Titre : Fun(gi) with codons. À quelles règles obéit l'organisation des codons ?

Encadrantes : Patricia Thébault (LaBRI), Raluca Uricaru (LaBRI) et Karine Dementhon (Laboratoire Microbiologie Fondamentale et Pathogénicité, Université de Bordeaux)

Mots clé : fungi, Candida, bioinformatique, data mining

Sujet :

Chaque année dans le monde, près de 1,5 million de personnes mourront d'une infection fongique invasive malgré la disponibilité de médicaments antifongiques. Cela représente plus de décès que ceux causés par la tuberculose ou le paludisme. Les médicaments antifongiques sont limités et des souches résistantes émergent. Aucune nouvelle molécule antifongique n'a été approuvée depuis 2006, soulignant la nécessité de mieux comprendre les mécanismes qui sous-tendent la pathogénicité de certains *fungi*, afin d'identifier de nouvelles cibles thérapeutiques potentielles. Au sein de ces pathogènes, les levures *Candida* représentent 70 à 90 % de toutes les infections fongiques invasives (dont le coût a atteint 1,5 milliard d'euros aux États-Unis en 2017), les candidoses invasives étant associées à un taux de mortalité dépassant 50% dans les unités de soins intensifs.

Plus de 200 espèces de *Candida* existent parmi les centaines d'espèces de levure (environ 1500), cependant, seule une douzaine est rapportée comme étant des pathogènes opportunistes humains. Notablement, les espèces de levure sont phylogénétiquement éloignées, et présentent une physiopathologie et des traits phénotypiques différents, ce qui soulève la question de l'existence d'un noyau commun de " traits de pathogénicité " qui pourraient représenter de nouvelles cibles antifongiques intéressantes.

Dans ce contexte, à partir de plus de 300 génomes de levures (dont 18 *Candida*) récupérés de 4 fournisseurs de données (NCBI, JGI, iGenolevures et RIKEN), une procédure informatique a été implémentée et appliquée afin d'extraire les régions codantes (CDS), d'identifier les codons pour chaque CDS et calculer certaines métriques en lien avec la composition des génomes et avec les codons. Le nombre de métriques étudiées se basent sur une littérature abondante et un total de 26 mesures ont été calculées allant de la simple mesure sur la composition en codons (GC% du génome) à des mesures plus sophistiquées pour prendre en compte les codons optimaux (par exemple le CAI Codon adaptation Index) ou encore le contexte (ex : Index of Preferred / Avoided Contexts, Balance of Indexes). Ces métriques ont été calculées à différentes échelles : au niveau de l'espèce et donc du génome, ou bien au niveau des CDS. Les valeurs obtenues sont organisées en différents tableaux qui fournissent un jeu de données original, extrêmement bien structuré et riche, avec la particularité d'être très propre et sans valeurs manquantes.

C'est ce jeu de données (orienté vers la composition en codons, et non pas vers la séquence comme c'est le cas habituellement), qui représente le point de départ de ce projet ayant comme but d'identifier du signal permettant la comparaison de ces différentes espèces. Le code génétique est classiquement étudié sur la base des tables de correspondances entre codons (triplets de bases nucléiques) et acides aminés. Ce code est universel mais peut comprendre certaines différences selon les espèces. En effet, chaque espèce peut en faire un usage différent, par exemple certains codons seront sur-représentés ou vice versa. L'étude globale des différences de l'usage des codons permettrait donc d'adresser de façon originale des questions en évolution des espèces et

notamment d'identifier des caractéristiques en fonction de groupes d'espèces (par exemple le groupe d'espèces pathogènes), des familles de gènes, du core-génome…
Concrètement, dans la suite du projet, et notamment grâce au future stagiaire M2, nous comptons développer (en utilisant les langages R et python), des approches bioinformatiques dédiées à la fouille de ces données de composition en codons, visant à réaliser :

- l'identification de *features* (i.e. caractéristiques, dans notre cas les différentes métriques) en se basant sur des méthodes de réduction de dimensions pour les données multivariées, combinées à des méthodes de sélection de *features* comme par exemple le sparse ACP [1] et NMF[2]. Cela permettrait par exemple d'identifier les métriques les plus adaptées pour distinguer un ensemble d'espèce (par exemple le groupe des pathogènes versus les non-pathogènes);
- le regroupement en familles de gènes ou en groupes d'espèces sur la base de leur phénotype, par des approches de clustering et en utilisant notamment des approches d'*ensemble clustering* [3] qui permettent l'intégration de multiples résultats de partitionnement exploitant diverses caractéristiques.

Ce projet interdisciplinaire a démarré il y a un an (encadrement de stage de M1) et réunit deux partenaires situés à Bordeaux au sein de deux départements distincts : Sciences biologiques et médicales et Sciences de l'ingénierie et du numérique (SIN). Chaque partenaire a une spécialité majeure telle que l'informatique/bioinformatique (R. Uricaru), ou la biologie des Candida (K. Démenthon). Les deux partenaires participeront à l'encadrement du stage de recherche en bioinformatique qui se déroulera au sein du LaBRI à partir du mois de janvier, et plus précisément dans l'équipe BKB (Bench to Knowledge and Beyond : https://www.labri.fr/bench-knowledge-and-beyond) dont les travaux sont centrés sur l'analyse de données (biologiques) massives.

[1] Shen H and Huang JZ, Sparse principal component analysis via regularized low rank matrix approximation, Journal of Multivariate Analysis, 2008
[2] Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics. 2016
[3] Guoxian Yu, Yuan Jiang, Jun Wang, Hao Zhang and Haiwei Luo, BMC3C: binning metagenomic contigs using codon usage, sequence composition and read coverage, Bioinformatics, 2018.