Genes, exons, introns, and splicing

- Gene - a segment of DNA or RNA that is transmitted from one generation to the next, and that carries genetic information such as exons for a protein.
- Exon - a region of a gene which is translated into protein.
- Introns - a region of a gene which is not translated into protein.
- Splicing - a process in which the introns are removed and exons are joined to be translated into a single protein.

Alternative splicing

- Alternative splicing is a process in which exons can be spliced out in different combinations named transcripts to generate the mature RNA molecule.
- Alternative splicing is a common mode of gene regulation within cells, being used by 90–95% of human genes.
- Alternative splicing can drastically alter the function of a gene in different tissue types or environmental conditions, or even inactivate the gene completely.
- Alternative splicing is implicated in many diseases.

Genome-wide analysis of alternative splicing

- Alternative splicing can be studied by:
  - microarray containing probes targeting individual exons or junctions.
  - large set of short sequence reads which have been randomly sampled from the cell's RNA sequences through a sequencing experiment.

The Transcriptome Reconstruction Problem

- **Given:** Set of transcripts, genome sequence and a collection of reads.
- **Find:** All transcripts from where reads were taken and their frequencies.

Algorithm for estimation of transcript frequencies

- We align the reads against the reference genome obtaining for each read a set of intervals representing the position of the read in the genome.
- For each read we need a set of transcripts from which the read could have originated.
- TopHat reads match equally well multiple transcripts, so their transcript of origin cannot be established unambiguously.
- Expectation Maximization (EM) algorithm to estimate the frequencies of the transcripts:
  - Step: Compute the expected number n(i) of reads that come from isoform i under the assumption that isoform frequencies f(i) are correct (under this assumption, the probability that read r is sampled from isoform i is given by f(i)*n(i)/sum f(i)*n(i)) is the sum of frequencies of isoforms that contain r.
  - Map: For each j, set the new value of f(j) = c(j)/(c(1) + ... + c(n)), where normalized coverages c(j) are based on expected counts computed in previous step.

Read Simulations

- Simulate a single or pair-end reads from a set of transcripts:
  - take a set of known transcripts from UCSC Genome Browser ( ), randomly select fragments of them, and then generate one or two reads from the ends of the fragments.
  - select the transcript based on different probability distributions.
  - transcripts are grouped into clusters, each cluster containing a number of unique transcripts.
  - clusters are chosen based on some known abundances observed in practice (UCSC).

Parameters for Simulations:

- **number of reads (millions):** 1, 5, 10, 15, 20, 25, 30, 50, 100
- **read length:** single reads: 25, 10, 25, 50, 100
- **Fragment length:** mean ~ 250, std. dev. ~ 25
- **Transcript probability distribution within a cluster:** uniform, geometric with ratio = 0.5
- **From UCSC Genome Browser:** list of known genes (list of exons starts and ends positions for each known transcript).
- **Cluster id:** list of exons for each known transcript.
- **Cluster abundance:** uniform at random from strings?
- **Clusters probability distribution:** based on clusters abundances and transcript length and distribution.

For each set of parameters running the simulations will generate a dataset consisting of:

- reads sequences and quality scores
- reads mapping positions onto the reference genome and reads sequences
- simulated transcripts frequencies

Exact + Unique :

- run Exact+EM/Genome+EM/Transcript+EM flow only on the subset of reads that are not ambiguous, (i.e., match perfectly only one transcript) and discard ambiguous reads.

Exact + Unique + EM:

- run Exact+EM/Genome+EM/Transcript+EM flow on all reads but for only one iteration starting from the estimates obtained from Exact+Unique flow.

Quality of the model

- **Given:** A spectrum of reads (including reads and their frequencies) from a sequencing machine and a set of transcripts from a database.
- **Find:** How well do they correspond to each other if reads are chosen uniformly at random from strings?
  - Question 1: Are some transcripts missing from the database or existing strings adequately explain the spectrum of reads?
  - Question 2: What would be the reads emitted by transcripts missing from the database?

The relationship between transcripts and reads can be represented by a weighted bipartite graph in which there is an edge between a transcript and a read if that transcript emits the read. This edge is weighted by the number of times that read can be emitted by the same transcript.

Measuring the quality by Mean Square Value

The spectrum of transcripts is obtained by EM. Having the frequency of transcripts and the weights of the edges results in expected spectrum of reads. The deviation (mean square value) between expected and observed reads measures the quality of the model.

- Using bootstrapping to get the 95% confidence interval of the deviations. If the real deviation is out of this confidence interval, some transcripts are missing from the database.
- For each read if the difference between observed and expected frequencies is not equal to zero, we remove that read (which is emitted by unknown transcripts) and repeat this process until the difference becomes zero for any single read.

Transcriptome Reconstruction from Sequence Reads

Marius Niculce, Serghei Mangul, Hamed Salooti, Ion Mandoiu and Alexander Zelikovsky

Department of Computer Science, Georgia State University, University of Connecticut

Supported by Molecular Basis of Disease Program, GSU, NSF Grant IIS-0916401