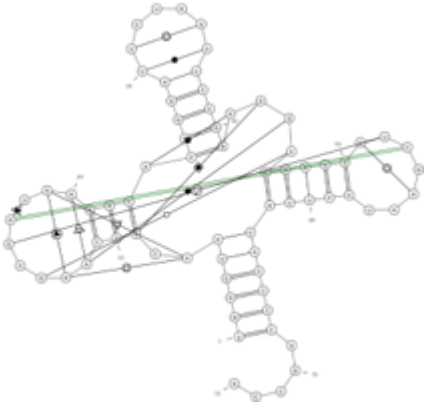
TD de Bioinformatique Structurale : Encodage des structures sous forme de graphe d'interactions et recherche de motifs

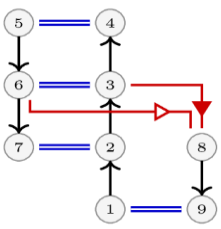
On définit la structure secondaire d'un ARN comme étant un graphe, les nucléotides formant les nœuds, et les interactions entre nucléotides les arêtes du graphe. Les interactions entre nucléotides incluent les liaisons phosphodiester qui unissent chaque nucléotide à ses deux voisins.
Parmi les autres interactions, on distingue notamment les interactions canoniques (formées par des liaisons hydrogène entre les cotés « Watson-Crick » de A et U ou de G et C, ou parfois de G et U).



Une nomenclature plus vaste, dite de « Leontis-Westhof », décrit 12 types d'interactions plus variées, incluant les canoniques mais décrites comme les « non-canoniques » (oui c'est paradoxal).
On les distingue sur les graphes de structure secondaire avec des petits symboles. Référez vous à l'article : « *Leontis, N. B. et Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs.RNA, 7(4):499–512* » pour comprendre ces notations.

Pour chaque chaine d'ARN, la base de données RNANet contient les positions et types d'interactions entre nucléotides de la chaine (et certaines avec des nucléotides extérieurs à la chaine, il faudra les ignorer).
Par exemple, pour un fichier « chaine d'ARN » de RNANet fourni au format CSV (aussi appelé « datapoint »), chaque ligne correspond à un nucléotide, et on a les champs :
`index_chain` : donne le numéro du nucléotide dans la chaine d'ARN, de 1 à N si N est la longueur de la chaine.
`paired` : donne une liste de valeurs numériques correspondant aux `index_chain` des nucléotides appariés avec le nucléotide en question. Si le nucléotide n'est pas apparié, on a la valeur `NaN`.
`pair_type_LW` : donne une liste d'étiquettes (de même taille que `paired` et ordonnée de la même façon) indiquant le type d'interaction que forment les deux nucléotides.

1) Créez un programme capable de construire un graphe modélisant la chaine d'ARN à partir du fichier CSV.
1.5) Appliquez votre code à tous les fichiers CSV pour charger tous les ARN de la base de données.



2) Ensuite, choisissez un ou plusieurs motifs d'ARN sur le site http://carnaval.lri.fr/Networks/all.html .
Le site propose des « RIN » pour Recurrent Interaction Networks. Il s'agit de sous-graphes que l'on souhaite identifier dans les structures d'ARN.

Utilisez un algorithme de recherche de sous-graphe pour identifier les structures d'ARN contenant le (ou les) motif(s) d'ARN que vous aurez choisi sur CaRNAval.

Exercice à réaliser en binôme ou trinôme, avant le 5 Mars 2021.
Le rendu comprendra :

- Une archive ou dépôt de code, code qui devra être commenté et contenir des instructions d'installation.
- Un rapport court expliquant le fonctionnement de votre algorithme.

Contact et questions : louis.becquey@univ-evry.fr ; fariza.tahi@univ-evry.fr