# *Méthodes pour la reconstruction du répertoire des transcrits d'un gène à partir de données RNA-seq de 3ème génération*

***Encadrement***

**VARRÉ Jean-Stéphane**
*jean-stephane.varre@univ-lille.fr*

***Résumé***

La recherche en bioinformatique a été dynamisée par les technologies de séquençage permettant d'obtenir de nombreuses données génomiques à faible coût. Ces avancées technologiques ont aussi permis de dynamiser la recherche en transcriptomique qui a pour objectif l'étude de l'ensemble des ARN présents dans un milieu : lesquels sont présents, en quelle quantité, y a-t-il de nouvelles isoformes ? L'intérêt de l'étude des ARN réside dans le fait qu'il est désormais admis que le mécanisme d'épissage alternatif, la capacité d'un gène à produire différents ARN, qu'on nommera transcrits isoformes, est le mécanisme majeur permettant une diversité de protéines produites dans la cellule.

Il est donc crucial de pouvoir répondre à la question de l'identification des transcrits isoformes, à la fois pour l'étude de leur diversité mais également parce que des dérèglements de l'épissage sont en cause dans un certain nombre de maladies, dont des cancers.

L'objectif de la thèse est d'appréhender la diversité de transcrits isoformes à partir d'expériences de séquençage dites de 3ème génération. Des travaux existent déjà autour de cette problématique mais dont l'objectif est une analyse à gros grain de l'ensemble de tous les ARN séquencés. L'angle de la thèse est d'aborder un problème plus précis d'analyse d'un ou de quelques gènes cibles permettant le développement de méthodes plus fines et plus appropriées à l'étude précise des isoformes de transcrits de gènes identifiés impliqués dans des maladies.

***Contexte scientifique***

Le sujet se positionne dans le domaine de la bioinformatique et plus précisément de la transcriptomique. Avec pour objectif de contribuer par une méthode d'analyse de données de séquençage RNA-seq à la connaissance de la diversité de transcrits au sein d'une cellule ou d'un milieu.

La transcription est la première étape du processus d'expression des gènes qui convertit un segment génomique (un gène) d'un génome en un ARN (ARNm). L'épissage alternatif est un mécanisme capable de fabriquer différents ARN, des transcrits isoformes, à partir d'un même segment génomique en joignant des parties appelées exons (par opposition aux introns) (boîtes sur la Figure 1). L'étape suivante consiste à traduire l'ARN en protéine. L'épissage alternatif permet de produire plusieurs types d'isoformes de protéines à partir d'un même segment génomique. Il permet ainsi de générer un certain nombre de protéines à partir d'un nombre limité de gènes. L'épissage alternatif est très répandu chez les eucaryotes multicellulaires, et joue un rôle important dans le développement des tissus et des espèces [1]. Il est observé à la fois chez les animaux et les plantes. L'épissage alternatif joue également un rôle important dans les maladies et notamment le cancer [4, 9] (Figure 1). Il est donc important de pouvoir décrire et comparer le répertoire des transcrits isoformes de différents milieux ou dans différents tissus ou dans des conditions expérimentales différentes.

Le RNA-seq (séquençage des ARN) permet d'obtenir la séquence des molécules d'ARN, dans n'importe quel organisme séquencé ou type de cellule unique, dans différentes conditions expérimentales. Cependant, il reste difficile de décrire l'ensemble du répertoire des transcrits pour un gène donné. Des difficultés subsistent en raison de la faible longueur des lectures (segment d'ADN séquencé) issues du séquençage de deuxième génération (NGS). En effet cela empêche d'obtenir des structures exoniques complètes [10] puisque que la reconstruction des ARN complets nécessite une étape d'assemblage des reads. Mais, avec le séquençage de troisième génération (TGS), on peut obtenir un ARN (presque) complet [8]. En supprimant l'étape d'assemblage on obtient sans ambiguïté la séquence de chaque transcrit séquencé permettant de gagner en précision sur la prédiction de la structure exonique.

Des technologies comme ONT Nanopore permettent de réaliser rapidement et à faible coût des expériences RNA-seq, notamment en utilisant des séquenceurs très compacts et portables tel que MinION. Ils ont l'inconvénient de produire des lectures avec un taux d'erreur assez important (autour de 8% à l'heure actuelle) et de tronquer les ARN séquencés. Elles restent néanmoins un outil de choix et les problèmes technologiques peuvent trouver des réponses méthodologiques dans le développement d'outils algorithmiques dédiés. Un certain nombre de méthodes pour l'analyse RNA-seq ont été conçues pour les NGS [5, 6], mais les méthodes pour les TGS, en particulier pour la technologie ONT, nécessitent encore des développements.

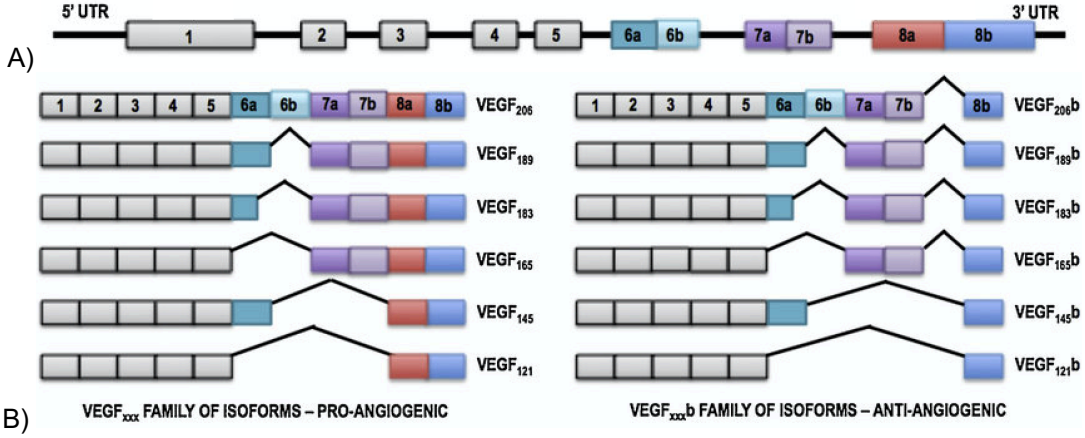


Figure 1 : A) exemple schématique de la structure exonique du gène VGEF impliqué dans la régulation de l'angiogénèse (formation de nouveaux vaisseaux sanguins), le gène est constitué de 11 exons. B) 12 transcrits isoformes. L'inclusion de certains exons modifie la fonction, favorisant ou non l'angiogénèse. Les isoformes VEGFxxxb sont habituellement régulés dans les tissus humains et dérégulés dans de nombreuses maladies, cancer compris. Tiré de [15].

Nous posons donc la problématique suivante. Etant donné une expérience RNA-seq TGS, un ou quelques gènes d'intérêt, sélectionner les lectures correspondantes et décrire précisément la structure exonique (c'est-à-dire les exons, les jonctions d'épissage) de celles-ci, en se donnant la possibilité d'intégrer des informations connues (séquence génomique, transcrits connus, protéines connues) de l'espèce ou d'une autre espèce.

Répondre à cette problématique permettra d'aborder des questions sur la diversité de transcrits au sein d'un milieu (microbiote par exemple). Mais un intérêt socio-économique majeur de ces recherches consiste en l'étude de gènes cibles pour les maladies. La possibilité de comparer rapidement et à l'échelle de l'individu les isoformes de transcrits peut permettre d'avancer sur les techniques de diagnostic ou de traitement personnalisé.

### *Contexte local*

Le sujet proposé sera développé au sein de l'équipe Bonsai de l'UMR CRIStAL 9189. Bonsai regroupe 9 membres permanents (1 DR CNRS, 1 PR, 2 CR CNRS, 4 MCF dont 1 HDR, 1 IR CNRS).

Bonsai est une équipe de recherche en bioinformatique dont une des thématiques principales est la conception d'algorithmes pour l'exploitation de données de séquençage, en particulier de 3ème génération. Elle bénéficie d'une reconnaissance internationale pour ses travaux sur les algorithmes et les structures de données pour l'analyse de données génomiques.

Sur la thématique de l'épissage alternatif, des travaux ont été menés depuis plusieurs années, notamment sur l'exploitation des données de séquençage de seconde génération. Plus récemment l'équipe Bonsai a été coordinatrice du projet ANR ASTER 2016-2020 (Algorithmes et outils logiciels pour le séquençage d'ARN de troisième génération) regroupant le Génoscope (Evry), le LBBE (Lyon) et l'Institut Pasteur de Lille. Des travaux sur la comparaison et la prédiction d'isoformes transcrits à partir de données d'autres espèces (sans recours à une expérience de séquençage) ont également été menés.

### *Travail à réaliser*

Ces dernières années, nous avons travaillé sur une méthodologie permettant de comparer le répertoire de transcrits isoformes entre deux ou plusieurs espèces [2, 7], démontrant que la génomique comparative peut aider à découvrir des transcriptions alternatives pour une espèce proche. Nous nous sommes ensuite penchés sur le problème de la description de la structure exonique d'un gène donné à partir d'un sous-ensemble de TGS sélectionnés, en ne connaissant que sa séquence génomique [3]. La connaissance de la séquence génomique permet, par comparaison des lectures et de cette séquence, de sélectionner un sous-ensemble pertinent de lectures formant un ensemble putatif d'ARN correspondants à ce gène. Une fois ce sous-ensemble construit, des étapes de comparaison plus précises et de filtrage permettent de prédire une structure exonique du gène.

L'objectif de la thèse est d'explorer de manière plus large l'ensemble des données connues qui peuvent être fournies en plus des lectures RNA-seq pour décrire la structure exonique d'un ou plusieurs gènes. Nous nous intéresserons à deux aspects qui se complèteront.

D'une part, au-delà de la connaissance de la séquence génomique nous intégrerons la connaissance de transcrits ou de protéines du gène en question pour étayer les prédictions de structure exonique. Nous explorerons des méthodes permettant d'intégrer l'un ou l'autre des données ou une combinaison de celles-ci. Une des difficultés dans l'intégration de transcrits connus réside dans l'absence de signaux permettant de connaître la position des jonctions intron/exon. L'intégration de données protéiques pose également des difficultés méthodologiques eu égard à la diversité des isoformes de transcrits pouvant produire la même isoforme et à la comparaison protéine-ARN. Il s'agira à chaque fois de tirer parti des connaissances actuelles sans restreindre la méthode afin de la garder apte à découvrir de nouveaux évènements (par exemple de nouveaux exons ou des exons tronqués, ce qui est par exemple le cas de l'exon 8 de VGEF, Figure 1). Enfin nous étendrons le travail à la prédiction sans connaissance a priori.

D'autre part, en se plaçant dans un contexte d'étude de gènes pour des espèces pour lesquelles on ne dispose pas d'information a priori (pas de séquençage du génome par exemple, ou information parcellaire parce que l'espèce est non modèle), on s'intéressera à la manière dont on peut projeter le connaissance d'autres espèces proches sur cette espèce. Cette approche de génomique comparative pourra également aider à découvrir des variants non connus de l'espèce d'intérêt (par exemple l'homme) en utilisant la connaissance d'autres espèces.

Outre les résultats méthodologiques, un des livrables de la thèse sera la mise en place d'un outil d'analyse, rendu disponible gratuitement à la communauté, permettant à un biologiste ou un bioinformaticien d'analyser ses données.

***Calendrier prévisionnel***

Suivant le cursus du candidat, le présent programme de travail pourra être adapté, en appuyant plus l'état de l'art sur la compréhension des mécanismes biologiques ou la compréhension des technologies ou des méthodes bioinformatique. L'aspect génomique comparative sera envisagé pour chaque type de données intégrées (tâches 5, 6 et 7).

Tâche 1 (M1 à M4) : état de l'art sur RNA-seq 3$^{ème}$ génération
Livrable : rapport sur l'état de l'art

Tâche 2 (M1 à M4) : état de l'art sur l'épissage alternatif
Livrable : rapport sur l'état de l'art

Tâche 3 (M3 à M5) : prise en main des outils usuels de traitement de données de lectures 3$^{ème}$ génération
Livrables : rapport sur les outils ; installation d'un environnement de travail

Tâche 4 (M3 à M10 et au fil de l'eau ensuite) : construction de jeux de données
Livrables : collecte de jeux de données RNA-seq ; identification de gènes d'intérêt

Tâche 5 (M4 à M14) : intégration de données de transcrits
Livrables : algorithme / pipeline d'analyse ; publication sous forme de poster dans une conférence nationale ; intégration dans le logiciel

Tâche 6 (M11 à M20) : intégration de données de protéines
Livrables : algorithme / pipeline d'analyse ; publication dans une conférence ; intégration dans le logiciel

Tâche 7 (M20 à M23) : intégration de données hétérogènes (génome, transcrits et protéines)
Livrables : logiciel ; publication dans un journal

Tâche 8 (M23 à M29) : analyse sans connaissance a priori
Livrables : logiciel ; publication dans un journal

Tâche 9 : rédaction du manuscrit de thèse
Livrables : manuscrit de thèse ; soutenance de thèse

***Profil recherché***

Étudiant.e en master 2 (ou équivalent) en informatique ou bioinformatique avec une bonne maîtrise de l'algorithmie et de la programmation. Une expérience dans le développement d'outils pour la bio-informatique et/ou la chemo-informatique est souhaitée.

***Bibliographie***


[1] Baralle, F. E., and Giudice, J. Alternative splicing as a regulator of development and tissue identity. Nature reviews. Molecular cell biology 18, 7 (July 2017), 437–451.

[2] Blanquart, S., Varré, J.-S., Guertin, P., Perrin, A., Bergeron, A., and Swenson, K. M. Assisted transcriptome reconstruction and splicing orthology. BMC Genomics 17, Suppl 10 (2016), 786.

[3] Borée, C., Antoine-Lorquin, A., Varré, J.-S., and Touzet, H. My RNA-tailor is rich : fine modeling of alternative transcripts from long reads, 2020. Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM).

[4] Cheng, S., Ray, D., Lee, R., Naripogu, K. B., Yusoff, P., Goh, P., Liu, Y., Suzuki, Y., Das, K., Chan, H. S., Wong, W. K., Chan, W. H., Chow, P. K.-H., Ong, H. S., Raj, P., Soo, K. C., Tan, P., Epstein, D. M., and Rozen, S. G. A functional network of gastric-cancer-associated splicing events controlled by dysregulated splicing factors. NAR Genomics and Bioinformatics 2, 2 (2020). lqaa013.

[5] Denti, L., Rizzi, R., Beretta, S., Della Vedova, G., Previtali, M., and Bonizzoni, P. ASGAL: aligning RNA-Seq data to a splicing graph to detect novel alternative splicing events. BMC Bioinformatics 19, 1 (2018), 444.

[6] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. Bioinformatics (Oxford, England) 29, 1 (2012), 15–21.

[7] Guillaudeux, N., Belleanne´e, C., Blanquart, S., and Varré J.-S. Predicting isoform transcripts: What does the comparison of known transcripts in human, mouse and dog tell us?, 2019. Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM).

[8] Hardwick, S. A., Joglekar, A., Flicek, P., Frankish, A., and Tilgner, H. U. Getting the Entire Message: Progress in Isoform Sequencing. Frontiers in Genetics 10 (2019), 11706.

[9] Mavrou, A., and Oltean, S.SRPK1 inhibition in prostate cancer: A novel anti-angiogenic treatment through modulation of VEGF alternative splicing. Pharmacological Research 107 (2016), 276–281.

[10] Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Hubbard, T. J., Guigo, R., Harrow, J., and Bertone, P. Assessment of transcript reconstruction methods for RNA-seq. Nature methods 10, 12 (2013), 1177–1184.