2

L’ARNs non codant joue un rôle vital dans beaucoup de processus cellulaire comme l’épissage de l’ARN, la translation ou la régulation des gènes. Hors la plupart des ARNs non codant ne sont pas annotés fonctionnellement.

Une des approches les plus connues pour assigner des fonctions putative est le clustering des transcrits selon la séquence et la structure secondaire. Pas le mieux car la séquence d’information est changée par les modifications post-transcriptions, et la structure secondaire est seulement un proxy un indicateur une prédiction pour la conformation 3D de l’ARN polymer.

Un type d’informations différent qui n’a pas le problème vu précédemment et qui peut être utilisé pour la détection des classes d’ARN, est le « pattern of processing » et ses traces dans les ‘reads data’.

BlockClust est une approche efficace pour la détection de transcrit avec le même système de fonctionnement. Les auteurs propose un nouveau moyen d’encoder les profils d’expression en structures discrètes compactes, qui peuvent alors être traitées en utilisant la technique du graph-kernel. Ils réalise à la fois un clustering non supervisé et développent des models discriminatoires spécifique aux familles. Et finalement ils montrent comment l’approche est précise, modulable et robuste selon les différents organismes, tissus et variétés de cellule.

Rappelons la méthode expérimentale : on récolte les ARN de nombreuses cellules à la fois, on les coupe en fragments qu'on amplifie par [PCR](http://fr.wikipedia.org/wiki/R%C3%A9action_en_cha%C3%AEne_par_polym%C3%A9rase) et qu'on donne à manger à un séquenceur. Celui-ci nous retourne de petits bouts de séquence génomique, des "reads", sous forme de texte avec des A,T,G,C. Vous héritez de dizaines de millions de ces petites phrases.

On part du principe que le nombre de reads est proportionnel à l'abondance des ARN correspondants dans la cellule, le but étant d'estimer cette abondance.

Le RNA-seq est une technique relativement récente et faire partie de ce qu'on appelle "[séquençage de seconde génération](http://bioinfo-fr.net/le-sequencage)" (next-generation sequencing) ou "séquençage à haut débit" (high-throughput sequencing), avec le [ChIP-seq et ses contemporains](http://bioinfo-fr.net/dnase-seq-faire-seq-chip-seq-3-outils-danalyse-de-la-regulation-de-lexpression-des-genes).

3

L’idée centrale de la méthode BlockClust est de caractériser les locis transcrits en utilisant les profils d’expression obtenues à partir d’un protocole expérimentale de séquençage. On extrait les attributs caractéristiques à partir des profils d’expressions.

On code alors la séquence de plusieurs attributs en structures compactes discrètes, que ns avons ensuite traités en utilisant un "graph-kernel« .

Voir doc Sup graph-kernel?

Les deux points essentiels de BlockClust sont :

- encodé les profile d’expression avec des attributs discretisés

- génération de caractéristiques combinatoires à partir de la séquence d’attributs.

4

Avec le terme profil d’expression on désigne l’ensemble de read sequences assemblés relative à un transcrit donné. Afin d’extraire ces profils, il est nécessaire d’aligner les ‘reads’ contre la référence génomique qui leurs correspond afin d’obtenir leurs coordonnées chromosomique.

L’information sur les reads alignés est

5

Dans BlockClust nous n'utilisons pas des techniques basées sur alignement pour comparer  
les groupes de blocks, car le temps de calcul serait trop long.  
Au lieu de ça on extrait des caractères explicits qui pourra etre utilisé plus efficacement. les caractéristiques considérées sont celles développées pour un ‘graph kernel’ appelé Neighborhood Subgraph Pairwise Distance Kernel (NSPDK).

Le shema suivant montre comment sont générés les caractères avec un rayon et une distance donnée.

6

La notion de similarité de NSPDK peut être employée directement en des algorithmes de clustering,  
cela se servent de l'information de similitude ou de distance par paires.

Comme un algorithme de clustering BlockClust emploie l’algorithme MCL (Markov Cluster Process). Donné pesé le graph G des poids des plus proche voisins entre les exemples à grouper, l'algorithme de MCL applique un processus algébrique paramétré à la matrice de ‘random walks’=‘mesures de proba’ sur  G. L'idée sous-jacente est de caractériser des groupes comme sous-graphes  
tels qu'une marche aléatoire sur le graphique ira rarement d'un sous-graphe à l'autre. Le MCL a été choisi comme il produit équilibré les groupes non hiérarchiques et lui ne fait ni l'un ni l'autre l'information de ensemencement du besoin ni un nombre défini par l'utilisateur de groupes. D'ailleurs il peut être utilisé dedans arrangements à grande échelle comme il peut fonctionner avec des réalisations clairsemées de graphique/matrice.

Dans notre arrangement d'application, le paramètre d'inflation, qui affecte groupez la granularité, a été sélectionné pour maintenir les groupes relativement petits.

7

En plus du clustering non supervisé, BlockClust fournit un mode de classification supervisé. Donné un ensemble deprofils d'expression pour connu famille de ncRNA ou classe et un ensemble d'exemples négatifs, c.-à-d. expression profils des ncRNAs avec une fonction différente ou inconnue, BlockClust peut efficacement établir un classificateurbinaire linéaire distinctif. Comme dans  
mode de groupement non surveillé, nousextrayons d'abord haut-dimensionnel explicite représentations de vecteur des codages deprofil d'expression. Plus tard BlockClust emploie des techniques linéaires rapides et extensibles comme les machines stochastiques de vecteur de soutien de descente de gradient (Bottou, 2010) pour induire un modèle distinctif.Notez cela même si nous employez les modèles linéaires pour laisser mesurer aux arrangements larges de données de génome, résulter le classificateur est en fait non linéaire dansl'espace original d'attribut.  
Les modèles en résultant sont précis etétonnant robustes : un modèle pour l'identification des gènes de tRNA peut être formée sur des données humaines avec lit extrait sous un protocole expérimental spécifique (dites Illumina GAII) et lui peuvent être employés pour annoter sûrement des profils d'expression à travers organismes divers (par exemple mouche ou usines), des données produites par différent expérimental

8