DNA2: Aligning ancient diversity
(Last week)

- Comparing types of alignments & algorithms
- Dynamic programming (DP)
- Multi-sequence alignment
- Space-time-accuracy tradeoffs
- Finding genes - motif profiles
- Hidden Markov Model (HMM) for CpG Islands

RNA1: Structure & Quantitation

- Integration with previous topics (HMM & DP for RNA structure)
- Goals of molecular quantitation (maximal fold-changes, clustering & classification of genes & conditions/cell types, causality)
- Genomics-grade measures of RNA and protein and how we choose and integrate (SAGE, oligo-arrays, gene-arrays)
- Sources of random and systematic errors (reproducibility of RNA source(s), biases in labeling, non-polyA RNAs, effects of array geometry, cross-talk).
- Interpretation issues (splicing, 5' & 3' ends, gene families, small RNAs, antisense, apparent absence of RNA).
- Time series data: causality, mRNA decay, time-warping

Discrete & continuous bell-curves

- Normal (m=20, s=4.47)
- Poisson (m=20, s=2m)
- Binomial (N=20, p=.01, m=Np)
- t-dist (m=20, s=4.47, dof=2)
- ExtrVal(u=20, L=1/4.47)

Modified bases & bps in RNA

<table>
<thead>
<tr>
<th>Base Name</th>
<th>Full Name of Nucleoside</th>
<th>Standard Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>2'-5' linked uracil</td>
<td>U</td>
</tr>
<tr>
<td>F</td>
<td>2'-fluorouracil</td>
<td>F</td>
</tr>
<tr>
<td>L</td>
<td>2'-deoxyinosine</td>
<td>I</td>
</tr>
<tr>
<td>i</td>
<td>2'-deoxyinosine</td>
<td>I</td>
</tr>
<tr>
<td>m</td>
<td>2'-deoxynosine</td>
<td>N</td>
</tr>
<tr>
<td>M</td>
<td>2'-deoxyinosine</td>
<td>N</td>
</tr>
<tr>
<td>Z</td>
<td>2'-deoxyinosine</td>
<td>N</td>
</tr>
</tbody>
</table>

Non-watson-crick bps

- CH₃
Covariance

\[ M_{ij} = \sum_{x_i,x_j} f_{x_i,x_j} \log \frac{f_{x_i,x_j}}{f_{x_i}f_{x_j}} \]  
\text{M}_{0 \text{ to } 2 \text{ bits; } x=\text{base type}}

xixj see Durbin et al p. 266-8.

Mutual Information

\[ M_{i,j} = \sum_{x_i,x_j} f_{x_i} \log \frac{f_{x_i}}{f_{x_i}f_{x_i}} \]  
\text{M}_{0 \text{ to } 2 \text{ bits; } x=\text{base type}}

xixj see Durbin et al p. 266-8.

\begin{align*}
\text{ACUUAU} & = (\text{AU} \log_2 (\text{AU} / (\text{A} \times \text{U}))) \\
\text{AGCUUAG} & = 4 \times 0.5 \log_2 (0.5 / (0.5 \times 0.5)) = 2 \\
\text{GUCUGC} & = 4 \times 0.5 \log_2 (0.5 / (0.5 \times 0.5 \times 1)) = 0
\end{align*}

See Shannon entropy, multinomial

\[ \pi(n;q) = \frac{n!}{n_1! n_2! \cdots n_m!} \prod_{i=1}^{m} \psi_i^{n_i} \]

RNA secondary structure prediction


Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.

Each set of 750 generated structures contains one structure that, on average, has 86% of known base-pairs.

Stacked bp & ss

Figure 10.10 An example \( \Delta G \) calculation for an RNA stem loop (the wild type B17 coat protein binding site)

Initial 1981 O(N^2) DP methods: Circular Representation of RNA Structure

Did not handle pseudoknots

RNA pseudoknots, important biologically, but challenging for structure searches
Dynamic programming finally handles RNA pseudoknots too.

Rivas E, Eddy SR J Mol Biol 1999 Feb 5;285(5):2053-68 A dynamic programming algorithm for RNA structure prediction including pseudoknots. (ref)

Worst case complexity of $O(N^6)$ in time and $O(N^4)$ in memory space.

Bioinformatics 2000 Apr;16(4):334-40 (ref)

CpG Island + in a ocean of -
First order Hidden Markov Model

MM=16, HMM= 64 transition probabilities (adjacent bp)

Small nucleolar (sno)RNA structure & function

Lowe et al. Science (ref)

Performance of RNA-fold matching algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>CPU bp/sec</th>
<th>True pos.</th>
<th>False pos.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRNASCAN’91</td>
<td>400</td>
<td>95.1%</td>
<td>0.4x10^-6</td>
</tr>
<tr>
<td>TRNASCAN-SE '97</td>
<td>30,000</td>
<td>99.5%</td>
<td>&lt;7x10^-11</td>
</tr>
<tr>
<td>SnoRNAs'99</td>
<td></td>
<td>&gt;93%</td>
<td>&lt;10^-7</td>
</tr>
</tbody>
</table>

(See p. 258, 297 of Durbin et al.; Lowe et al 1999)

Putative Sno RNA gene disruption effects on rRNA modification

Primer extension pauses at 2’O-Me positions forming bands at low dNTP.

Lowe et al. Science 1999 283:1168-71 (ref)
RNA1: Structure & Quantitation

- Integration with previous topics (HMM & DP for RNA structure)
- Goals of molecular quantitation (maximal fold-changes, clustering & classification of genes & conditions/cell types, causality)
- Genomics-grade measures of RNA and protein and how we choose and integrate (SAGE, oligo-arrays, gene-arrays)
- Sources of random and systematic errors (reproducibility of RNA source(s), biases in labeling, non-polyA RNAs, effects of array geometry, cross-talk).
- Interpretation issues (splicing, 5' & 3' ends, gene families, small RNAs, antisense, apparent absence of RNA).
- Time series data: causality, mRNA decay

RNA (array) & Protein/metabolite (MS) quantitation

RNA measures are closer to genomic regulatory motifs & transcriptional control
Protein/metabolite measures are closer to Flux & growth phenotypes.

Integrating 8 regulon measures

In vitro array binding or selection
Protein fusions

In vivo crosslinking & selection (1-hybrid)

Conserved operons
Phylogenetic profiles
Coregulated sets of genes
Microarray data
Metabolic pathways

Check regulons from conserved operons (chromosomal proximity)

B. subtilis
purE purK purF purG purL purM purN purH purD

C. acetobutylicum
purE purC
purF purM purN purH purP

In E. coli, each color above is a separate but coregulated operon:

Predicting the PurR regulon by piecing together smaller operons

E. coli
purE purK
purM purN
purH purD

M. tuberculosis
purE
purM
purH

P. horokoshii
purM
purH

C. jejuni
purE
purM

M. janaschii
purM
purH

P. furiosus
purQ
purY

The above predicts regulon connections among these genes:

(Whole genome) RNA quantitation objectives

RNAs showing maximum change
minimum change detectable/meaningful

RNA absolute levels (compare protein levels)
minimum amount detectable/meaningful

Network-- direct causality - motifs

Classify (e.g. stress, drug effects, cancers)
**Subcellular inhomogeneity**

Dissected tissues have mixed cell types.

Cell-cycle differences in expression.

XIST RNA localized on inactive X-chromosome.

---

**Fluorescent in situ hybridization (FISH)**

- Time resolution: 1 msec
- Sensitivity: 1 molecule
- Multiplicity: >24
- Space: 10 nm (3-dimensional, in vivo)

10 nm accuracy with far-field optics energy-transfer fluorescent beads nanocrystal quantum dots, closed-loop piezo-scanner (ref).

---

**RNA1: Structure & Quantitation**

- Integration with previous topics (HMM & DP for RNA structure)
- Goals of molecular quantitation (maximal fold-changes, clustering & classification of genes & conditions/cell types, causality)
- Genomics-grade measures of RNA and protein and how we choose and integrate (SAGE, olio-arrays, gene-arrays)
- Sources of random and systematic errors (reproducibility of RNA source(s), biases in labeling, non-polyA RNAs, effects of array geometry, cross-talk).
- Interpretation issues (splicing, 5' & 3' ends, gene families, small RNAs, antisense, apparent absence of RNA).
- Time series data: causality, mRNA decay, time-warping.

---

**Steady-state population-average RNA quantitation methodology**

- Microarrays
  - ~1000 bp hybridization
  - R/G ratios
  - R, G values
  - quality indicators
- Affymetrix
  - 25-bp hybridization
  - Averaged PM-MM
  - "presence"
- SAGE
  - sequence counting
  - Counts of SAGE 14 to 22-mers sequence tags for each ORF
- MPSS

---

**GeneChip Expression Analysis**

GeneChip expression analysis probe array

- Each probe cell contains millions of copies of a specific oligonucleotide probe
- Biotinylated RNA from experiment
- Streptavidin-phycoerythrin conjugate

---

**Most RNAs < 1 molecule per cell.**

Reproducibility confidence intervals to find significant deviations. (ref)
**RNA array analyses**

- Image analysis: dCHIP, Wong/Affy, ArrayExos, NC-I-RRB
- Array viewer: TIGR, FP-SCAN, NIH
- Database & management: GeneX, MAPS, AMAD
- Statistics: AFM, AMIADA, Churchill, R-SMA, SAM, VERA-SAM ISB, R-Bioconductor

**Free vs Open:** OSI, FSF, ISCB

**Database & management**

- GeneX
- MAPS
- AMAD
- Cluster&Visualize
- CLUSFAVOR
- Cluster-TreeView
- GeneCluster WI
- J-Express
- PatE
- Fluid
- SVDBAN
- LANL
- TreeArrange
- Waterloo
- XCluster
- Stanfoed

**Statistics**

- AFM
- AMIADA
- Churchill
- R-SMA
- SAM
- VERA-SAM ISB
- R-Bioconductor

**“Significant” distributions**

- t-test: \( t = \frac{\text{Mean} \times \text{SD}}{\sqrt{N}} \)
- Degrees of freedom = N-1
- H0: The mean value of the difference = 0. If difference distribution is not normal, use the Wilcoxon Matched-Pairs Signed-Ranks Test

**Independent Experiments**


**RNA quantitation**

- Is less than a 2 fold RNA ratio ever important?
  - Yes; 1.5 fold in trisomies.
- Why oligonucleotides rather than cDNAs?
  - Alternative splicing, 5' & 3' ends; gene families.
- What about using a subset of the genome or ratios to a variety of control RNAs?
  - It makes trouble for later (meta) analyses.
**Whole genome RNA quantitation methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes immobilized labeled RNA</td>
<td>Chip manufacture</td>
</tr>
<tr>
<td>RNAs immobilized labeled genes</td>
<td>RNA sizes</td>
</tr>
<tr>
<td>Northern gel blot</td>
<td>Sensitivity $10^{-10}$</td>
</tr>
<tr>
<td>QRT-PCR</td>
<td>No crosshybridization</td>
</tr>
<tr>
<td>Reporter constructs</td>
<td>Spatial relations</td>
</tr>
<tr>
<td>Fluorescent In Situ Hybridization</td>
<td>Gene discovery</td>
</tr>
<tr>
<td>Tag counting (SAGE)</td>
<td>&quot;Selective&quot; discovery</td>
</tr>
<tr>
<td>Differential display &amp; subtraction</td>
<td></td>
</tr>
</tbody>
</table>

**Genomic oligonucleotide microarrays**

- 295,936 oligonucleotides (including controls)
- Intergenic regions: ~6bp spacing
- Genes: ~70 bp spacing
- Not polyA (or 3' end) biased

**Strengths:** Gene family paralogs, RNA fine structure (adjacent promoters), untranslated & antisense RNAs, DNA-protein interactions.

- E. coli 25-mer array
- Non-coding sequences (12% of genome)
- Protein coding 25-mers
- tRNAs, rRNAs

**Random & Systematic Errors in RNA quantitation**

- Secondary structure
- Position on array (mixing, scattering)
- Amount of target per spot
- Cross-hybridization
- Unanticipated transcripts

**Detection of Antisense and Untranslated RNAs**

- Expression Chip
- Reverse Complement Chip

- b0671 - ORF of unknown function, tiled in the opposite orientation
- "intergenic region 1725" - is actually a small untranslated RNA (csrB)
Mapping deviations from expected repeat ratios

Li & Wong

RNA1: Structure & Quantitation

Integration with previous topics (HMM & DP for RNA structure)
Goals of molecular quantitation (maximal fold-changes, clustering & classification of genes & conditions/cell types, causality)
Genomics-grade measures of RNA and protein and how we choose and integrate (SAGE, oligo-arrays, gene-arrays)
Sources of random and systematic errors (reproducibility of RNA source(s), biases in labeling, non-polyA RNAs, effects of array geometry, cross-talk).
Interpretation issues (splicing, 5' & 3' ends, gene families, small RNAs, antisense, apparent absence of RNA).
Time series data: causality, mRNA decay, time-warping

Independent oligos analysis of RNA structure

Selinger et al

Predicting RNA-RNA interactions

Selinger et al

of the human genome using microarray technology.
Time courses

- To discriminate primary vs secondary effects we need conditional gene knockouts.
- Conditional control via transcription/translation is slow (>60 sec up & much longer for down regulation)
- Chemical knockouts can be more specific than temperature (ts-mutants).

Beyond steady state: mRNA turnover rates (rifampicin time-course)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chip half life (min)</th>
<th>Northern half life (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>capE</td>
<td>2.4</td>
<td>&gt;20</td>
</tr>
<tr>
<td>lpp</td>
<td>&gt;20</td>
<td>&gt;300</td>
</tr>
</tbody>
</table>

Chip metric = $S_{max}$

TimeWarp: pairs of expression series, discrete or interpolative

Asch & Church

TimeWarp: cell-cycle experiments

RNA1: Structure & Quantitation

- Integration with previous topics (HMM & DP for RNA structure)
- Goals of molecular quantitation (maximal fold-changes, clustering & classification of genes & conditions/cell types, causality)
- Genomics-grade measures of RNA and protein and how we choose and integrate (SAGE, oligo-arrays, gene-arrays)
- Sources of random and systematic errors (reproducibility of RNA, biases in labeling, non-polyA RNAs, effects of array geometry, cross-talk)
- Interpretation issues (splicing, 5' & 3' ends, gene families, small RNAs, antisense, apparent absence of RNA)
- Time series data: causality, mRNA decay, time-warping