Algorithme EM

Philippe Veber

18 février 2020

1 Quantification des isoformes en RNA-seq

En RNA-seq, on cherche à quantifier l'expression des gènes en séquençant les ARNs messagers après extraction dans un échantillon d'intérêt. En comptant le nombre de lectures provenant de chaque messager, on peut ainsi évaluer l'abondance *relative*[1] de chacun d'eux. Le type de séquençage encore massivement utilisé en 2020 produit des lectures de longueur largement inférieure à la longueur d'un transcrit, et pour cette raison les ARNs messagers sont fragmentés préalablement au séquençage. Cela complique toutefois beaucoup les choses lorsque plusieurs messagers partagent des séquences communes : il devient impossible d'identifier sans ambiguïté le messager d'origine de certaines lectures. Cette situation survient lorsqu'un gène possède présente plusieurs isoformes ou dans le cas de paralogies (duplications de gènes) récentes.

1. Supposons un instant que l'on réalise une expérience de RNA-seq avec un séquenceur capable de séquencer les messagers entiers.
   (a) Existe-t-il aujourd'hui de tels séquenceurs ? Si oui, pourquoi ne sont-ils pas utilisés en routine pour du RNA-seq, si non quel est l'obstacle technique à leur réalisation ?
   (b) Proposez un modèle pour les données issues du séquenceur et une manière d'estimer l'abondance relative des messagers à partir d'elle (ne donner que les résultats, sans démonstration). On notera ρ_t l'abondance relative d'un transcrit t.

On revient maintenant au problème de départ (lectures courtes, gènes avec plusieurs isoformes), mais en se concentrant dans un premier temps sur un exemple, donné en figure 1. On supposera pour cet exemple que les trois isoformes sont de longueurs égales. Dans ce cas, la proportion d'isoformes A dans un échantillon est en moyenne égale à la proportion de lectures provenant de l'isoforme A après fragmentation.

2. En supposant que les trois isoformes sont présentes en concentrations égales, donnez intuitivement la probabilité que chaque lecture viennent de chaque isoforme.

De manière absolument heuristique, on peut tenter d'assimiler la quantification des isoformes à un scrutin, où chaque lecture dispose d'un suffrage. Contrairement aux types de scrutins en vigueur sous notre constitution, on peut imaginer que les lectures aient le droit de « partager » leur vote entre plusieurs isoformes, et que ce partage soit donné par les probabilités calculées à la question précédente.

1. Il faut bien insister sur ce point : l'information « brute » fournie par une expérience de RNA-seq n'est pas une concentration, c'est-à-dire l'abondance absolue de chaque messager, mais simplement sa proportion dans l'ensemble des messagers. Cela a notamment des implications au moment de l'analyse différentielle.

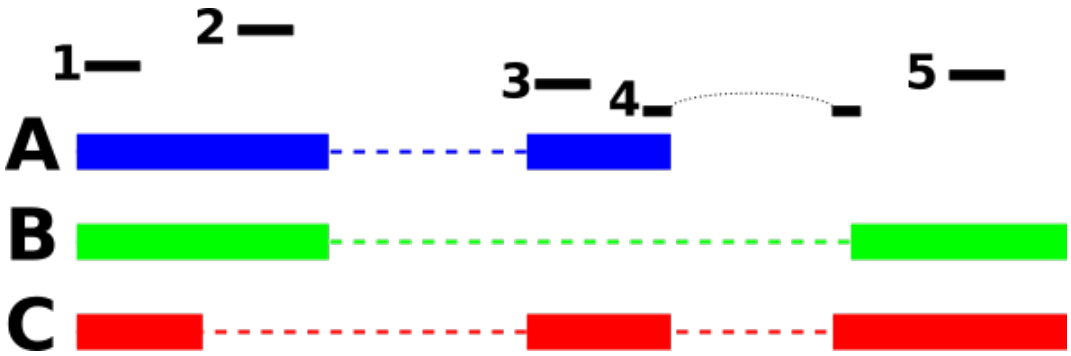

FIGURE 1 – Illustration du problème de quantification des isoformes (adapté de [1]). On considère un gène présentant trois isoformes A, B et C. Le séquençage d'un échantillon d'ARN messager produit 5 lectures.

3. À l'aide des probabilités calculées à la question précédente, proposez une estimation de l'abondance relative des différentes isoformes.
4. Avec le résultat précédent, proposez une nouvelle estimation de la probabilité pour chaque lecture de provenir de chaque isoforme.
5. Utilisez ces probabilités mises à jour pour produire une nouvelle estimation de la proportion de chaque isoforme.
6. Réalisez une itération supplémentaire des deux étapes précédentes. À quoi pourrait-on reconnaître le moment où l'on peut stopper les itérations ?
7. Proposez une description formelle de l'algorithme déroulé aux questions précédentes.

À ce stade, rien ne garantit que l'algorithme que vous venez de décrire conduise à une évaluation fiable des abondances relatives d'isoformes. Pour clore cette partie, nous allons éprouver expérimentalement son comportement, en implémentant un simulateur de données. On définit un gène G comme l'ensemble $\{1, \ldots, m\}$ de ses exons et une isoforme comme un sous-ensemble de G.

8. Montrez sur un exemple comment on peut coder l'ensemble des transcrits d'un gène par une matrice de 0 et de 1.
9. Implémentez un simulateur de données RNA-seq, prenant en entrée le nombre de lectures souhaité, l'ensemble des transcrits et leur abondance relative. Pas besoin ici de simuler les séquences nucléotidiques ou la qualité de séquençage, une représentation très simple des lectures fera très bien l'affaire : chaque lecture pourra être représentée par l'exon où elle s'aligne (pour simplifier on supposera l'absence de jonctions).
10. Implémentez l'algorithme décrit en question 7 et évaluez sur quelques simulations sa capacité à retrouver les véritables abondances relatives d'isoformes à partir des lectures séquencées. Pensez bien à vous limiter à des isoformes de tailles égales pour rester dans les hypothèses de l'algorithme.

2 Algorithme EM

L'algorithme EM (pour *Expectation-Maximization*, espérance-maximisation) est un algorithme permettant de calculer le maximum de vraisemblance d'un modèle probabiliste sur des observations **incomplètes**. Soit un modèle probabiliste définissant deux groupes de variables aléatoires, X et Z, et paramétré par des paramètres θ. On suppose que l'on observe X (on note x sa réalisation) mais pas Z, qui constitue les données manquantes du problème (c'est en cela que l'on parle d'observations incomplètes). Si l'on observait aussi Z, on pourrait estimer θ par maximum de vraisemblance, en cherchant numériquement ou analytiquement la valeur de θ rendant les données les plus vraisemblables. Comme on n'observe pas Z, on ne peut simplement pas évaluer la vraisemblance, et encore moins chercher à la maximiser. L'algorithme EM est une alternative populaire dans cette situation.
En voici une description formelle :

1. choisir arbitrairement une valeur de départ $\theta^{(0)}$ pour les paramètres
2. répéter jusqu'à convergence ou un nombre maximal d'itérations, pour i partant de 1 et incrémentant à chaque iteration
   (a) calculer la distribution de Z sachant X et $\theta^{(i)}$, c'est-à-dire $\mathbb{P}[Z \mid X = x, \theta = \theta^{(i)}]$
   (b) calculer
   $$\theta^{(i+1)} \leftarrow \text{argmax}_\theta \, \mathbb{E}_{Z|X=x,\theta=\theta^{(i)}}[\log L(\theta \mid X = x, Z)]$$

Cette formulation très générale a de quoi intimider. Notre objectif est de savoir l'appliquer sur des problèmes particuliers, tout en constatant qu'elle conduit à des algorithmes simples et somme toute assez intuitifs. On admettra que cette procédure converge vers un maximum (possiblement local) de la fonction de vraisemblance.

La formulation ci-dessus fait apparaître une notation nouvelle avec le terme $\mathbb{E}_{Z|X=x,\theta=\theta^{(i)}}[\dots]$. Il s'agit d'une façon de préciser la distribution de la variable aléatoire apparaissant dans le terme dont on souhaite prendre l'espérance. Dans le cas où Z est une variable discrète, cette notation signifie :

$$\mathbb{E}_{Z|X=x,\theta=\theta^{(i)}}[\log L(\theta \mid X = x, Z)] = \sum_z \mathbb{P}[Z = z \mid X = x, \theta = \theta^{(i)}] \log L(\theta \mid X = x, Z = z)$$

2.1 Retour sur la quantification d'isoformes

Nous allons ici appliquer la « recette » générale de l'algorithme au problème de quantification d'isoforme. Comme d'habitude, on part de la description du modèle en termes de variables et lois de distributions, puis on déroule les formules générales dans ce cas particulier.

1. Rédigez un modèle probabiliste spécificiant le simulateur que vous avez développé en question 9 du paragraphe précédent, en précisant bien les variables observées, non-observées, les paramètres (rappel : on se limite au cas où toutes les isoformes ont le même nombre d'exons).
2. En utilisant le théorème de Bayes, en déduire une formule pour $\mathbb{P}[Z = z \mid X = x, \theta = \theta^{(i)}]$
3. Donnez l'expression de $\log L(\theta \mid X = x, Z)$.
4. Donnez l'expression de $\mathbb{E}_{Z|X=x,\theta=\theta^{(i)}}[\log L(\theta \mid X = x, Z)]$

5. Écrivez l'algorithme EM complet avec les formules trouvées et comparez avec l'algorithme écrit en première partie.
6. Serait-il difficile d'autoriser les isoformes à avoir un nombre différent d'exons ?

2.2 Estimation de la diversité génétique

Lors du premier cours, vous avez vu un modèle simple pour travailler sur la diversité génétique, comprise comme la fréquence avec laquelle on trouve des variations génétiques (ici ponctuelles) entre deux individus d'une même espèce. Le modèle comprend la séquence observée d'un individu (via un séquenceur, elle est donc possiblement bruitée), son vrai génotype et une séquence de référence. Il implique également un paramètre ε de fréquence d'erreurs de séquençage, et un paramètre θ donnant la fréquence des variations génétiques.

Comme suggéré à la fin du cours, nous allons ici appliquer l'algorithme EM pour montrer comment déterminer le paramètre θ, en supposant la valeur de ε connue.

1. Écrivez le modèle complet, en précisant variables observées, variables non-observées (on dit souvent variables cachées) et paramètres.
2. En déduire une formule pour $\mathbb{P}[Z = z \mid X = x, \theta = \theta^{(i)}]$
3. Donnez l'expression de $\log L(\theta \mid X = x, Z)$.
4. Donnez l'expression de $\mathbb{E}_{Z|X=x,\theta=\theta^{(i)}}[\log L(\theta \mid X = x, Z)]$
5. Écrivez l'algorithme EM complet avec les formules trouvées ci-dessus.
6. Écrivez un simulateur de données selon le modèle de la question 1.
7. Essayez votre algorithme sur un grand nombre de simulations, et représentez sur un graphique l'ensemble des résultats par un nuage de points : chaque point a pour abscisse la vraie valeur de θ, et pour ordonnée la valeur trouvée par l'algorithme.

2.3 Une application en écologie

Des collègues écologues souhaitent évaluer l'abondance relative de deux espèces de goujons dans le Limousin. Problème, les deux espèces sont très proches en apparence, et l'identification requiert un examen minutieux et long. En revanche, on sait que les individus adultes sont en moyenne plus grands dans une espèce que dans l'autre. À l'issue d'une campagne de pêche, les collègues disposent de la taille d'un nombre n de spécimens adultes. Pouvez-vous les aider à estimer la proportion d'individus de chaque espèce ?

Références

[1] L. Pachter, "Models for transcript quantification from rna-seq," 2011.