CSE 549: RNA-Seq aided gene finding
# Finding Genes

We’ll break gene finding methods into 3 main categories.

<table>
<thead>
<tr>
<th>ab initio</th>
<th>comparative</th>
<th>combined / extrinsic</th>
</tr>
</thead>
<tbody>
<tr>
<td>latin — “from the beginning” w/o experimental evidence</td>
<td>make use of knowledge across species</td>
<td>Make use of experimental evidence (e.g. RNA-seq)</td>
</tr>
<tr>
<td>based on predictive modeling</td>
<td>a known human gene is strong evidence for a chimp gene</td>
<td>Evidence highlights transcribed regions</td>
</tr>
<tr>
<td>how well do genomic sequences score under our “gene model”?</td>
<td>many “housekeeping” genes are incredibly similar across highly divergent species</td>
<td>Gene structure extracted from evidence (potentially combined with model predictions)</td>
</tr>
</tbody>
</table>
Typical Approaches to Annotation are “Hybrid” Methods

Consider *ab initio* predictions

Refine with homological & experimental evidence

Combine predictions of many different tools

Manually curate the most promising results
Hybrid Gene Finding “Pipelines”

(A) ab initio gene finding using a selection of the following software tools: GeneMarkHMM, FGENESH, Augustus, and SNAP, GlimmerHMM.

(B) protein homology detection and intron resolution using the GeneWise software and the uniref90 non-redundant protein database.

(C) alignment of known ESTs, full-length cDNAs, and most recently, Trinity RNA-Seq assemblies to the genome.

(D) PASA alignment assemblies based on overlapping transcript alignments from step (C)

(E) use of EvidenceModeler (EVM) to compute weighted consensus gene structure annotations based on the above (A, B, C, D)

(F) use of PASA to update the EVM consensus predictions, adding UTR annotations and models for alternatively spliced isoforms (leveraging D and E).

(G) limited manual refinement of genome annotations (F) using Argo or Apollo

http://pasa.sourceforge.net
What is “experimental” data?

There are many ways we can obtain “experimental” evidence of a gene.

- Sequencing of protein product (mass spec.)
  - Expensive & slow, but provides direct evidence of protein coding genes (reverse translation)
- Expressed Sequence Tags (EST)
  - Targeted sequencing (typically Sanger sequencing) of expressed transcripts
- RNA-Seq
  - High throughput sequencing of the “transcriptome”
1. fragmentation of RNA
2. random priming to make sscDNA (first-strand synthesis)
3. construction of dscDNA (second-strand synthesis)
4. size selection
5. sequencing
6. mapping

How is RNA-seq Useful?

RNA-Seq reads come from a spliced transcript — if we can map them back to the genome, they give us evidence of transcribed regions.

Human genome contains > 14,000 pseudogenes [Pei et al. Genome Biology 2012]

RNA-seq Alignment is Hard

RNA-seq reads can span exon boundaries. Thus, contiguous regions of the read may map kilobases apart when aligning to the genome.

Improving the quality, sensitivity and speed of “spliced” alignment is still an active area of research. How can we be confident in a spliced-alignment when only a small portion of a read maps to an exon?

Assemblers have to deal with non-uniform coverage and a mixture of different but highly-related isoforms in the same sample. We’ll talk more about RNA-seq assembly later.
Does Experimental Evidence Help?

There are (as we’ll see over the next 2 lectures) many uses of RNA-seq apart from helping \textit{ab initio} gene prediction.

Nonetheless, such evidence may be a powerful tool in helping us predict the existence of new genes.

Next, we’ll compare the unsupervised (no experimental evidence) and semi-supervised (with the help of RNA-seq) versions of a popular eukaryotic gene predictor — GeneMark
Unsupervised Gene Finding w/o Experimental Evidence

Start with heuristic / uninformative parameters and train via the Viterbi training algorithm we discussed previously.

D. mel exon lengths

C. intestinalis intron lengths

Lomsadze et al. 2005
### Table 1. Values of several categories of sensitivity and specificity (Sn/Sp) and (Sn+Sp)/2 characterizing the accuracy of gene predictions produced for the group of ‘well-studied’ genomes by the eukaryotic GeneMark.hmm with models derived by both unsupervised and supervised training

<table>
<thead>
<tr>
<th></th>
<th>A. thaliana</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unsupervised</td>
<td>Supervised</td>
<td>Unsupervised</td>
<td>Supervised</td>
<td>Unsupervised</td>
<td>Supervised</td>
<td></td>
</tr>
<tr>
<td>Nucleotide</td>
<td>97.7</td>
<td>96.3</td>
<td>97.2</td>
<td>95.8</td>
<td>99.1</td>
<td>96.4</td>
<td>97.8</td>
</tr>
<tr>
<td></td>
<td>94.8</td>
<td>94.3</td>
<td>94.0</td>
<td>93.6</td>
<td>94.0</td>
<td>92.7</td>
<td>90.8</td>
</tr>
<tr>
<td>Internal exons</td>
<td>91.2</td>
<td>89.5</td>
<td>87.3</td>
<td>89.9</td>
<td>91.3</td>
<td>94.0</td>
<td>91.3</td>
</tr>
<tr>
<td>Initiation sites</td>
<td>80.1</td>
<td>78.3</td>
<td>80.1</td>
<td>76.0</td>
<td>85.8</td>
<td>68.9</td>
<td>79.2</td>
</tr>
<tr>
<td></td>
<td>76.5</td>
<td>71.9</td>
<td>85.8</td>
<td>77.4</td>
<td>68.9</td>
<td>73.3</td>
<td>83.9</td>
</tr>
<tr>
<td>Termination sites</td>
<td>87.5</td>
<td>85.3</td>
<td>88.3</td>
<td>83.7</td>
<td>85.1</td>
<td>85.2</td>
<td>89.6</td>
</tr>
<tr>
<td></td>
<td>83.1</td>
<td>81.6</td>
<td>78.6</td>
<td>75.3</td>
<td>85.1</td>
<td>85.2</td>
<td>89.6</td>
</tr>
<tr>
<td>Donor sites</td>
<td>94.0</td>
<td>92.2</td>
<td>94.0</td>
<td>91.9</td>
<td>96.2</td>
<td>93.5</td>
<td>93.7</td>
</tr>
<tr>
<td></td>
<td>90.3</td>
<td>89.8</td>
<td>90.0</td>
<td>90.8</td>
<td>91.4</td>
<td>90.8</td>
<td>90.0</td>
</tr>
<tr>
<td>Acceptor sites</td>
<td>94.0</td>
<td>92.1</td>
<td>93.0</td>
<td>91.4</td>
<td>97.3</td>
<td>94.5</td>
<td>93.0</td>
</tr>
<tr>
<td></td>
<td>90.2</td>
<td>89.2</td>
<td>91.6</td>
<td>91.4</td>
<td>95.2</td>
<td>92.8</td>
<td>90.0</td>
</tr>
</tbody>
</table>

Boldface highlights the higher value in comparison of unsupervised and supervised modes (ES-3.0 versus E-3.0).
Learning Site-Specific Features

Donor

Acceptor

1st iter

convergence

Lomsadze et al. 2005
Incorporating Experimental Evidence

GeneMark ET Procedure

Genome sequence → RNA-seq reads → Heuristic parameters → Spliced alignment

Parameters of donor and acceptor states, intron state → Initial HMM parameters → Mapped splice sites

GeneMark.hmm gene prediction → Parameter re-estimation

Incorporating Experimental Evidence

**Figure 3.** Selection of elements of training set in GeneMark-ET for the next iteration. The new training set of protein-coding regions is comprised from exons with at least one ‘anchored splice site’ as well as long exons predicted *ab initio* (>800 nt).
Table 4. Assessment of gene prediction accuracy of GeneMark-ES (ES) and GeneMark-ET (ET) gene finders using unsupervised (genomic based) and semi-supervised (genomic and transcriptomic based) training, respectively.

<table>
<thead>
<tr>
<th></th>
<th>D. melanogaster</th>
<th>A. aegypti</th>
<th>A. gambiae</th>
<th>A. stephensi</th>
<th>Culex q.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES</td>
<td>ET</td>
<td>ES</td>
<td>ET</td>
<td>ES</td>
</tr>
<tr>
<td>Internal exon</td>
<td>Sn</td>
<td>86.7</td>
<td>87.2</td>
<td>69.3</td>
<td>91.7</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>76.9</td>
<td>82.9</td>
<td>60.7</td>
<td>75.9</td>
</tr>
<tr>
<td>Intron</td>
<td>Sn</td>
<td>82.6</td>
<td>84.8</td>
<td>67.9</td>
<td>89.6</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>75.3</td>
<td>79.2</td>
<td>64.6</td>
<td>80.3</td>
</tr>
<tr>
<td>Donor site</td>
<td>Sn</td>
<td>85.3</td>
<td>87.0</td>
<td>74.6</td>
<td>92.8</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>84.5</td>
<td>86.5</td>
<td>76.2</td>
<td>86.8</td>
</tr>
<tr>
<td>Acceptor site</td>
<td>Sn</td>
<td>86.2</td>
<td>88.2</td>
<td>74.3</td>
<td>94.1</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>85.5</td>
<td>87.0</td>
<td>79.0</td>
<td>89.6</td>
</tr>
<tr>
<td>Initiation site</td>
<td>Sn</td>
<td>71.0</td>
<td>75.1</td>
<td>62.5</td>
<td>79.6</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>83.1</td>
<td>81.5</td>
<td>77.1</td>
<td>83.9</td>
</tr>
<tr>
<td>Termination site</td>
<td>Sn</td>
<td>77.3</td>
<td>84.2</td>
<td>68.1</td>
<td>88.0</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>90.7</td>
<td>90.0</td>
<td>91.3</td>
<td>96.0</td>
</tr>
<tr>
<td>Nucleotide</td>
<td>Sn</td>
<td>91.5</td>
<td>92.1</td>
<td>87.0</td>
<td>98.1</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>98.3</td>
<td>97.4</td>
<td>95.2</td>
<td>96.2</td>
</tr>
<tr>
<td>Gene</td>
<td>Sn</td>
<td>57.9</td>
<td>63.6</td>
<td>40.3</td>
<td>66.7</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>57.3</td>
<td>61.0</td>
<td>42.6</td>
<td>64.3</td>
</tr>
<tr>
<td>Partial gene</td>
<td>Sn</td>
<td>59.9</td>
<td>67.2</td>
<td>41.2</td>
<td>69.0</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>59.3</td>
<td>64.5</td>
<td>43.6</td>
<td>66.5</td>
</tr>
</tbody>
</table>

Bold font highlights the higher accuracy value in a given category and given species. Partial gene level accuracy is computed without taking into account a difference in annotation and prediction of translation starts. Spliced alignments for GeneMark-ET were produced by UnSplicer.
## Effect of Using Spliced-Alignments

**Table 4.** Assessment of gene prediction accuracy of GeneMark-ES (ES) and GeneMark-ET (ET) gene finders using unsupervised (genomic based) and semi-supervised (genomic and transcriptomic based) training, respectively

<table>
<thead>
<tr>
<th></th>
<th>D. melanogaster</th>
<th>A. aegypti</th>
<th>A. gambiae</th>
<th>A. stephensi</th>
<th>Culex q.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ES</td>
<td>ET</td>
<td>ES</td>
<td>ET</td>
<td>ES</td>
</tr>
<tr>
<td>Internal exon</td>
<td>Sn</td>
<td>86.7</td>
<td>87.2</td>
<td>69.3</td>
<td>91.7</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>76.9</td>
<td>82.9</td>
<td>60.7</td>
<td>75.9</td>
</tr>
<tr>
<td>Intron</td>
<td>Sn</td>
<td>82.6</td>
<td>84.8</td>
<td>67.9</td>
<td>89.6</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>75.3</td>
<td>79.2</td>
<td>64.6</td>
<td>80.3</td>
</tr>
<tr>
<td>Donor site</td>
<td>Sn</td>
<td>85.3</td>
<td>87.0</td>
<td>74.6</td>
<td>92.8</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>84.5</td>
<td>86.5</td>
<td>76.2</td>
<td>86.8</td>
</tr>
<tr>
<td>Acceptor site</td>
<td>Sn</td>
<td>86.2</td>
<td>88.2</td>
<td>74.3</td>
<td>94.1</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>85.5</td>
<td>87.0</td>
<td>79.0</td>
<td>89.6</td>
</tr>
<tr>
<td>Initiation site</td>
<td>Sn</td>
<td>71.0</td>
<td>75.1</td>
<td>62.5</td>
<td>79.6</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>83.1</td>
<td>81.5</td>
<td>77.1</td>
<td>83.9</td>
</tr>
<tr>
<td>Termination site</td>
<td>Sn</td>
<td>77.3</td>
<td>84.2</td>
<td>68.1</td>
<td>88.0</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>90.7</td>
<td>90.0</td>
<td>91.3</td>
<td>96.0</td>
</tr>
<tr>
<td>Nucleotide</td>
<td>Sn</td>
<td>91.5</td>
<td>92.1</td>
<td>87.0</td>
<td>98.1</td>
</tr>
<tr>
<td></td>
<td>Sp</td>
<td>98.3</td>
<td>97.4</td>
<td>95.2</td>
<td>96.2</td>
</tr>
</tbody>
</table>

Bold font highlights the higher accuracy value in a given category and given species. Partial gene level accuracy is computed without taking into account a difference in annotation and prediction of translation starts. Spliced alignments for GeneMark-ET were produced by UnSplicer.
Effect of Using Spliced-Alignments

Figure 4. Observed dynamics of change in iterations of the mean of Sn and Sp internal exon prediction values for the GeneMark-ET and GeneMark-ES algorithms in cases of *Drosophila melanogaster* (A) and *Anopheles aegypti* (B) genomes.
Figure 1. The dot plot graph depicting average lengths of exons, introns and intergenic regions against the value of percentage of non-coding DNA in a given genome was made for the five insect genomes used in the GeneMark-ET tests as well as for several other eukaryotic species. The average lengths of intron and intergenic regions correlate with the genome length while the average length of protein-coding exons (CDS) does not show dependence on the genome size.

Next 6 Slides
Combining Evidence from Multiple Predictions

Figure 3
Venn diagrams contrasting correctly predicted rice gene structure components by ab initio gene finders. Percentages are shown for the fraction of 1,058 cDNA verified rice genes and gene structure components that were predicted correctly by each ab initio gene predictor. The cDNA-verified gene structure components consist of 7,438 total exons: 86 single, 5408 internal, 972 initial, and 972 terminal.
Combining Evidence from Multiple Predictions

Figure 1
Rice Ab initio gene prediction accuracies. Gene prediction accuracies are shown for GeneMark.hmm, Fgenesh, and GlimmerHMM ab initio gene predictions based on an evaluation of 1058 cDNA-verified reference rice gene structures. The accuracy of EvidenceModeler (EVM) consensus predictions from combining all three ab initio predictions using equal weightings (weight = 1 for each) is also provided.
Combining Evidence from Multiple Predictions

Figure 4
Exon prediction accuracy limited to consensus complete exon calls. Exon sensitivity (eSn) and exon specificity (eSp) were determined by comparing ab initio predicted exons. Exons were restricted to those perfectly agreed upon by either two or three different gene predictors. Only those predicted exons found within 500 base pairs flanking the 1,058 reference gene structures were considered for the specificity calculations.
Combining Evidence from Multiple Predictions
Combining Evidence from Multiple Predictions

Figure 6
Response of EVM prediction accuracy to varied evidence types and weights. Iterations (30) of randomly weighted evidence types were evaluated by EvidenceModeler (EVM). Iterations 1 to 10 included only ab initio predictors GlimmerHMM, Fgenesh, and GeneMark.hmm. Iterations 11 to 20 additionally included AAT-nap alignments of nonrice proteins, GeneWise predictions based on nonrice protein homologies, and AAT-gap2 alignments of other plant expressed sequence tags. Iterations 21 to 30 included Program to Assemble Spliced Alignments (PASA) alignment assemblies and corresponding suplement of PASA long-open reading frame (ORF)-based terminal exons. Exon and complete gene prediction sensitivity values resulting from EVM using the corresponding weight combinations are plotted below.
Figure 9
Human consensus gene prediction accuracy by EVM. The consensus gene prediction accuracy by EvidencerModeler (EVM) is shown based on trained evidence weights and the corresponding combination of evidence, as applied to the GENCODE test regions of the human genome. The accuracies for the inputted gene predictions obtained from the ENCODE Genome Annotation Assessment Project (EGASP) dataset are provided for reference sake, including GENSCAN, TWINSCAN, GlimmerHMM, GeneMark.hmm on the repeat-masked genome, GeneID, and SGPgene. EVM-GF corresponds to EVM applied to these gene prediction tiers alone (GF), and serves as the baseline evidence for the subsequent entries. EVM-GeneWise includes GeneWise predictions based on nonhuman protein homologies; EVM-nap includes AAT-nap spliced alignments of nonhuman proteins; the EVMgap2_"_series includes AAT-gap2 alignments of corresponding transcripts from the Dana Farber Gene Indices (CINGI, Ciona intestinalis [sea squirt]; XGI, Xenopus tropicalis [frog]; GGGL, Gallus gallus [chicken]; DOGGL, Canis familiaris [dog]; SGGI, Sus scrofa [pig]; RGI, rat; MGI, mouse); EVM-alignAsm includes Program to Assemble Spliced Alignments (PASA) alignment assemblies and corresponding terminal exon supplement; and EVM:All includes all evidence described (GF, gap2, nap, GeneWise, and PASA). Sn, sensitivity; Sp, specificity.
Other Ways of Using Experimental Evidence

Experimental evidence (RNA-seq, in particular) is a great help in improving gene prediction. However, its uses stretch far beyond assisting *ab initio* gene prediction.

**Transcript quantification**
- Differential expression, alternative splicing analysis

**Fusion/chimera detection**

**Variant (SNP, SV, CNV) detection**

**Transcript assembly**
- Genome guided & de novo

**Build higher-level models of transcription**
- co-expression networks -> regulatory networks
Other Ways of Using Experimental Evidence

Transcript quantification

How much of each gene (or transcript / isoform) is present in a particular experiment?

Differential expression, alternative splicing analysis

What are the \textit{statistically significant} differences in expression or splicing between experimental conditions?

These tools can be used to study e.g. differences between healthy / diseased tissue, or how gene expression differs across tissue types.
Other Ways of Using Experimental Evidence

Fusion/chimera detection

![Diagram of gene fusion](http://biome.ewha.ac.kr:8080/FusionGene/)

Variant (SNP, SV, CNV) detection

Find small (SNP) or large (SV) variation in how reads map back to their genes of origin

Find differences in the number of copies of a gene in the DNA (CNV)

image from: [http://biome.ewha.ac.kr:8080/FusionGene/](http://biome.ewha.ac.kr:8080/FusionGene/)
Other Ways of Using Experimental Evidence

Transcript assembly

With sufficiently deep sequencing, we can hope to assemble transcripts present in an experiment in a manner similar to how we assemble DNA.

This often lets us find previously undiscovered genes, as well as novel splice variants (combination of exons that make up an isoform of the gene).

Genome guided and de novo

Assembly can either rely on knowing the reference genome (making the problem much easier), or can be done directly from the RNA-seq reads without the reference (or via hybrid approaches).
Other Ways of Using Experimental Evidence

Build higher-level models of transcription

Co-expression networks -> regulatory networks

By looking at how the expression of different genes covaries across many different experimental conditions and tissue types, we can begin to view the set of genes as a network.

Which genes tend to be “turned on” when others are “turned off”.

Use such information to try and determine regulatory relationships between genes — which genes control others, and how.
Other Ways of Using Experimental Evidence

Build higher-level models of transcription

Co-expression networks -> regulatory networks

By looking at how the expression of different genes covaries across many different experimental conditions and tissue types, we can begin to view the set of genes as a network.

Which genes tend to be “turned on” when others are “turned off”.

Use such information to try and determine regulatory relationships between genes — which genes control others, and how.
Genome-guided Assembly, an Example

We’ll take a look at how Cufflinks, one of the most popular tools for RNA-seq-based transcript discovery & quantification assembles transcripts from data.

Transcript (rather than just “gene”) discovery lets us explore the different variants of a gene that are actually expressed in a cell.

Caveat: Cufflinks solves the identification and quantification in largely separate phases. This turns out to discard a lot of useful information. Newer approaches attempt to combine these two phases, which results in both better identification and better quantification. Still, Cufflinks is an incredibly useful (and widely used) tool (cited ~2200 times).

![Diagram of transcript structures illustrating 11 distinct types of alternatively included regions (AIRs) within the genes.](image)
Transcript (rather than just “gene”) discovery lets us explore the different variants of a gene that are actually expressed in a cell.
Cufflinks Transcript Assembly


* Slide from Carl Kingsford
Def. A pair \((S, \leq)\) is a partial order if, for all \(x, y \in S\):

- (transitivity) \(x \leq y \) and \(y \leq z \Rightarrow x \leq z\)
- (reflexivity) \(x \leq x\)
- (antisymmetry) \(x \leq y \) and \(y \leq x \Rightarrow x = y\)
**Def.** A pair \((S, \leq)\) is a partial order if, for all \(x, y \in S\):

- (transitivity) \(x \leq y \text{ and } y \leq z \implies x \leq z\)
- (reflexivity) \(x \leq x\)
- (antisymmetry) \(x \leq y \text{ and } y \leq x \implies x = y\)

*Slide from Carl Kingsford*
Def. A pair \((S, \leq)\) is a partial order if, for all \(x, y \in S\):

- (transitivity) \(x \leq y \) and \( y \leq z \) \(\Rightarrow x \leq z\)
- (reflexivity) \(x \leq x\)
- (antisymmetry) \(x \leq y \) and \( y \leq x \) \(\Rightarrow x = y\)

*Slide from Carl Kingsford*
Cufflink's Partial Order

- sequenced fragment:

- $x \preceq y$ if $x$ aligns to the left of $y$ and $x$ and $y$ have compatible intron structure

- $y_1 \preceq x_1$: incompatible b/c the right end of $y_2$ is split-mapped, implying an intron where there is no intron in $x_2$.

- $x_3$ and $y_3$ are nested, and so are merged into a single fragment.

- $x_4$ is uncertain because it could be compatible with either $y_4$ or $y_5$; $x_4$ is therefore thrown away.

(Trapnell et al., 2010)

* Slide from Carl Kingsford
Cufflinks’ Assembly Algorithm

Partitioning partial order into smallest # of chains → “parsimonious” set of transcripts that explains the observed reads

Dilworth’s Theorem

Dilworth ≡ König

König’s Theorem

Solvable in $O(E\sqrt{V})$

A vertex cover is a subset of the vertices such that each edge is adjacent to at least one vertex from the subset.
Cufflinks’ Assembly Algorithm

Partitioning partial order into smallest # of chains → “parsimonious” set of transcripts that explains the observed reads

Smallest # of chains

Largest antichain

Dilworth’s Theorem

Dilworth ≡ König

Vertex cover

König’s Theorem

Maximum bipartite matching

A vertex cover is a subset of the vertices such that each edge is adjacent to at least one vertex from the subset.

Solvable in $O(E\sqrt{V})$

* Slide from Carl Kingsford
Dilworth’s Theorem

**Thm (Dilworth).** In a poset, the size of the largest antichain = the size of the minimum cover by chains.

*Proof intuition.*

- The largest antichain must hit every chain (otherwise it could be made larger).

- It can’t hit any chain twice, otherwise it would contain two comparable items.
König’s Theorem

**Thm (König).** In a bipartite graph, the # of edges in a maximum matching = # of vertices in the smallest vertex cover.

**Proof intuition.**

- In a maximum matching, every edge must be covered.
- Otherwise, if both endpoints are not matched, we could add that edge to the matching and increase its size.

* Slide from Carl Kingsford*
Let $M$ be the maximal matching.

By König’s theorem, there is a (minimal) vertex cover $C$ of the same size as $M$.

Let $T$ be the elements of the poset that are not in $C$.

$T$ is an antichain. Why?

Make a set $W$ of chains by $u \equiv v$ if $(u, v) \in M$.

These equivalence classes are chains. Why?
Using Matching to Find a Minimal Chain Cover

Let $M$ be the maximal matching.

By König’s theorem, there is a (minimal) vertex cover $C$ of the same size as $M$.

Let $T$ be the elements of the poset that are not in $C$.

$T$ is an antichain. Why?

Make a set $W$ of chains by $u \equiv v$ if $(u,v) \in M$.

These equivalence classes are chains. Why?

* Slide from Carl Kingsford
Using Matching to Find a Minimal Chain Cover

Let $M$ be the maximal matching.

By König’s theorem, there is a (minimal) vertex cover $C$ of the same size as $M$.

Let $T$ be the elements of the poset that are not in $C$.

$T$ is an antichain. Why?

Make a set $W$ of chains by $u \equiv v$ if $(u,v) \in M$.

These equivalence classes are chains. Why?

If $u$ and $v$ were comparable, there would be an edge between them, and since neither $u$ or $v$ used in $M$, we could add that edge to $M$.

Every pair of items in each equivalence class had an edge between them, meaning they were comparable.

* Slide from Carl Kingsford
\[ |W| = |T| \]

**M** = maximal matching.

**C** = vertex cover of the same size as **M**.

**T** = antichain elements of poset that are not in **C**.

**W** = set of chains formed from edges of **M**.

Size of **T** is \( n - m \). Why?

Size of **W** is \( n - m \). Why?

\[ n = \# \text{ elements in poset} \]

\[ m = \# \text{ of edges in matching} \]

* Slide from Carl Kingsford*
\[ |W| = |T| \]

\( M = \) maximal matching.
\( C = \) vertex cover of the same size as \( M \).
\( T = \) antichain elements of poset that are not in \( C \).
\( W = \) set of chains formed from edges of \( M \).

Size of \( T \) is \( n - m \). Why?

\( n = \# \) elements in poset
\( m = \# \) of edges in matching

Every edge uses up exactly one element on the LHS of the bipartite graph.

Size of \( W \) is \( n - m \). Why?

* Slide from Carl Kingsford
$|W| = |T|$  

$M = \text{maximal matching.}$  

$C = \text{vertex cover of the same size as } M.$  

$T = \text{antichain elements of poset that are not in } C.$  

$W = \text{set of chains formed from edges of } M.$  

$n = \# \text{ elements in poset}$  

$m = \# \text{ of edges in matching}$  

Size of $T$ is $n - m$. Why?  

Every edge uses up exactly one element on the LHS of the bipartite graph.  

Size of $W$ is $n - m$. Why?  

Consider set of $n$ "chains" each consisting of a single element of poset.  

Each edge $(u,v)$ that we use to put $v$ into the same poset as $u$ reduces the number of chains by 1.  

$\Rightarrow$ Number of equivalence-class chains $= n - m$  

* Slide from Carl Kingsford
All antichains must be of size $\leq$ all chain covers.

Suppose not, and let $A$ be an antichain bigger than cover $Q$.

Then, by pigeonhole, $A$ must contain at least 2 elements $x, y$ from the same chain in $Q$.

But $x, y$ are comparable because they are in the same chain.

$\Rightarrow$ the pair $(T,W)$ must be a largest antichain and a smallest $W$ because they are the same size.

* Slide from Carl Kingsford
A Matching-Covering Example

Double circles denote vertex cover

(Trapnell et al., 2010)

* Slide from Carl Kingsford
Selecting From Among Many Minimum Solutions

Idea: exons included in same transcript should have similar expression

Estimate Percent Spliced In (PSI, \( \psi \)): # of reads crossing exon x that are compatible with x divided by # of reads overlapping x (divided by length of x).

\[
\text{weight}(x, y) = - \log(1 - |\psi_x - \psi_y|)
\]

* Slide from Carl Kingsford*
Selecting From Among Many Minimum Solutions

Idea: exons included in same transcript should have similar expression

Correct

Wrong

Estimate Percent Spliced In (PSI, \(\psi\)): # of reads crossing exon \(x\) that are compatible with \(x\) divided by # of reads overlapping \(x\) (divided by length of \(x\)).

\[
\text{weight}(x, y) = -\log(1 - |\psi_x - \psi_y|)
\]

* Slide from Carl Kingsford
**Discovery of Novel Isoforms**

<table>
<thead>
<tr>
<th>Category</th>
<th>Transfrags</th>
<th>% of total transfrags</th>
<th>Assembled reads (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match to known isoform</td>
<td>39,857</td>
<td>13.5</td>
<td>76.7</td>
</tr>
<tr>
<td>Novel isoform of known gene</td>
<td>18,565</td>
<td>6.3</td>
<td>11.3</td>
</tr>
<tr>
<td>Contained in known isoform</td>
<td>71,029</td>
<td>24.1</td>
<td>4.6</td>
</tr>
<tr>
<td>Repeat</td>
<td>41,906</td>
<td>14.2</td>
<td>0.6</td>
</tr>
<tr>
<td>Intronic</td>
<td>32,658</td>
<td>11.1</td>
<td>0.6</td>
</tr>
<tr>
<td>Polymerase run-on</td>
<td>18,522</td>
<td>6.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Intergenic</td>
<td>48,604</td>
<td>16.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Other artifacts</td>
<td>22,483</td>
<td>7.7</td>
<td>4.5</td>
</tr>
<tr>
<td>Total transfrags</td>
<td>293,624</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

**TABLE 2.** Classification of all transfrags produced at any time point with respect to annotated gene models and masked repeats in the mouse genome. Transfrags that are present in multiple time point assemblies are multiply counted to preserve the relative distribution of transfrags among the categories across the full experiment.

(Trapnell et al., 2010)