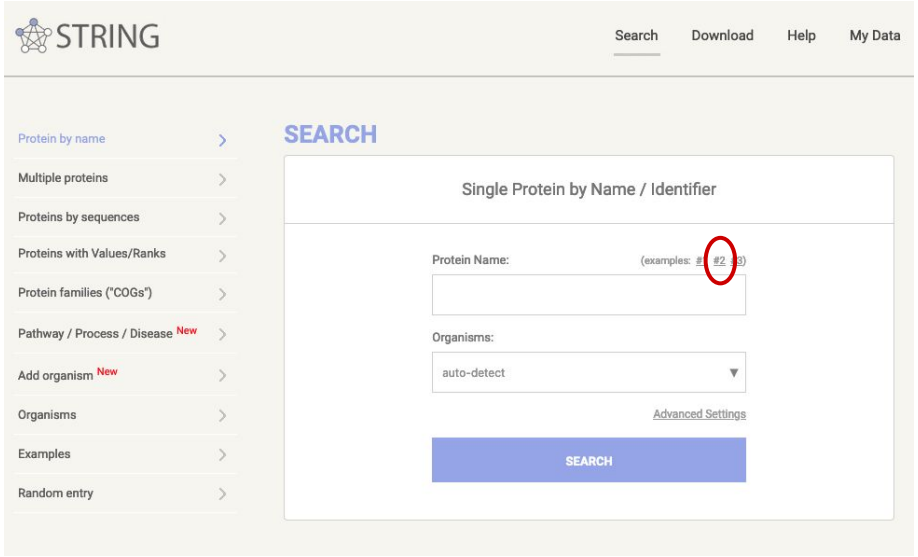
Un aperçu des statistiques sur le contenu de la base de données est disponible sur cette page : STRING database - Statistics.

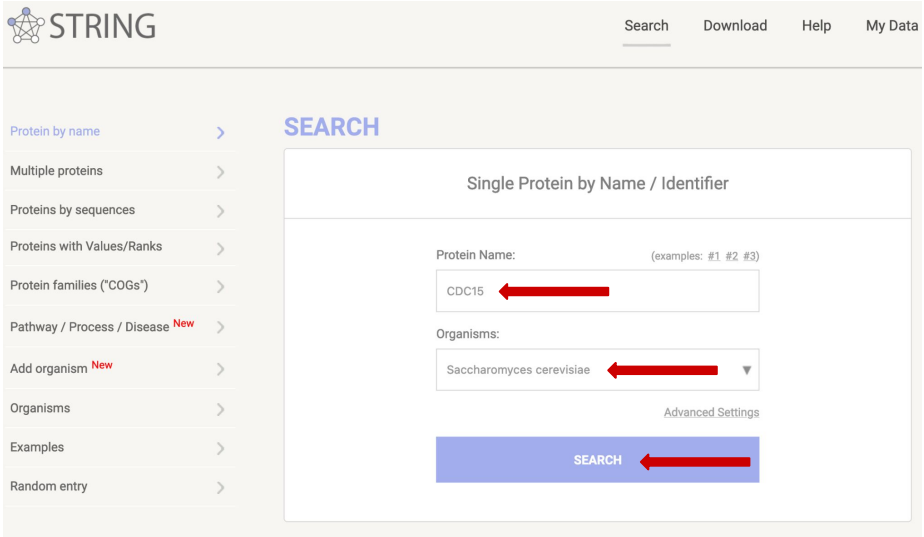
von Mering, C. V. STRING: a database of predicted functional associations between proteins. Nucleic Acids Research 31, 258–261 (2003). doi.org/10.1093/nar/gkg034
Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research 51, D638–D646 (2023). doi.org/10.1093/nar/gkac1000

Tutoriel - Prise en main de la base de données STRING

Tutoriel – Prise en main de la base de données STRING – Exemple de réseau

- Sur la page d'accueil de **STRING** (string-db.org), sélectionnez la modalité de recherche "**Protein by name**".
- Cliquez sur l'**exemple "#2"** proposé par STRING. Cela vous permettra d'extraire le voisinage de la protéine CDC15 dans le réseau d'interactions de la levure du boulanger *Saccharomyces cerevisiae*.
- Cliquez sur "**Search**".





Tutoriel – Prise en main de la base de données STRING

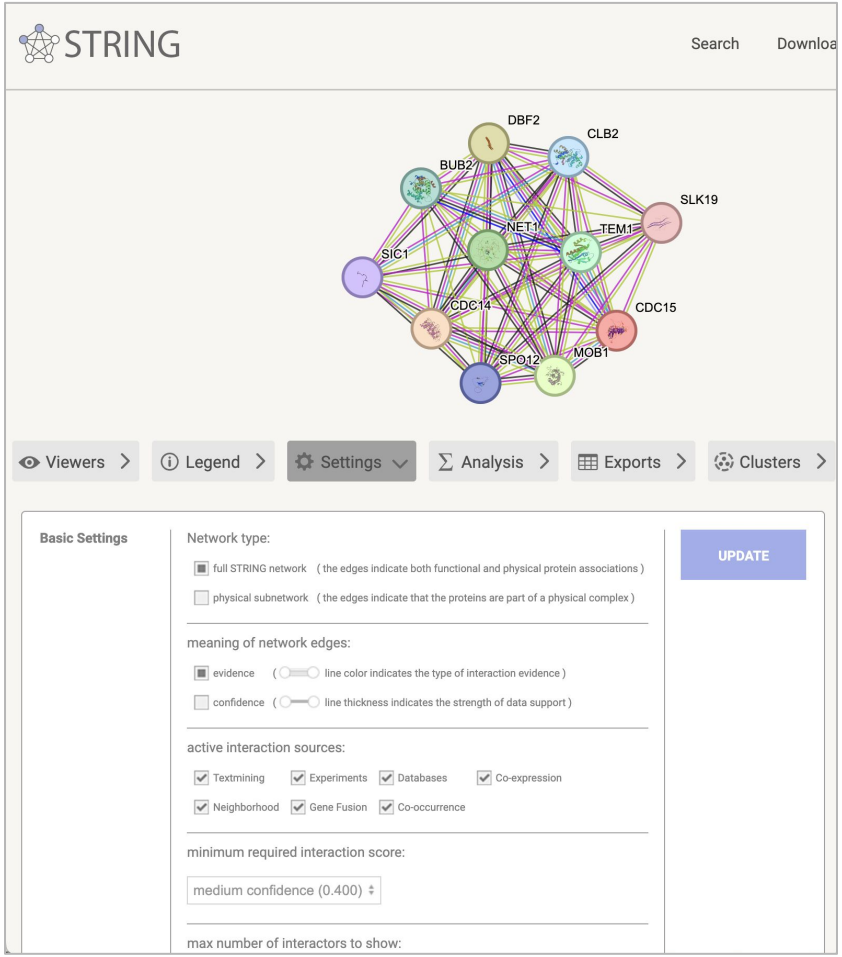
La recherche par nom de protéine collecte par défaut 10 noeuds au sein du **voisinage de premier ordre** de CDC15 et affiche le résultat sous forme d'un réseau.

En dessous du réseau, vous pouvez consulter different onglets qui permettent de

- changer la modalité de visualisation des données (**Viewers**). Pour ce TP on utilisera exclusivement la modalité "*Network*" ;
- afficher la légende avec les détails sur la signification de chaque élément visuel dans le réseau (**Legend**) ;
- changer les paramètres d'affichage des données sur le réseau (**Settings**) ;
- visualiser des propriétés topologiques du réseau et les résultats de l'analyse d'enrichissement fonctionnel (**Analysis**) ;

Les buttons "***More***"et "***Less***" permettent d'ajouter ou retirer des nœuds du réseau. Par default, le nombre maximal de nœuds ajoutés/retirés est égal à 10 (ce paramètre peut être changé dans l'onglet **Settings**).

Les deux autres onglets (**Exports** et **Clusters**) ne seront pas utilisés dans ce TP.



Tutoriel – Prise en main de la base de données STRING – Paramètres de base

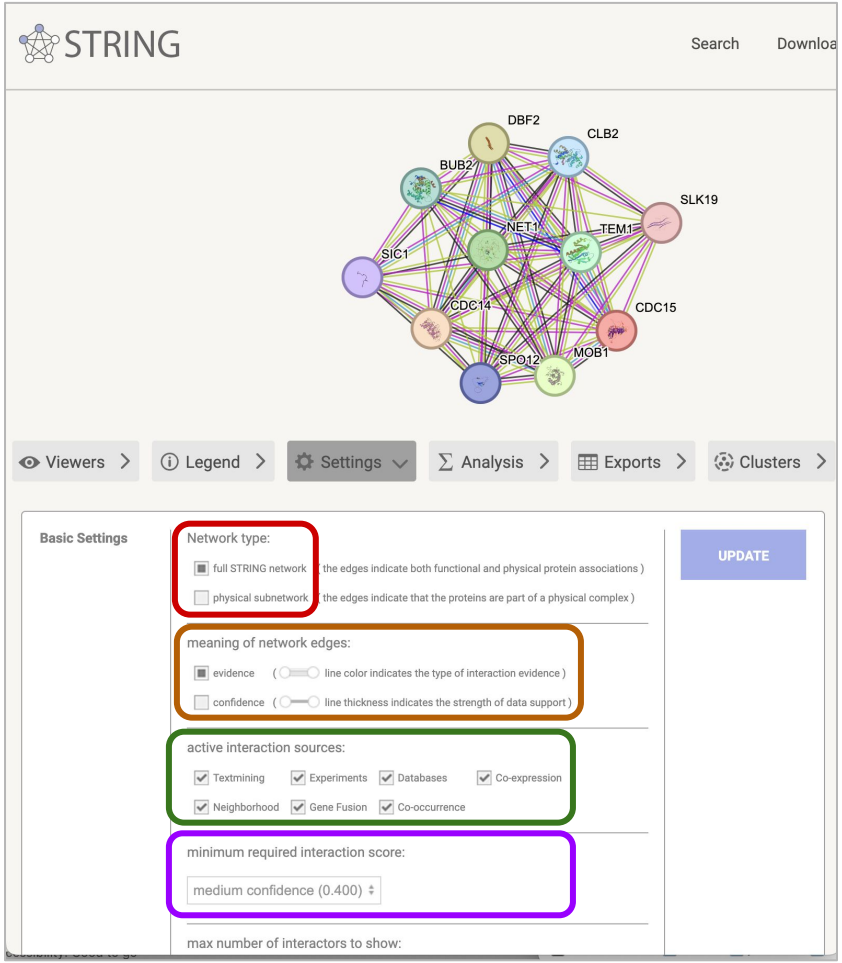
L'onglet **Settings** est selectionné par default.

Parmi les paramètres de base (Basic Settings), l'option "**Network type**" offre le choix entre afficher le réseau complet (l'ensemble des interactions physique et fonctionnelles, *full STRING network*) ou restreindre l'affichage au réseau généré à partir des données d'interaction physiques (*physical subnetwork*).

Active interaction sources. Par default, STRING affiche les arêtes du réseau sous forme de liens multiples. Chaque lien représente une preuve ou indication à l'appui d'une interaction donnée. Chaque indication provient d'une source de données différente.

Le **score de fiabilité (confidence score)** peut avoir des valeurs comprises entre 0 et 1. Le score est d'autant plus élevé qu'on dispose d'indications fiables pour établir que deux protéines interagissent. Par default, STRING affiche les interactions avec un score égale ou superieur à 0.4 (*medium confidence score*).

Grace à l'option "**meaning of network edges**", il est possible de changer la modalité d'affichage des arêtes en représentant soit le nombre d'indications, soit le score de fiabilité (*confidence score*).



Tutoriel – Prise en main de la base de données STRING – Légende du graphique

- Cliquez sur l'onglet **Legend**.

Cet onglet fournit des informations détaillées sur les éléments et les codes de couleur utilisés dans la visualisation des réseaux d'interactions.

Les Nœuds

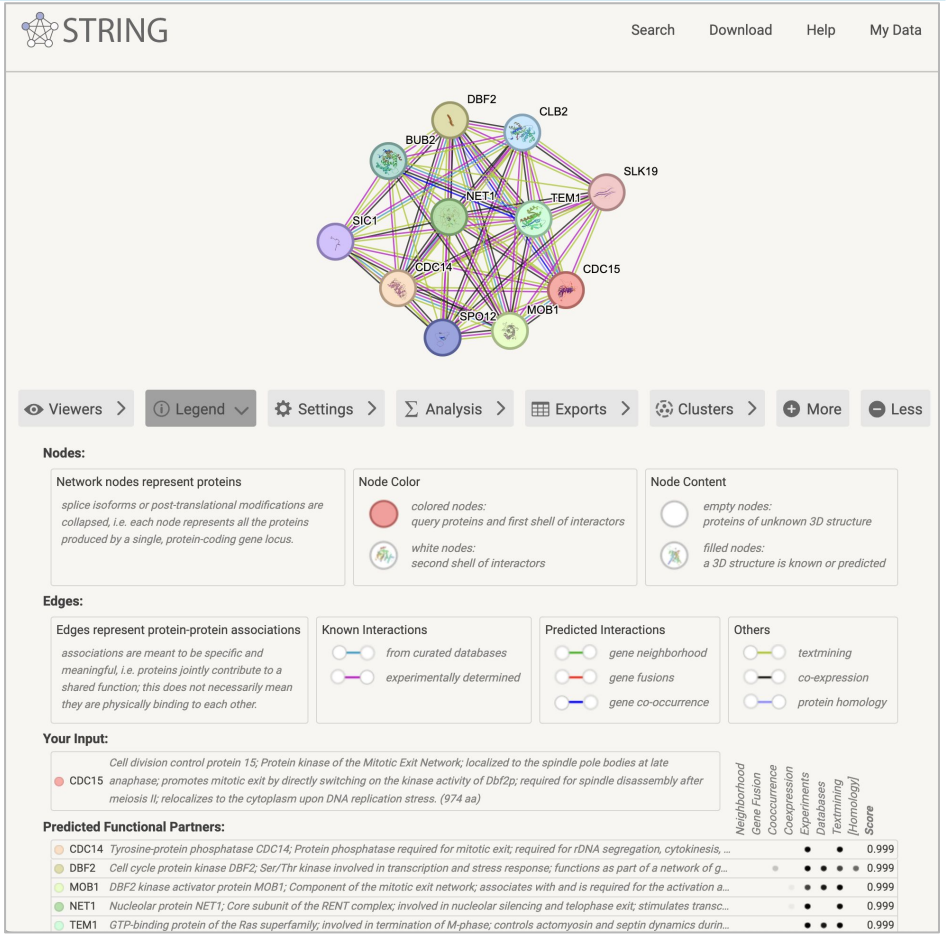
Les nœuds colorés représentent notre protéines d'intérêt et les voisin de première ordre. Les voisins de deuxieme ordre, si presents, son colorés en blanc. Si dans le nœud il y a une image, cela veut dire qu'une structure tridimensionnelle, soit expérimentalement déterminée soit prédite, est disponible.

Les Arêtes

Chaque couleur de lien représente une source d'information. Dans ce TP, les sources que nous interessent sont les suivantes:

- Données expérimentales, bases de données de référence (Known Interactions).
- Données de co-expression génique, fouille de texte (Others)

Les interactions prédites ne seront pas prises en compte.



Tutoriel – Prise en main de la base de données STRING – Statistiques du sous-réseau sélectionné

- Cliquer sur l'onglet **Analysis**

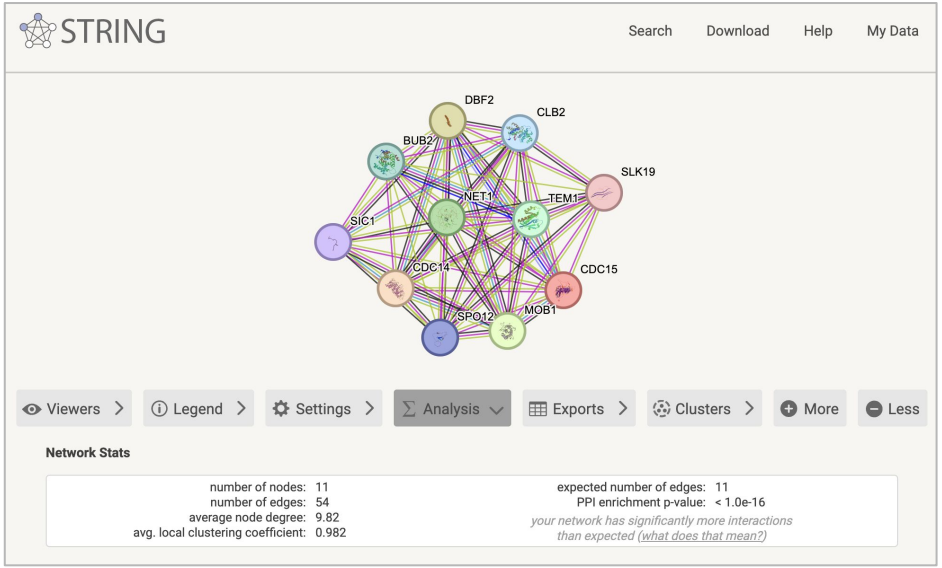
La première analyse qui est affichée est l'analyse de réseau (***Network Stats***), qui indique

- les nombres de nœuds (**nodes: 11**) et d'arêtes (**edges: 54**) dans le réseau sélectionné ;
- des propriétés générales de ce réseau telles que degré moyen des nœuds et coefficient de regroupement moyen.

Cette section indique également si le réseau visualisé est significativement plus connecté que ce que l'on pourrait attendre par hasard.

- **expected number of edges :** si on avait sélectionné aléatoirement le même nombre de noeuds dans le réseau complet on s'attendrait à observer 11 interactions entre elles.
- La **p-value** (significativité critique) indique la probabilité d'obtenir un nombre au moins aussi élevé d'arêtes dans un réseau aléatoire comportant le même nombre de noeuds et d'arêtes que le réseau complet de *Saccharomyces cerevisiae*. Une p-valeur est faible indique que les protéines sélectionnées ont un nombre significativement élevé d'interactions, ce qui suggère que ces protéines forment un groupe connecté biologiquement, au moins partiellement.

ATTENTION : Lorsque vous interrogez STRING en partant d'une protéine comme dans cet exemple, STRING ajoute par défaut un total de 10 protéines parmi ses premières voisines dans le réseau. L'enrichissement de connectivité résulte donc du processus de collecte du sous-réseau. Il faut donc ignorer cette statistique pendant ce TP, elle n'est informative que quand pour d'autres types de requêtes, où l'on extrait un sous-réseau sur base de critères indépendants de la connectivité.



Tutoriel – Prise en main de la base de données STRING - Enrichissement fonctionnel

Sous les statistiques du réseau, STRING fournit différents tableaux contenant les résultats des analyses d'enrichissement fonctionnel. STRING utilise différentes sources d'annotations pour cette analyse, et notamment : Gene Ontology, les voies de signalisation de KEGG, et les localisations cellulaires.

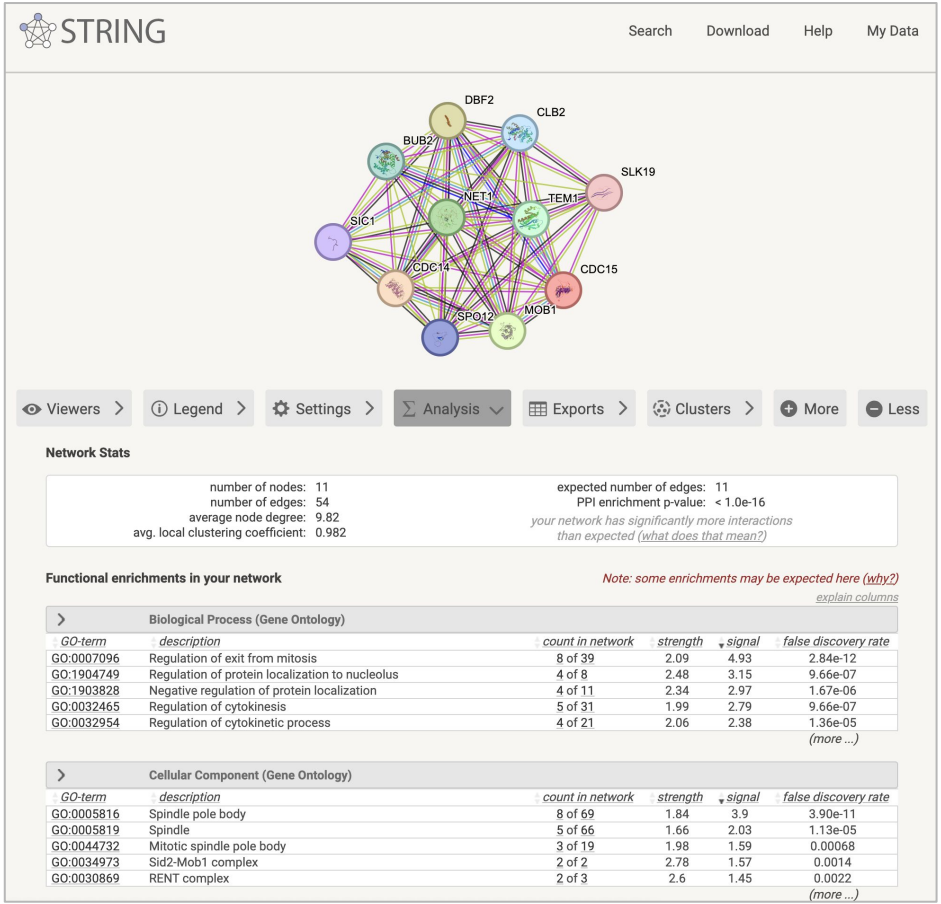
Pour chaque terme d'annotation enrichi (indiqué avec son identifiant et sa description dans les deux premières colonnes), les information à retenir sont les suivantes:

- ***Count in network***: le nombre de protéines ayant cette annotation dans notre sélection, comparée à l'ensemble des protéines ayant la même annotation dans le réseau complet. Par exemple, le réseau d'interactions de Saccharomyces cerevisiae comporte 39 protéines annotées "Regulation of exit from mitosis", et 8 d'entre elles se retrouvent dans le sous-réseau de 11 protéines autour de CDC15 (incluse).
- ***Strength***: une mesure qui décrit l'ampleur de l'enrichissement.

  $strength = log_{10}(observed / expected)$

  observed : nombre de protéines du sous-réseau annotées pour cette classe fonctionnelles

  expected : nombre attendu si on avait sélectionné le même nombre de noeuds au hasard
- ***False Discovery Rate (FDR)***: une mesure décrivant la significativité de l'enrichissement. Plus petite est la FDR, plus l'enrichissement est significatif.
- ***Signal*** : fait le compromis entre deux autres statistiques. Nous l'ignorerons pendant ce TP.



Enrichissement fonctionnel – Principe

L'analyse de l'**enrichissement fonctionnel** est une approche très couramment utilisée en bioinformatique pour associer des propriétés fonctionnelles à un groupe de gènes ou de protéines d'intérêt.

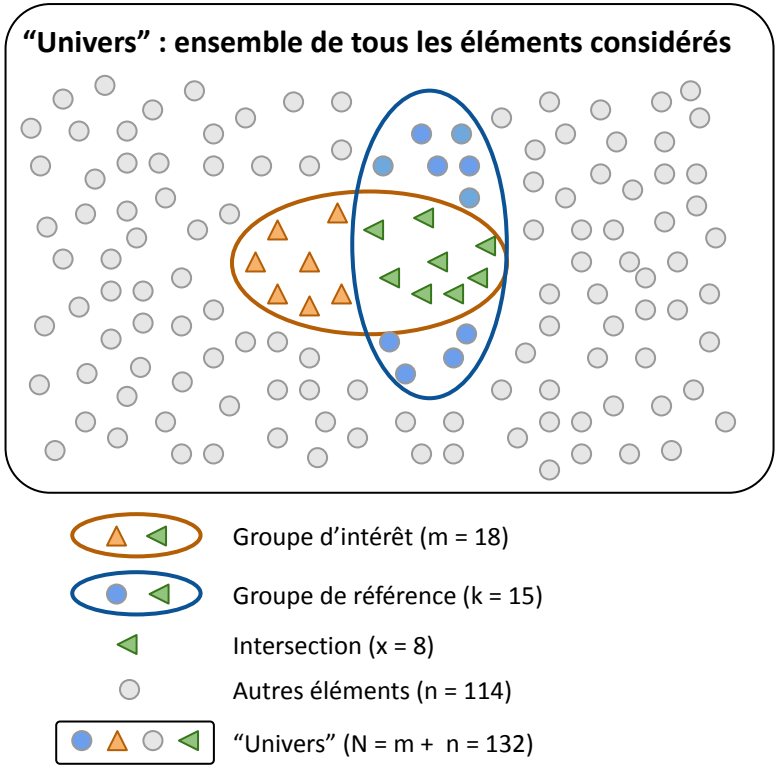
Le **groupe d'intérêt** (gènes ou protéines) peut provenir de différents types de données

- Gènes co-exprimés (données transcriptomiques)
- Gènes appartenant au même opéron (bactéries)
- Protéines faisant partie d'un sous-réseau d'interaction
- …

On compare ce groupe d'intérêt à un **groupe de référence** (gènes ou protéines), par exemple

- Gènes impliqué dans une même voie métabolique (bases de données métaboliques)
- Gènes-cibles d'un facteur de transcription (bases de données de facteurs transcriptionnels)
- Gènes impliqués dans un même processus de la Gene Ontology
- …

Question : l'intersection entre le groupe d'intérêt et le groupe de référence est-elle plus élevée que ce à quoi on s'attendrait au hasard ?



Enrichissement fonctionnel –Statistiques

On peut calculer diverses statistiques pour mesurer le degré d'enrichissement et sa significativité

- Espérance aléatoire (*expectation,* ***exp***) : nombre d'éléments marqués auxquels on s'attendrait si on avait sélectionné le même nombre d' éléments au hasard.

  Dans l'exemple: $exp = m / N = 18 / 132 = 2.04$

- Rapport observations / espérance (***obs/exp ratio***, églaement appellé ***fold change***) : rapport entre le nombre d'éléments marqués dans la sélection ($x = obs = 8$)

  Dans l'exemple, le ratio vaut $r = x / exp = 8 / 2.04 = 3.91$

  Notre sélection est donc enrichie d'un facteur ~4 par rapport à l'espérance aléatoire.

- Le score ***strength*** de STRING-DB est le logarithme en base 10 de ce ratio.

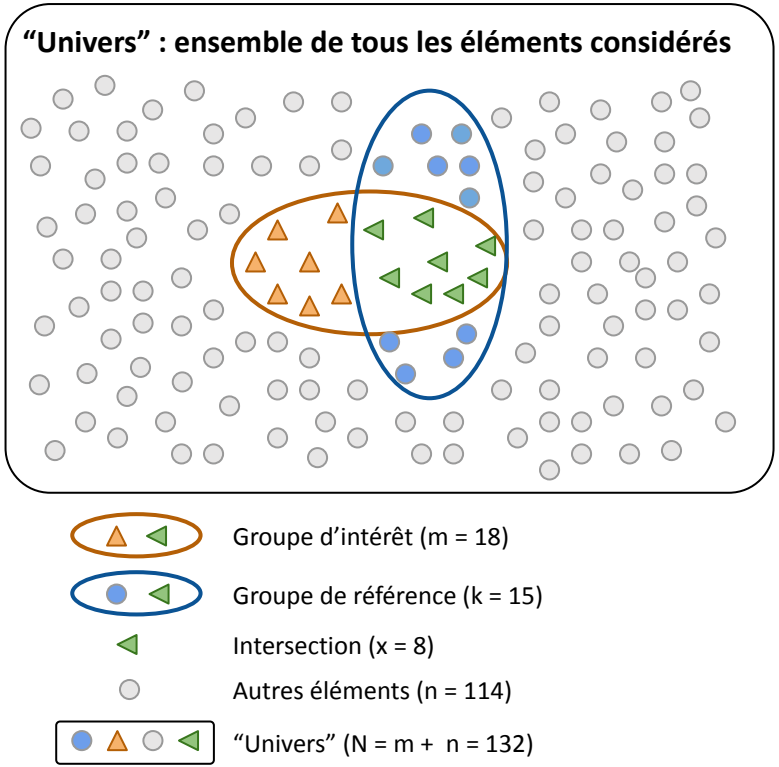
  $strength = log10(obs/exp) = log10(3.91) = 0.59$

- On peut également calculer une p-valeur, qui indique la probabilité d'obtenir un enrichissement au moins aussi important que celui observé.

  Dans l'exemple, cette p-valeur vaut $P=9\ 10^{-5}$

  Comme cette p-valeur est **nettement inférieure à 1**, on peut considérer que l'enrichissement est **statistiquement significatif**.

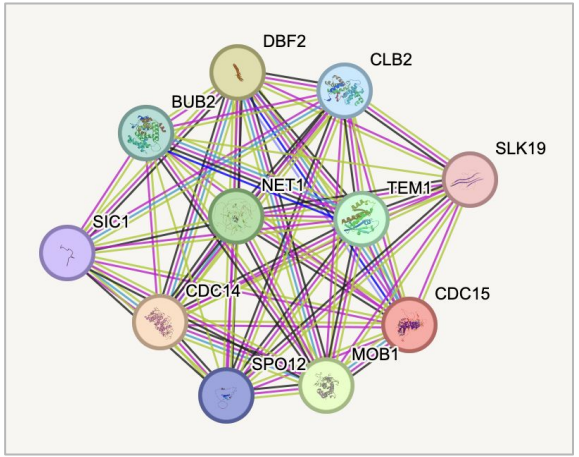
Note: STRING-DB ne retourne pas directement la P-valeur mais un **False Discovery Rate (FDR)**, qui inclut une correction pour tenir compte des test multiples. Les concepts de correction de test multiple ne sont pas abordés ici, il vous faut simplement savoir que ***plus le FDR est faible, plus l'enrichissement est significatif.***



Exercice 1 - Réseau de voisinage de CDC15 chez la levure *Saccharomyces cerevisiae*

Exercice 1 - Sous-réseau de voisinage de CDC15 chez la levure *Saccharomyces cerevisiae*

Sur Ametice, répondez aux Questions 1 à 6 du TP7.



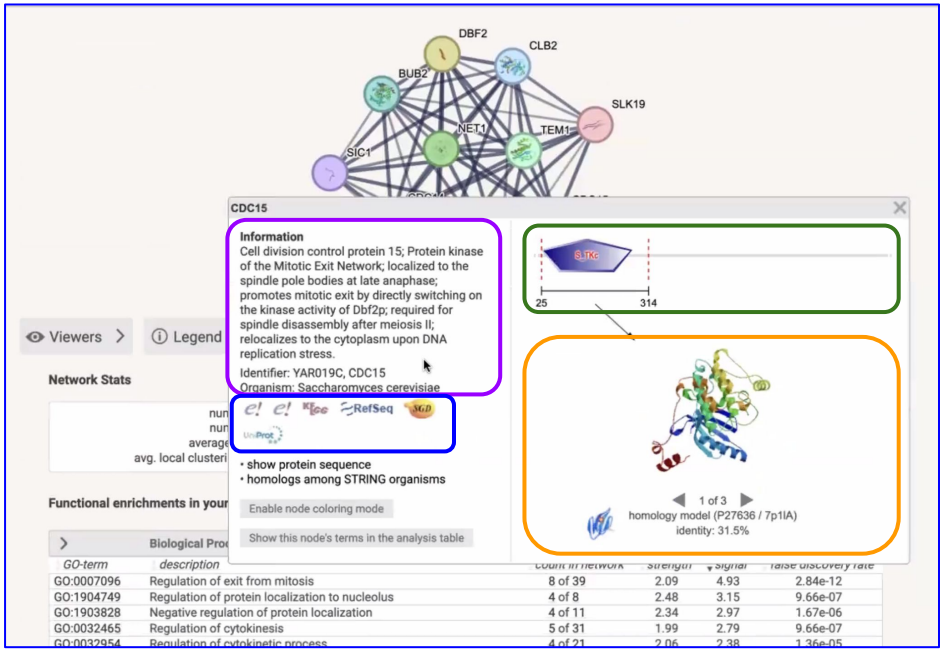
1. Combien des nœuds y a-t-il dans le sous-réseau autour de CDC15 (inclus) ?
2. Combien d'arêtes y a-t-il dans ce sous-réseau ?
3. Quel est le coefficient de regroupement moyen (*avg. local clustering coefficient*) de ce sous-réseau ?
4. Que signifie le fait que le coefficient de regroupement moyen d'un réseau soit élevé, mais inférieur à 1) ?
   a. Cela signifie que le réseau est complet avec des liens directs entre tous les nœuds.
   b. Cela indique que les nœuds du réseau ont tendance à former des groupes fortement connectés entre eux.
   c. Cela montre que le réseau a peu de liens.
   d. Cela signifie que les nœuds sont connecté exclusivement au nœud d'intérêt (CDC15 dans notre cas).
5. Quelle est l'annotation "Biological Process" (Gene Ontology) avec l'enrichissement les plus significatif (copier & coller l'identifiant GO).
6. Après avoir consulté tous les tableaux d'enrichissement, quels sont les processus cellulaires dans lequel les protéines de ce sous-réseau sont vraisemblablement impliquées ? (une ou plusieurs réponses)
   a. Division cellulaire
   b. Apoptose
   c. Glycolyse
   d. Transcription

Tutoriel - Prise en main de la base de données STRING – Fiche d'info sur une protéine

- Cliquez sur un noeud (par exemple CDC15) pour consulter l'information sur la protéine correspondante.

La fenêtre d'information contient

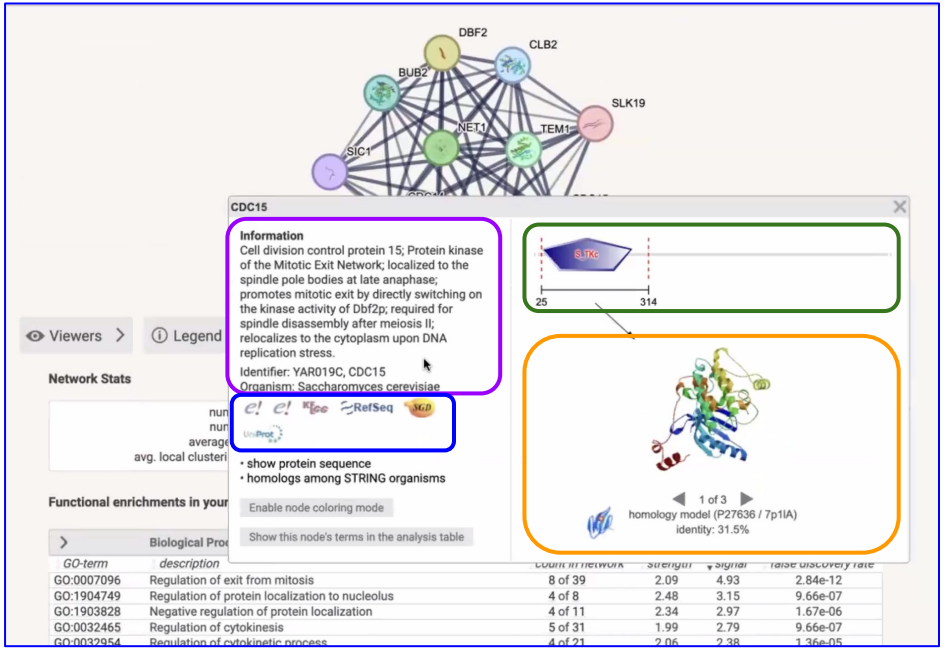
- une **description** de la fonction de la protéine,
- des **hyperliens** vers de bases des données externes (UniProt, Ensembl, KEGG, …).
- la **composition en domaines**
- une visualisation de la **structure tridimensionnelle**, si disponible.



Exercice 1 - Sous-réseau de voisinage de CDC15 chez la levure *Saccharomyces cerevisiae* – Info sur une protéine

Sur Ametice, répondez aux Questions 7 à 9 du TP7.

7. A quelles catégories fonctionnelles appartient la protéine CDC15 ? (une ou plusieurs réponses)
   a. Ubiquitin ligase
   b. Oxidoreductase
   c. Cytokine
   d. Proteine kinase
8. En cas de stress lié à la réplication de l'ADN, dans quel(s) compartiment(s) cellulaire(s) la protéine CDC15 se relocalise-t-elle ? (une ou plusieurs réponses)
   a. Lysosome
   b. Pôle du fuseau mitotique
   c. Noyau
   d. Cytoplasme
   e. Centriole
9. Parmi les structures tridimensionnelles disponibles pour CDC15, quelles méthodes ont été utilisées ? (une ou plusieurs réponses)
   a. Modèle AlphaFold
   b. Modélisation par homologie
   c. Structure expérimentale de PDB
   d. Aucun de ces types



Exercice 1 - Sous-réseau de voisinage de CDC15 chez la levure *Saccharomyces cerevisiae* – Interactions physiques

- Cliquer sur l'onglet **Settings**
- Dans la section "**meaning of network edges**", choisissez la modalité d'affichage ***confidence score***, puis cliquez sur le button "**UPDATE**".

Les interactions entre protéines sont désormais représentées par un seule arête, dont l'épaisseur est proportionnelle au score de fiabilité.

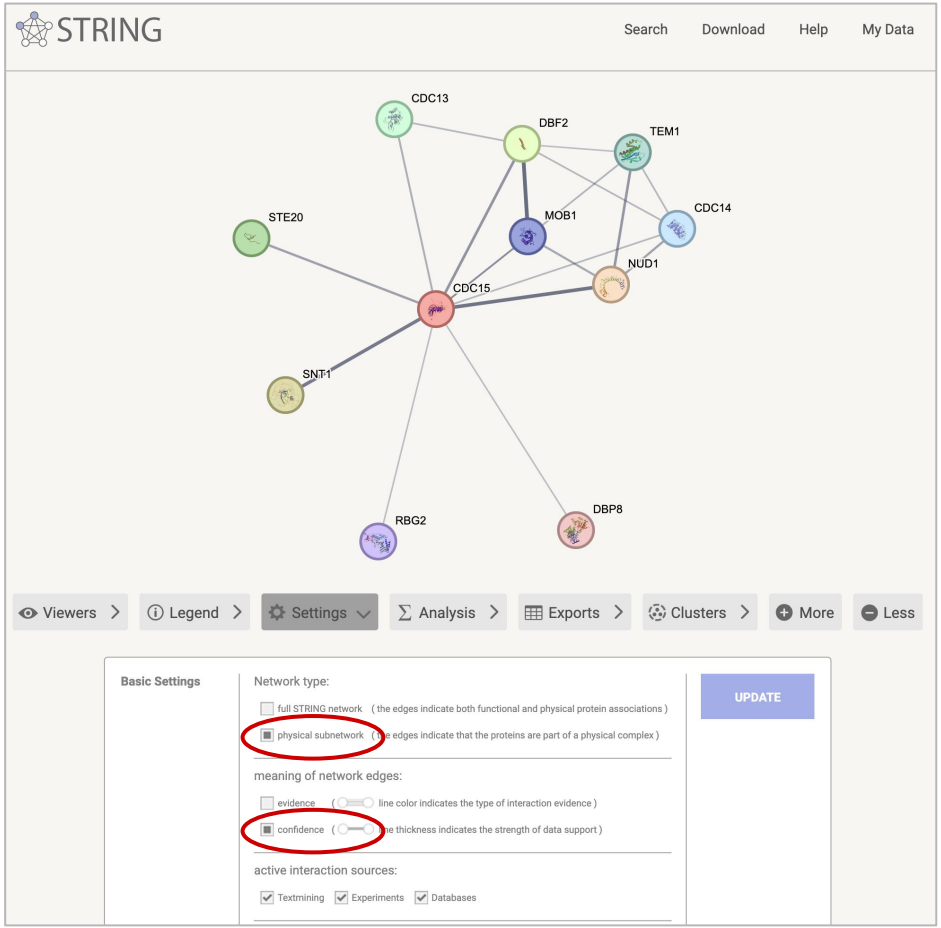
- Dans "**Network type**" choisissez d'afficher que les interactions physiques (*physical subnetwork*). Cliquez sur le button "UPDATE".
- Cliquez sur l'onglet **Analysis,** et observez les nouvelles statistiques du réseau.

Sur Ametice, répondez aux Questions 10 à 12 du TP7.

10. Combien des nœuds il y a dans le sous-réseau d'interactions physiques autour de CDC15 ?
11. Combien d'arêtes il y a dans le sous-réseau d'interactions physiques autour de CDC15 ?
12. Quel est le coefficient de regroupement moyen du sous-réseau d'interactions physiques autour de CDC15 (*avg. local clustering coefficient*) ?

Tentez de comprendre la raison pour laquelle ce sous-réseau obtient un coefficient relativement élevé.

Info: Le coefficient de regroupement moyen est obtenu en calculant, pour chaque noeud, son coefficient de regroupement (rapport entre le nombre d'arêtes et le nombre maximal d'arêtes possibles entre ses premiers voisins).



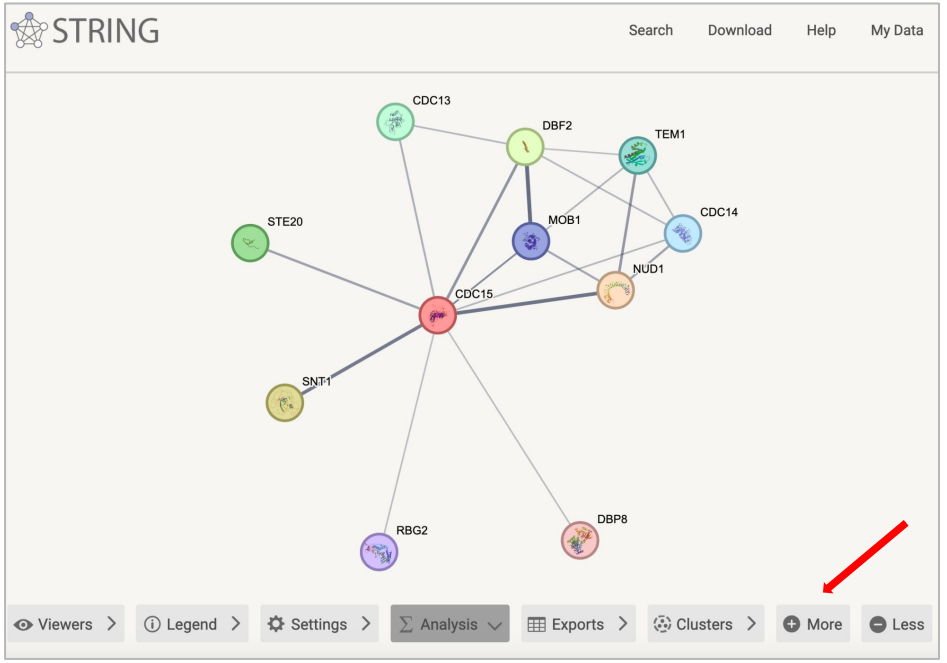
- Cliquer sur le button **More**

STRING ajoute les voisins de premier ordre des voisin de CDC15, ce qui revient à collecter son voisinage de deuxième ordre.

Astuce : sur le réseau étendu, les voisins de deuxième ordre sont colorés en blanc.

Sur Ametice, répondez aux Questions 13 à 15 du TP7.

13. Combien des nœuds ont été ajoutés ?
14. Combien d'arêtes y a-t-il dans ce sous-réseau ?
15. Consulter à nouveau tous les tableaux d'enrichissement. Y retrouve-t-on le(s) processus cellulaire(s) précédemment identifiés avec le “full STRING network” du voisinage d’ordre 1 ?



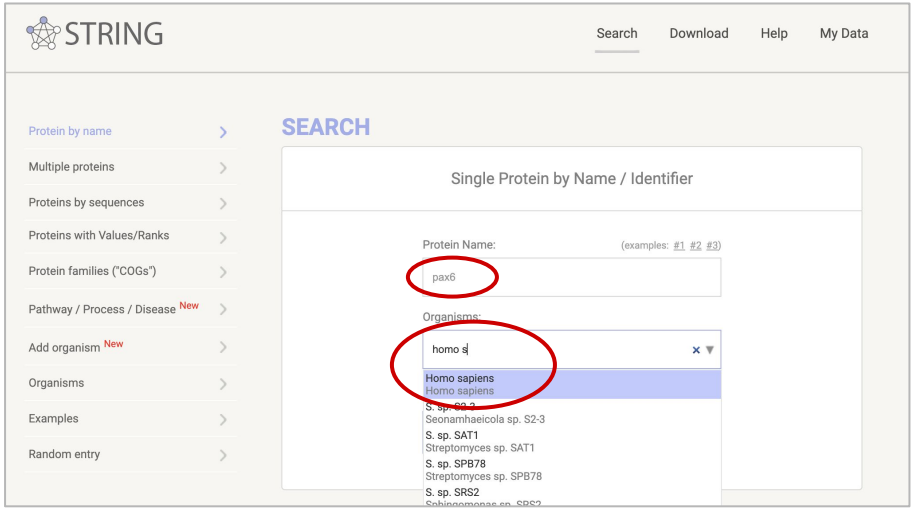
Exercice 2 - Exploration sous-réseau de voisinage de la protéine PAX6

Exercice 2 - Exploration du sous-réseau de voisinage de PAX6

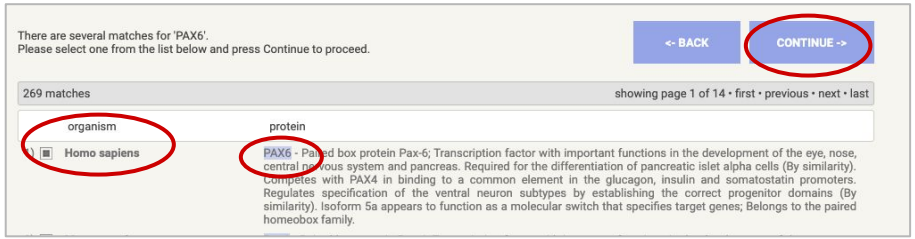
En partant du facteur transcriptionnel humain PAX6, nous utiliserons la base de données STRING pour extraire, visualiser et analyser son sous-réseau d'interactions moléculaires (voisinage de premier ordre), qui inclut des interactions physiques et fonctionnelles.

Ceci nous amènera à faire le lien entre les fonctions des protéines du réseau de PAX6, et les pathologies associées à une mutation de ce gène.

- Sur la page d'accueil de **STRING** (string-db.org), sélectionnez la modalité de recherche "**Protein by name**"
- Entrez *PAX6* pour **Protein name**
- Dans la boîte **Organisms**, tapez *Homo sapiens*. Quand vous commencerez la saisie, l'interface vous proposera la suite du mot .



STRING retourne une liste de protéines correspondant à votre recherche. Notez que plusieurs de ces protéines sont sélectionnées non pas en vertu de leur nom, mais parce que PAX6 apparaît dans leur description.

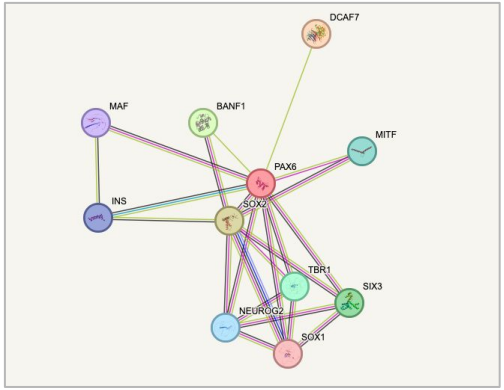


- Sélectionnez la protéine ***PAX6*** et cliquez sur "**CONTINUE**".
- Cliquez sur l'onglet **Analysis**.
- Ordonez les **False Discovery Rate** des processus biologiques en ordre croissant
- Observez les statistiques du sous-réseau de PAX6 et explorez les résultats de l'enrichissement fonctionnel.

Exercice 2 - Exploration du sous-réseau de voisinage de PAX6

Sur Ametice, répondez aux questions 16 à 18 du TP7

16. Ce sous-réseau comporte 11 molécules. Pourquoi ?
    a. PAX6 interagit avec 11 protéines
    b. PAX6 interagit avec 10 protéines
    c. C'est un effet des paramètres par défaut de STRING
17. Combien d'arêtes comporte ce sous-réseau ?
18. Quel est le processus biologique le plus significativement enrichi ?
    a. Cell fate commitment
    b. Eye development
    c. Double-stranded DNA binding
    d. Negative regulation of cellular biosynthetic process



Network Stats

number of nodes: 11
number of edges: 23
average node degree: 4.18
avg. local clustering coefficient: 0.835

expected number of edges: 13
PPI enrichment p-value: 0.00838
your network has significantly more interactions than expected (what does that mean?)

Functional enrichments in your network

Note: some enrichments may be expected here (why?)

explain columns

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0001654	Eye development	6 of 382	1.45	1.37	0.00037
GO:0002088	Lens development in camera-type eye	4 of 81	1.95	1.68	0.00041
GO:0001708	Cell fate specification	4 of 85	1.93	1.67	0.00041
GO:0045165	Cell fate commitment	5 of 250	1.55	1.44	0.00041
GO:0031327	Negative regulation of cellular biosynthetic process	8 of 1592	0.95	0.83	0.00055

(more ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0003690	Double-stranded DNA binding	9 of 1661	0.99	0.98	5.11e-05
GO:0000978	RNA polymerase II cis-regulatory region sequence-specific DNA bin...	8 of 1196	1.08	1.09	6.54e-05
GO:0001228	DNA-binding transcription activator activity, RNA polymerase II-spe...	6 of 458	1.37	1.45	8.49e-05
GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	8 of 1368	1.02	1.0	8.49e-05

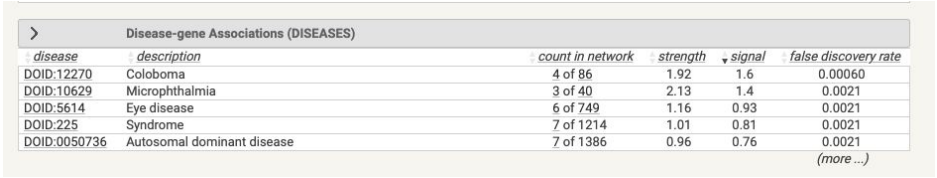
Cellular Component (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0000785	Chromatin	7 of 1285	0.99	0.8	0.0016
GO:0005694	Chromosome	8 of 1850	0.89	0.71	0.0016

Exercice 2 - Exploration du sous-réseau de voisinage de PAX6

Parmi les catégories d'annotations testées dans l'enrichissement fonctionnel, STRING propose aussi des annotations d'associations gènes-maladies extraites des bases de connaissances DISEASE and Monarch.

Dans les résultats d'enrichissement fonctionnel, explorez la section **Disease-gene Associations (DISEASE)**.



Disease-gene Associations (DISEASES)

disease	description	count in network	strength	signal	false discovery rate
DOID:12270	Coloboma	4 of 86	1.92	1.6	0.00060
DOID:10629	Microphthalmia	3 of 40	2.13	1.4	0.0021
DOID:5614	Eye disease	6 of 749	1.16	0.93	0.0021
DOID:225	Syndrome	7 of 1214	1.01	0.81	0.0021
DOID:0050736	Autosomal dominant disease	7 of 1386	0.96	0.76	0.0021

(more ...)

Sur Ametice, répondez aux questions suivantes.

19. Quelle est l'annotation maladie avec la valeur de "strength" la plus élevée ?
    a. Hypoplastic iris stroma
    b. Microphthalmia
    c. Eye disease
    d. Coloboma

Exercice 2 - Exploration du sous-réseau de voisinage de PAX6

- Dans le tableau **DISEASE**, cliquer sur "**Coloboma**": une pastille colorée apparaît dans le tableau. Noter que dans le sous-réseau de PAX6, les nœuds annotés avec "Coloboma" sont marqués de la même couleur.
- Dans le tableau **Monarch**, l'annotation *Coloboma* est également enrichie. Cliquer sur cette annotation. Les nœuds annotés avec *Coloboma* dans la base de données Monarch sont également colorés.
- Observer le sous-réseau.

Sur Ametice, répondez aux questions suivantes.

20. Combien de nœuds sont annotés comme étant liés au *Coloboma* à la fois dans DISEASE et Monarch ?
21. Quelle est la protéine liée au *Coloboma* dans une seule source d'annotation ?

Disease-gene Associations (DISEASES)

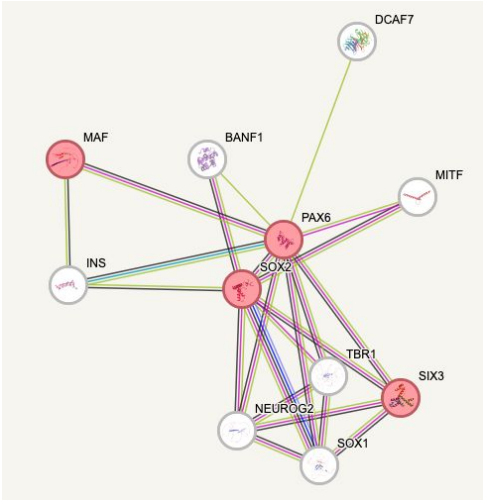
disease	description	count in network	strength	signal	false discovery rate
DOID:12270	Coloboma	4 of 86	1.92	1.6	0.00060
DOID:10629	Microphthalmia	3 of 40	2.13	1.4	0.0021
DOID:5614	Eye disease	6 of 749	1.16	0.93	0.0021
DOID:225	Syndrome	7 of 1214	1.01	0.81	0.0021
DOID:0050736	Autosomal dominant disease	7 of 1386	0.96	0.76	0.0021

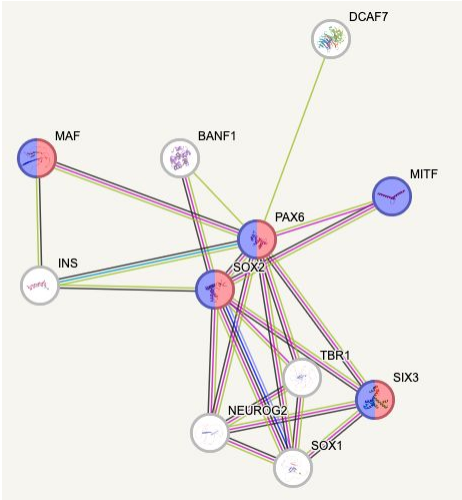
(more ...)

Human Phenotype (Monarch)

phenotype	description	count in network	strength	signal	false discovery rate
HP:0000589	Coloboma	5 of 205	1.64	1.42	0.00070
HP:0001488	Bilateral ptosis	3 of 48	2.05	1.12	0.0074
HP:0007990	Hypoplastic iris stroma	2 of 8	2.65	1.05	0.0124
HP:0001104	Macular hypoplasia	2 of 9	2.6	1.05	0.0124
HP:0000525	Abnormality iris morphology	5 of 352	1.41	0.99	0.0049

(more ...)





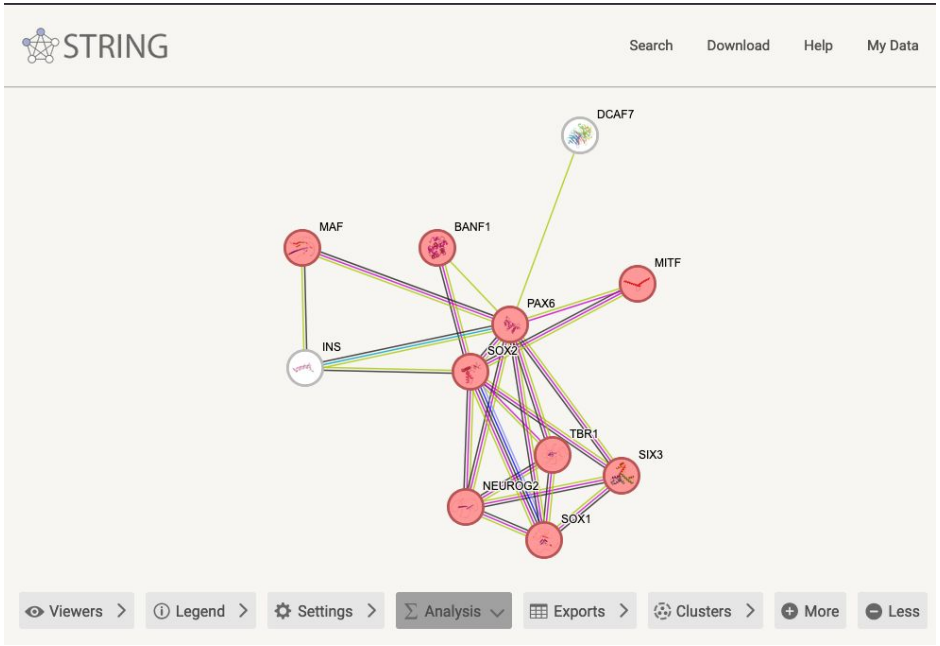
Exercice 3 - Identifications des variants de PAX6 associés aux maladies des yeux

Exercice 3 - Identification des variants de PAX6 associés aux maladies des yeux

L'analyse du sous-réseau de PAX6 nous a permis de découvrir que la protéine produite par ce gène interagit avec d'autres protéines impliquées dans des maladies des yeux, et dont la fonction moléculaire est principalement la liaison à l'ADN double-brin*. En effet, la majorité des protéines du sous-réseau sont des facteurs de transcription, comme c'est le cas de PAX6.

Ceci pourrait suggérer que des variants liés à une ou plusieurs des ces maladies, comme le *Coloboma* ou la *Microphthalmia*, pourraient avoir un impact sur la fonction moléculaire de PAX6 et de ses voisins.

L'objectif de ce dernier exercice est d'identifier la présence d'un ou plusieurs variants de PAX6 associés à des maladies, et de faire le lien avec un impact potentiel sur sa fonction.

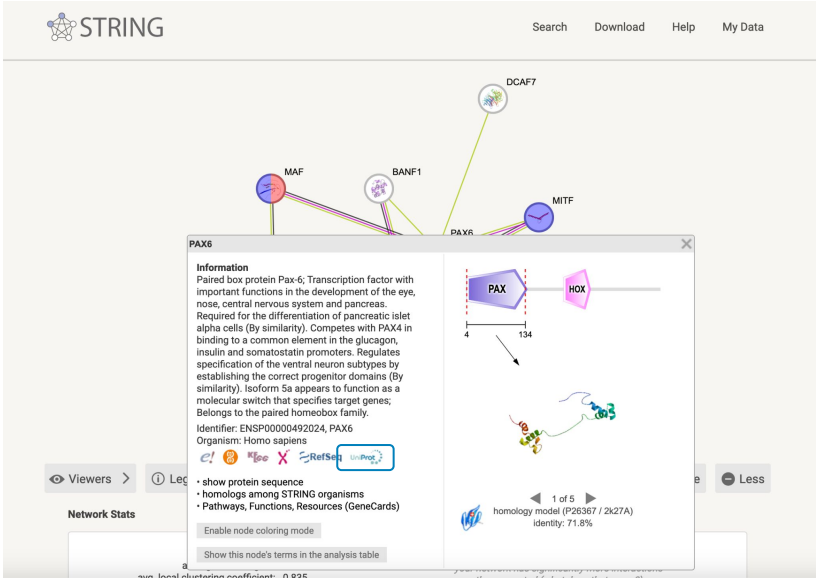


Molecular Function (Gene Ontology)					
GO-term	description	count in network	strength	signal	false discovery rate
GO:0003690	Double-stranded DNA binding	9 of 1661	0.99	0.98	5.11e-05
GO:0000978	RNA polymerase II cis-regulatory region sequence-specific DNA bin...	8 of 1196	1.08	1.09	6.54e-05
GO:0001228	DNA-binding transcription activator activity, RNA polymerase II-spe...	6 of 458	1.37	1.45	8.49e-05
GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	8 of 1368	1.02	1.0	8.49e-05

** Cette affirmation s'appuie sur les résultats de l'enrichissement fonctionnel présenté dans le tableau "Molecular Function". Voir capture d'écran ici à droite.*

Exercice 3 - Identification des variants de PAX6 associés aux maladies des yeux

- Cliquez sur le nœud PAX6
- Lisez les informations relatives à ses fonctions
- Cliquez sur l'icone ***UniProt*** pour ouvrir le lien dans un nouvel onglet du navigateur
  Astuce: cliquez sur le logo en appuyant la touche *Ctrl* sur le clavier, sur le ordinateur Mac, il faut appuyer sur la touche *command*.



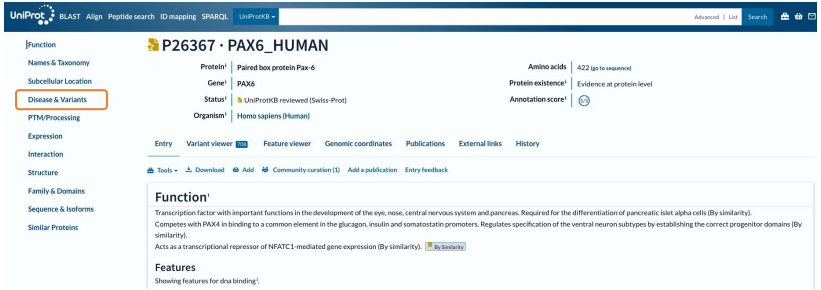
Cela vous permettra de visualiser la page de la proteine PAX6 (identifiants: P26367, PAX6_HUMAN)

- Sur le menu à gauche cliquer sur ***Disease & Variants***.

Cette section fournit des informations sur les maladies, les phénotypes et les variants associés à une protéine.

Sur Ametice, répondez aux questions suivantes.

22. Dans la sub-section "*Involvement in disease",* à combien de maladies PAX6 est-il associé ?
23. Y a-t-il une ou plusieurs de ces maladies qui ne soit pas liée à l'œil ?

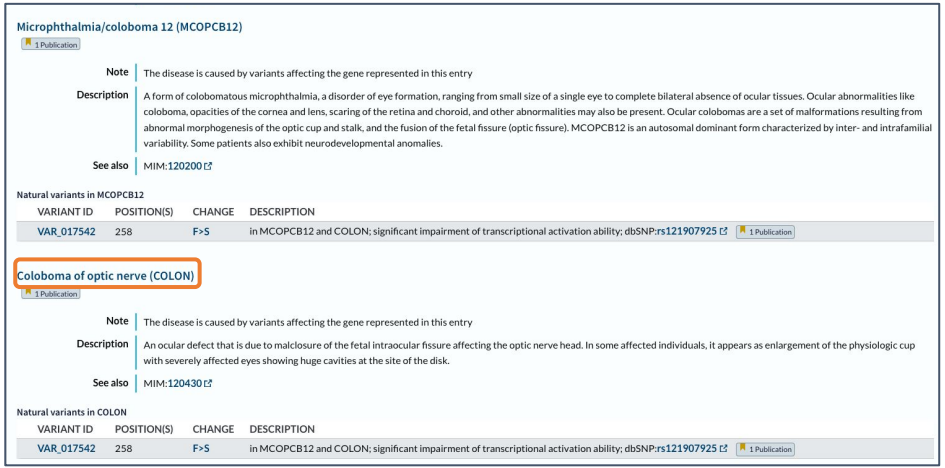


Exercice 3 - Identification des variants de PAX6 associés aux maladies des yeux

- Parmi les maladie associée à PAX6, cherchez ***Coloboma of optic nerve (COLON)***
- Lisez la description de cette maladie et vérifiez si il y a un ou plusieurs variants dans PAX6 associés à ***COLON***.

Sur Ametice, répondez aux questions suivantes.

24. Existe-t-il un variant de la séquence PAX6 associé au Coloboma ?
    a. Oui, il y a un variant qui cause une changement d'acide aminé (une phénylalanine, F, à la place d'une sérine, S) sur la position 258 de PAX6.
    b. Oui, il y a un variant qui cause une changement d'acide aminé (une sérine, S, à la place d'une phénylalanine, F) sur la position 258 de PAX6.
    c. Non, il n'y a pas de variant associé.
25. Le cas échéant, quel est l'identifiant du SNP de ce variant ?
    a. rs121907919
    b. rs121907928
    c. rs121907926
    d. rs121907925
26. En lisant la description associé à ce variant, pouvez-vous conclure qu'il a bien un impact sur la fonction de PAX6 ?



Microphthalmia/coloboma 12 (MCOPCB12)
1 Publication

Note	The disease is caused by variants affecting the gene represented in this entry
Description	A form of colobomatous microphthalmia, a disorder of eye formation, ranging from small size of a single eye to complete bilateral absence of ocular tissues. Ocular abnormalities like coloboma, opacities of the cornea and lens, scaring of the retina and choroid, and other abnormalities may also be present. Ocular colobomas are a set of malformations resulting from abnormal morphogenesis of the optic cup and stalk, and the fusion of the fetal fissure (optic fissure). MCOPCB12 is an autosomal dominant form characterized by inter- and intrafamilial variability. Some patients also exhibit neurodevelopmental anomalies.
See also	MIM:120200

Natural variants in MCOPCB12

VARIANT ID	POSITION(S)	CHANGE	DESCRIPTION
VAR_017542	258	F>S	in MCOPCB12 and COLON; significant impairment of transcriptional activation ability; dbSNP:rs121907925 1 Publication

Coloboma of optic nerve (COLON)
1 Publication

Note	The disease is caused by variants affecting the gene represented in this entry
Description	An ocular defect that is due to malclosure of the fetal intraocular fissure affecting the optic nerve head. In some affected individuals, it appears as enlargement of the physiologic cup with severely affected eyes showing huge cavities at the site of the disk.
See also	MIM:120430

Natural variants in COLON

VARIANT ID	POSITION(S)	CHANGE	DESCRIPTION
VAR_017542	258	F>S	in MCOPCB12 and COLON; significant impairment of transcriptional activation ability; dbSNP:rs121907925 1 Publication

Conclusions du TP

Au cours de ce TP nous avons utilisé la base de données STRING-DB, qui rassemble des informations sur les interactions fonctionnelles entre protéines ou gènes d'un organisme.
Ceci nous a amenés à définir des **concepts de la théorie des graphes** (degré, coefficients de centralité et de regroupement, cliques, chemins, voisinage) qui permettent de caractériser les **propriétés topologiques des réseaux**.
Nous avons également appliqué une **analyse d'enrichissement** fonctionnel pour tenter d'inférer les propriétés biologiques (processus biologique, lien avec des pathologies humaines) à partir du sous-réseau extrait au voisinage d'une protéine d'intérêt.
Ce type d'analyse est mis à contribution pour analyser les résultats d'analyses omiques de différents types (protéomique, transcriptomique, interactomique, métabolomique, …) afin de pouvoir interpréter les données massives produites par les technologies correspondantes.