IMPAQT Isoform Abundance Estimation

Bayesian HMM that learns as it processes? Think about it. Bayes is an interative process. If we process from 3’ to 5’, its going to adjust it’s sensitivity.

**The idea is to place a prior on the emission probailities that is constantly learning. Basically, it’s an empirical bayes variant of the HMM that penalizes clusters downstream of the original. The probaility of it’s inverse state relative to itself is the proportion of total reads. :)**

Here me out. We set the state probabilities to init values:

Cluster = (n \* num reads / span of window)

No cluster = 1 - clusters

Each emission probability is the beta binomial distribution. The uncertainty of the theta parameter (probability, I guess) is governed by what it’s discovered previously. There should be a heuristic way to implement the idea that a huge peak downstream of it increases the uncertainity of it’s own theta estimate.

But the transition probabilities are consistently learned. They are initially set to uniform, but with a beta prior.

Brainstorm:

* what if we model DBSCAN parameters THEN cluster?
* Does the beta distribution
* What if we account for PCR duplicates? Like really just probabilistically remove counts due to PCR duplicates? Are narrow clusters more likely to be full of PCR duplicates? This idea kind of sucks?
* Probability of a read being generated from a transcript is proportional to the transcripts expression AND the probability the read was selected, which is impacted by sequence bias (GC bias? Bias of the random primers?)
* Also, theoretically, isn’t reasonable to assume only reads upstream of the cluster should be considered as “potentially belonging to each transcript?” What does this mean? Like heirarchy?
  + WAIT THIS MIGHT BE IT
  + abundance of transcript cluster should depend on the abundance of the cluster downstream?
  + So this essentially prioritizes the 3’ most clusters, is that real? I mean technically that makes sense. So think of this case, two clusters with equal expression. It’s more likely that at least one of the reads originated from the more 3’ most cluster. Also think of the instance when the 5’ most cluster is more highly expressed? Given the assumption that the majority of the reads happen at the 3’ most region of the transcript, the downstream cluster is very unlikely to have transcript not belonging to itself. Conversely, a more highly expressed cluster that is the 3’ most cluster, is much more likely to contribute more reads to the second cluster existing, so it’s detection might be inflated just by chance. How do we establish removing these clusters?
  + Maybe fit a beta binomial model for each cluster (is in cluster vs is not).

Side note, what if we took a different approach? What if we did this in reverse? Like what if we modeled widths of clusters first and used

The problem of identifying distinct transcript isoforms and estimating their gene expression from TAGseq data requires careful consideration of the sequencing read generating mechanism (library preparation protocol). Random priming along the 3’ UTR of a polyadenylated transcript results in distributions of read coverage that are mostly concentrated immediately upstream of the polyA tail. We also know that the more highly expressed isoformss are more likely it is to be sequenced and thus, reads originating from regions farther upstream of the 3’ UTR are more likely to be observed. However, empirical evidence pertaining to the

In the case where multiple peaks in the read coverage density can be identified, correctly assigning these reads to there appropriate isoforms is impossible, but a probablistic approach that considers initial evidence of transcript expression as well as coverage distribution dispersion could provide an appropriate method of quantifiying expression of transcripts with distinct 3’ ends.

To demonstrate this approach, our model will be defined using a two isoform example.

Formally, our model seeks to infer the proportion of a sample of N reads belonging to a transcript isoform, ***I***, given a set of read alignments and the uncertainty surrounding an alternative isoform. Intuitively, we know that the probability of a read originating from an isoform depends on the isoform’s level of expression and the dispersion of its read coverage arising from random biological or techincal effects. The

DO NOT MODEL DISTANCE, MODEL SPAN OF CLUSTER.

the proportion of a sample of N reads that originated from a transcript isoform. This proportion can also be interpretted as the probability that

***θ ~ Beta(a,b)***

***I | θ ~ Bern(θ)***

S | ***I*** = span of region given expression, seems to be lognormal. So we are not modeling th

***R | I,S ~ Function of the dist***

Differentiating between transcript isoforms using

a probabalistic solution.

Ok so the issue is that I was saying P(R|GC(I)) impacts estimation and not actually what it represents which is it’s update of read assignment. This is dumb or at least not necessarily our problem. GC bias should probably be used to estimate the prior.

P( ***Ψ*** | ***R*** ) = P( ***R*** | ***Ψ*** ) \* P( ***Ψ*** )

------------------------

P( ***R*** )

P( ***R*** | ***Ψ*** ) = P( ***R*** | ***I*** ) \* P( ***I*** | ***Ψ*** )

P( ***Ψ*** | ***R*** ) = P( ***R*** | ***I*** ) \* P( ***I*** | ***Ψ*** ) \* P( ***Ψ*** )

-------------------------------------

P( ***R*** )

P( ***R*** | ***I*** ) = Some function of it’s distance to the read

P( ***I*** | ***Ψ*** ) = Basically the relative expression of the clusters updated for GC bias (will have to model GC bias for this)

P( ***Ψ*** ) = Uniform Distribution

Let ***I*** represent the abundance of a transcript isoform and ***R*** represent a read potentially belonging to the isoform ***I***. Intuitively, we know that the true abundance of ***I*** depends upon the probability that ***R*** originated from that isoform. So, from Bayes Rule we get…

P( ***I*** | ***R*** ) = P( ***R*** | ***I*** ) \* P( ***I*** )

----------------------

P( ***R*** )

Our understanding of the underlying process by which reads are generated from a particular isoform in TAGseq experiments tells that the likelihood of ***R*** given ***I*** can be described as ***R***’s dependence on two random variables representing biases introduced during sequencing, namely random priming along the 3’ UTR for a given isoform, which we will denote as ***D(I),*** and the under-representation of sequences with greater proportions of G and C nucleotides, which we will denote as ***GC(I)***. So, our likelihood term can be rewritten as…

P( ***R*** | ***I*** ) = P( ***R*** | ***GC(I)*** ∩ ***D(I)*** )

Given the rules for joint probabilities, we can rewrite our updated likelihood term as the joint probability of ***R***, ***GC(I)***, and ***D(I)***.

P( ***R*** | ***GC(I)*** ∩ ***D(I)*** ) = P( ***R*** ∩ ***GC(I)*** ∩ ***D(I)*** )

------------------------------

P( ***GC(I)*** ) \* P( ***D(I)*** )

However, given that ***R*** is conditional on ***GC(I)*** and ***D(I)*** but ***GC(I)*** and ***D(I)*** can be reasonably assumed to be independent, the joint probability of the three events can be rewritten as...

P( ***R*** ∩ ***GC(I)*** ∩ ***D(I)*** ) = P( ***R*** | ***GC(I***) ) \* P( ***GC(I)*** ) \* P( ***R*** | ***D(I)*** ) \* P( ***D(I)*** )

Substituting this term back into our joint probability formula, our definition of the likelihood of ***R*** given ***GC(I)*** and ***D(I)*** reduces to…

P( ***R*** | ***GC(I)*** ∩ ***D(I)*** ) = P( ***R*** | ***GC(I***) ) \* P( ***R*** | ***D(I)*** )

Finally, we can rewrite our definition of the posterior probability as…

P( ***I*** | ***R*** ) = P( ***R*** | ***GC(I***) ) \* P( ***R*** | ***D(I)*** ) \* P( ***I*** )

--------------------------------------------

P( ***R*** )

This part is potentially not right, give it some more thought.

P( ***R*** | ***GC(I***) ) = modeling the GC content of the isoform (cluster of reads) as binomial distribution, what is the likelihood of getting a GC count of the read given the distribution of GC content in the cluster.

So potentially, GC is modelled as number of bases that are modeled as G or C (successes). When getting probabilty of successes, we scale the GC% of read by length of cluster to get propotion equivalent.

GC(I) ~ Bin(N,P)

N = Length of cluster

P = Probability of getting G or C (% GC)

P( ***R*** | ***D(I)*** ) = the likelihood of the read given it’s distance to the bounds of the cluster will be defined as the distance to the bounds of the cluster divided by the sum of the distance to all clusters subtracted from one divided again by the number of clusters - 1. Formula below:

[ 1 - ( d / ∑d) ] / [ NumClusters – 1 ]

P( ***I*** ) = prior distribution of isoform abundnance that is initially estimated as the ratio of core points in that cluster to the sum of all core points in every isoform. Formula below:

Ci / ∑Cj

References:

MISO: https://www.nature.com/articles/nmeth.1528#Sec13

MISO Supps: https://static-content.springer.com/esm/art%3A10.1038%2Fnmeth.1528/MediaObjects/41592\_2010\_BFnmeth1528\_MOESM151\_ESM.pdf

QuantSeq: <https://www.nature.com/articles/nmeth.f.376>