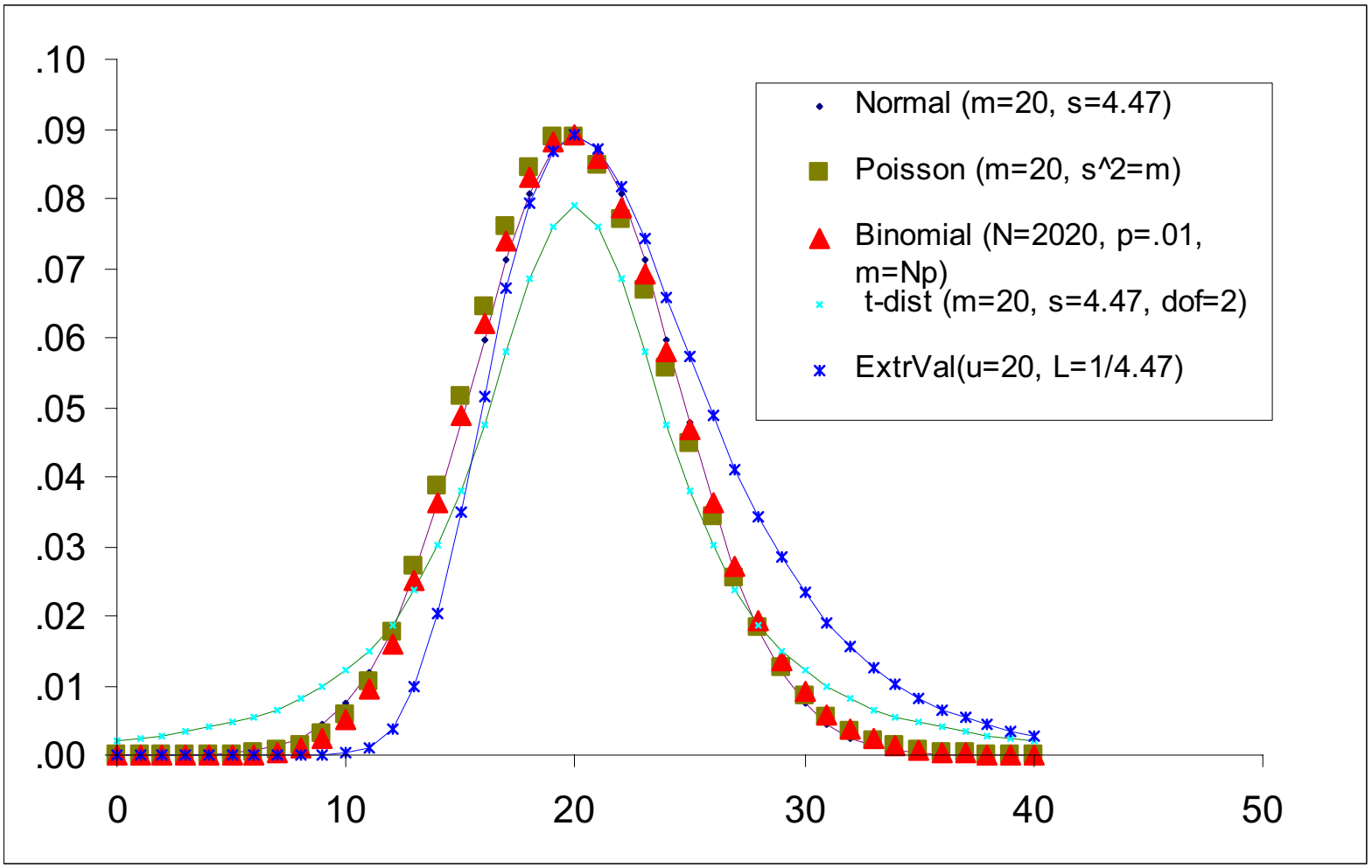
DNA2: Last week's take home lessons

- 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: Today's story & goals

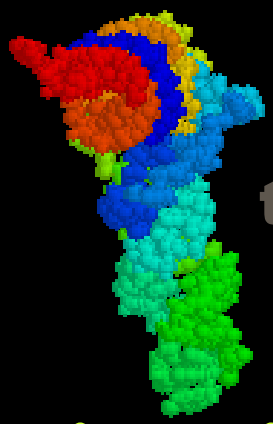
- Integration with previous topics (HMM & DP for **RNA structure**)
- Goals of molecular **quantitation** (maximal fold-changes, clustering & classification of genes & conditions/cell types, causality)
- Genomics-grade **measures** of RNA and protein and how we choose and integrate (SAGE, oligo-arrays, gene-arrays)
- Sources of random and systematic **errors** (reproducibilty 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



gggatttagctcagtt

gggagagcgccagact

gaa  gat

Primary
to tertiary
structure

ttg  gag

gtcctgtgttcgatcc

acagaattcgcacca

Non-watson-crick bps

See **(http://www.imb-jena.de/IMAGE_BPDIR.html)**

Modified bases & bps in RNA

See http://info.bio.cmu.edu/Courses/BiochemMols/tRNA_Tour/tRNA_Tour.html

ref

Covariance

see Durbin et al p. 266-8.

Mutual Information

ACUUA**U**
CCUUA**G**
GCUUG**C**
UCUUG**A**
i=1 j=6

$M_{1,6} = \sum_{x_1 x_6} = fAU \log_2[fAU/(fA*fU)] \ldots$

$= 4*.25\log_2[.25/(.25*.25)] = 2$

$M_{1,2} = 4*.25\log_2[.25/(.25*1)] = 0$

$M_{ij} = \sum_{x_i x_j} fx_i x_j \log_2[fx_i x_j/(fx_i fx_j)]$

M=0 to 2 bits; x=base type
see Durbin et al p. 266-8.

See Shannon entropy, multinomial Grendar (http://xxx.lanl.gov/pdf/math-ph/0009020)

$$\pi(\mathbf{n}|\mathbf{q}) = \frac{n!}{n_1!\, n_2! \ldots n_m!} \prod_{i=1}^{m} q_i^{n_i}$$

RNA secondary structure prediction

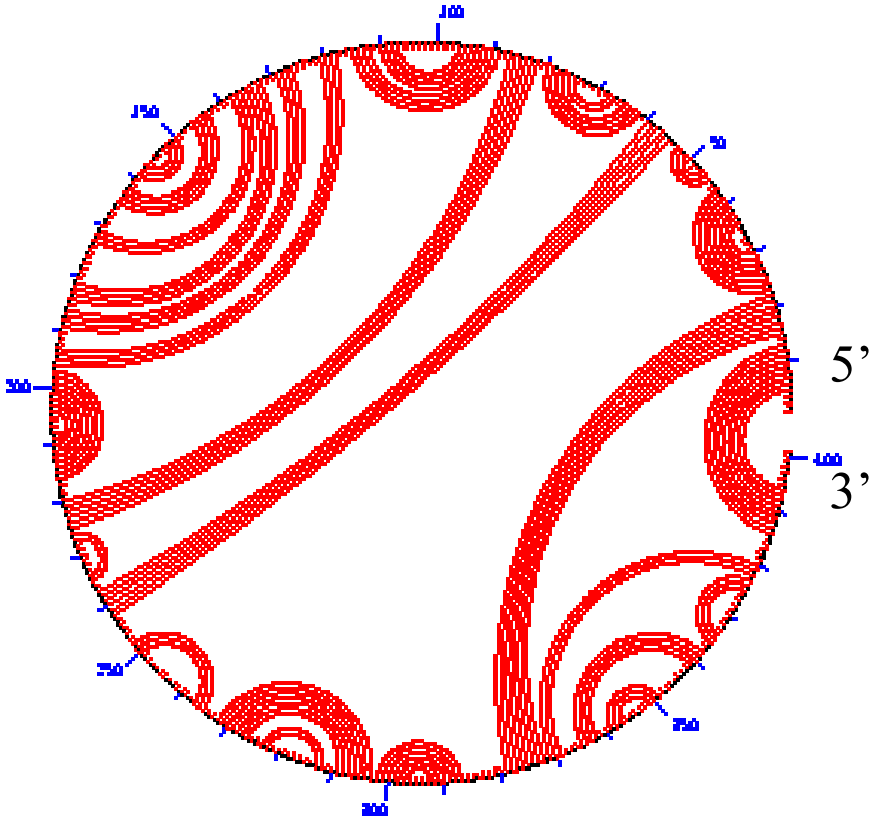
Mathews DH, Sabina J, Zuker M, Turner DH J Mol Biol 1999 May 21;288(5):911-40

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.

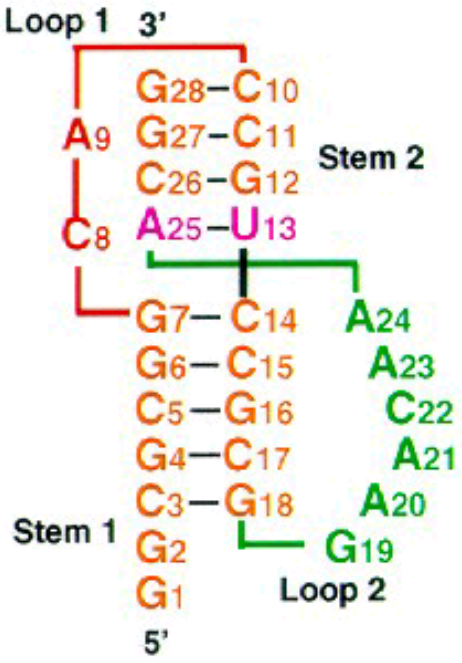
Stacked bp & ss

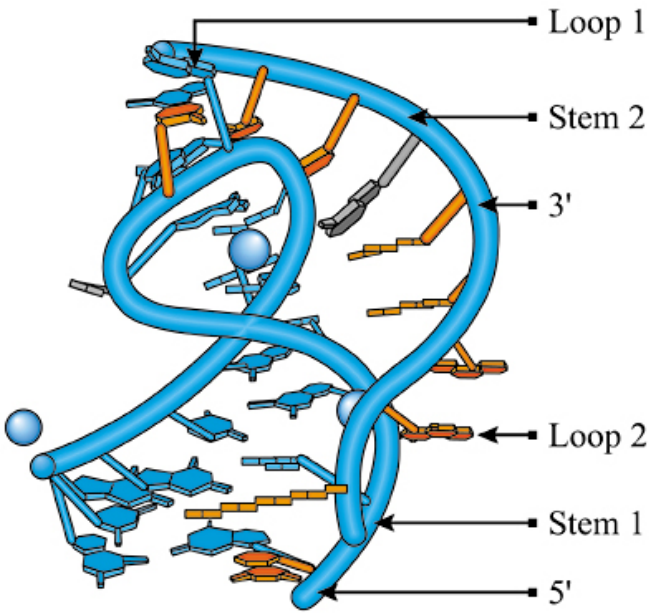
Initial 1981 O(N^2) DP methods: Circular Representation of RNA Structure



Did not handle pseudoknots

RNA pseudoknots, important biologically, but challenging for structure searches





Dynamic programming finally handles RNA pseudoknots too.

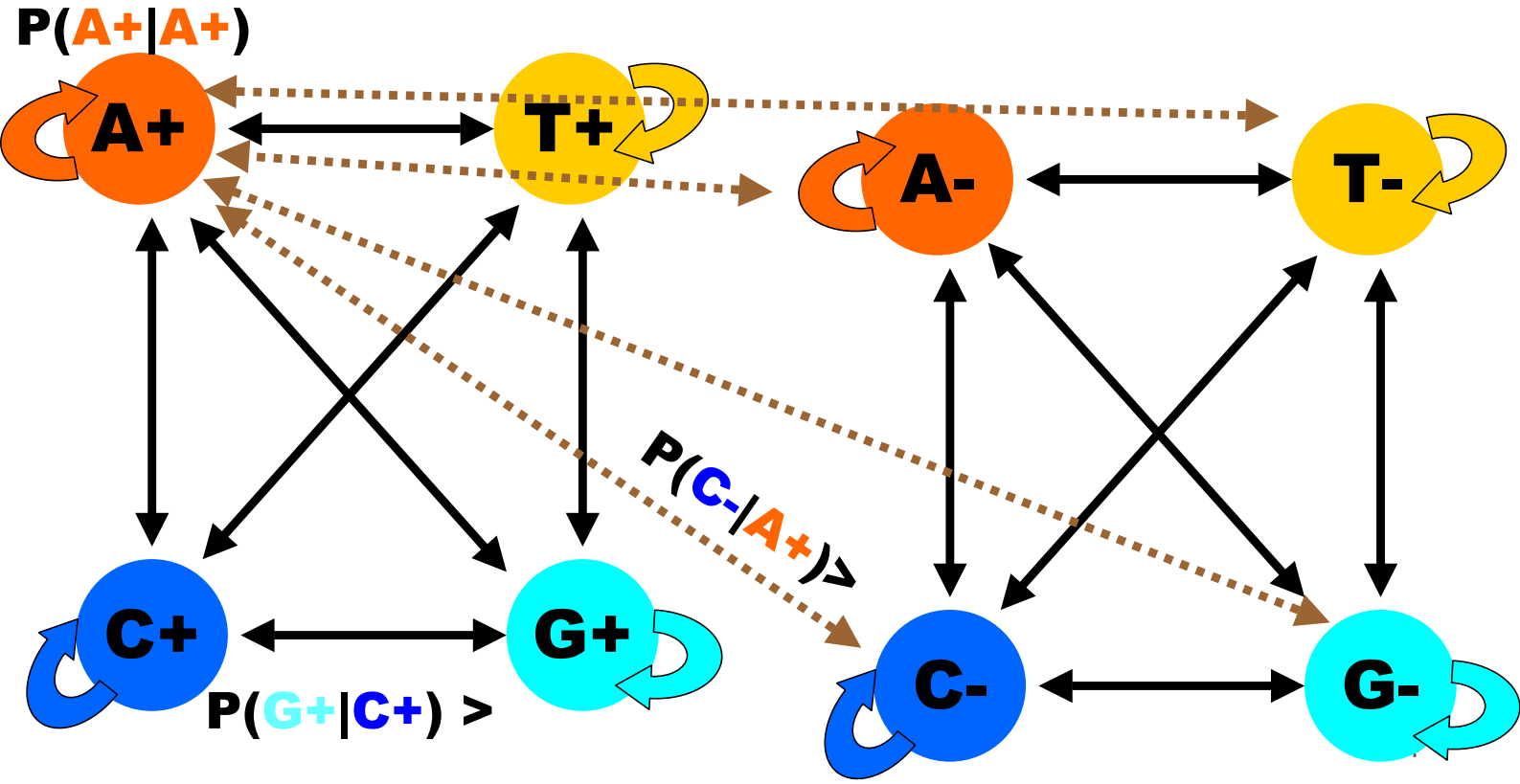
Rivas E, Eddy SR J Mol Biol 1999 Feb 5;285(5):2053-68 A dynamic programming algorithm for RNA structure prediction including
pseudoknots. (ref)
(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=9925784&dopt=Abstract)

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

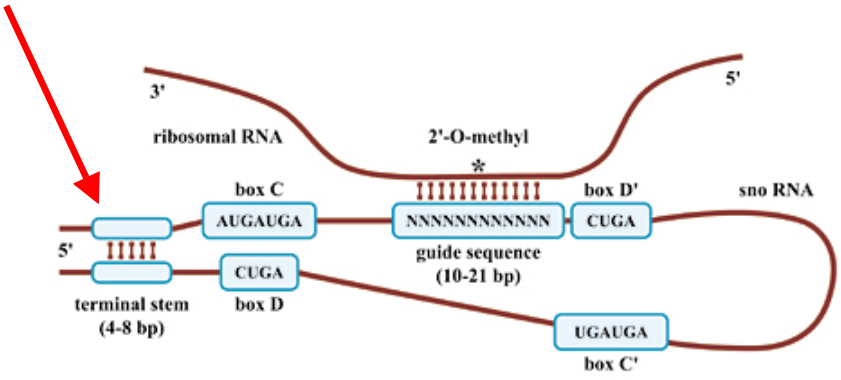
Bioinformatics 2000 Apr;16(4):334-40 (ref)
(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=10869031&dopt=Abstract)

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

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

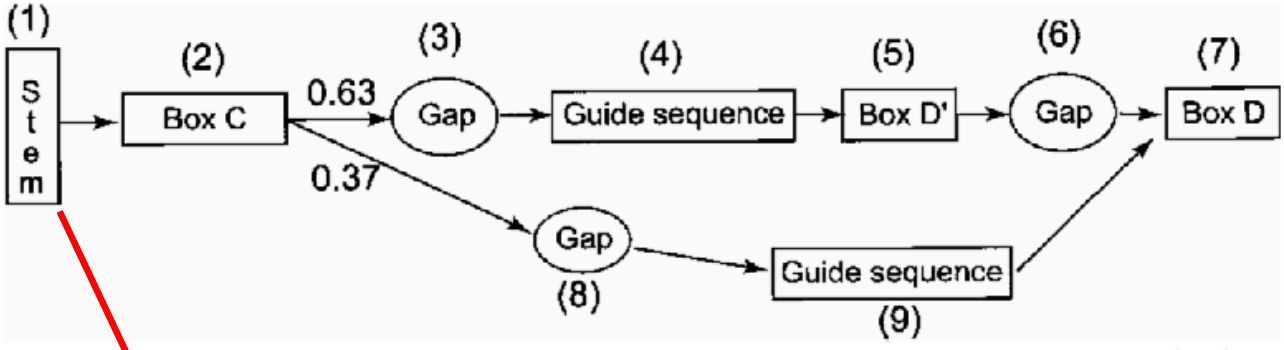


Small nucleolar (sno)RNA structure & function



See Lowe et al. Science (ref)

SnoRNA Search



State number	Feature	Model	Consensus	Feature score (bits)		
				Best	Average	Worst
1	Terminal stem	SCFG, 4 to 8 bp	6 bp (when present)	7.60	3.09	0.35
2	Box C	7-bp ungapped HMM	AUGAUGA	12.73	11.63	5.84
3	Gap	Duration model	Length: 6 to 10 bp	−1.59	−2.09	−4.76
4	Guide sequence	HMM	12-bp duplex	15.67	11.11	2.54
5	Box D′	4-bp ungapped HMM	CUGA	7.34	4.85	−3.74
6	Gap	Duration model	Length: 36 to 45 bp	−1.59	−2.43	−5.36
7	Box D	4-bp ungapped HMM	CUGA	8.05	7.92	5.43
8	Gap	Duration model	Length: 56 to 75 bp	−1.50	−2.10	−4.17
9	Guide sequence	HMM	14-bp duplex	18.96	13.98	9.95

Performance of RNA-fold matching algorithms

Algorithm	CPU bp/sec	True pos.	False pos.
TRNASCAN'91	400	95.1%	0.4×10^{-6}
TRNASCAN-SE '97	30,000	99.5%	$< 7 \times 10^{-11}$
SnoRNAs'99		>93%	$< 10^{-7}$

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

Putative Sno RNA gene disruption effects on rRNA modification

See Lowe et al. Science 1999 283:1168-71 (ref)
(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=10024243&dopt=Abstract)

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

RNA1: Today's story & goals

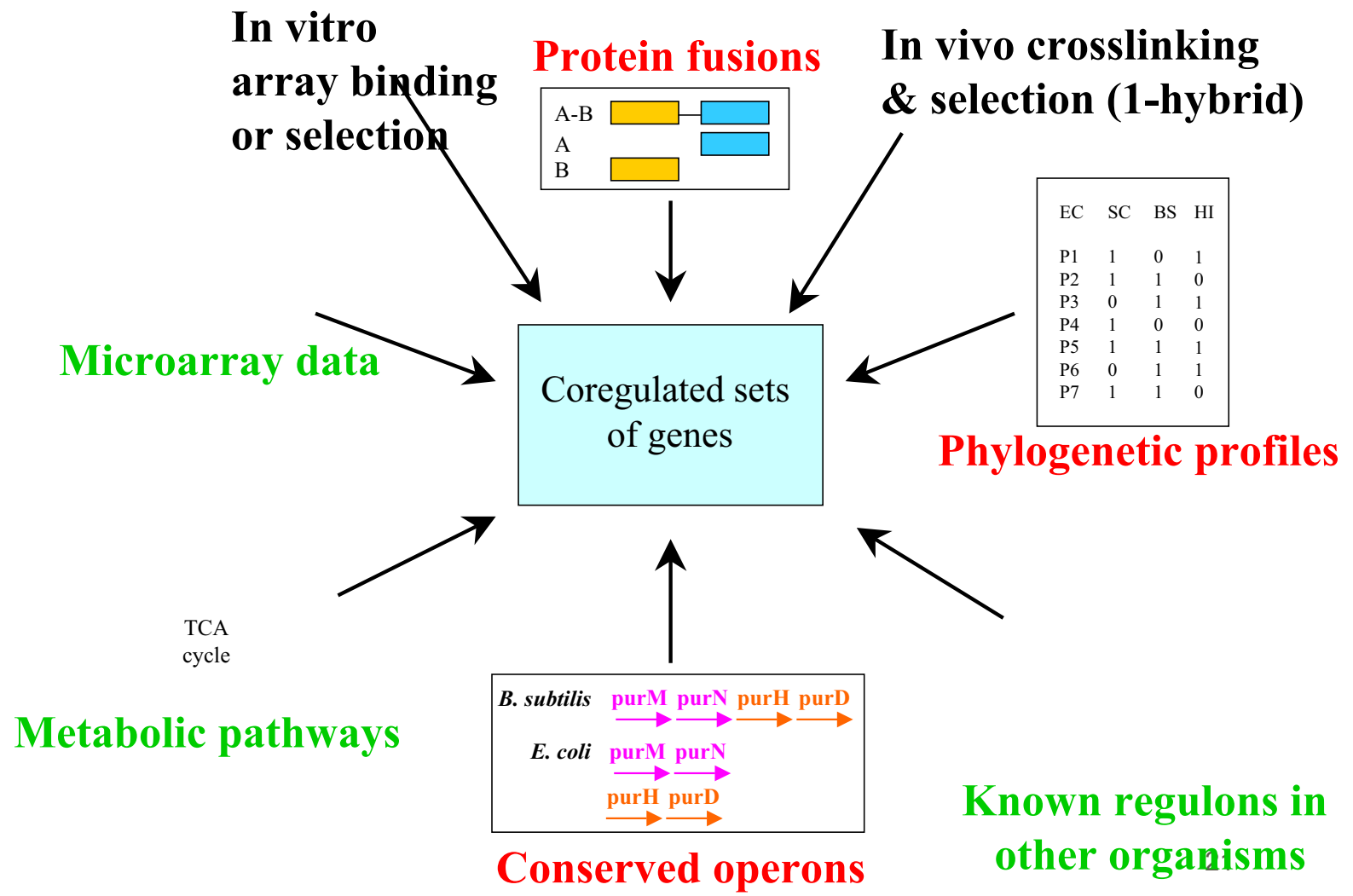
- Integration with previous topics (HMM & DP for **RNA structure**)
- Goals of molecular **quantitation** (maximal fold-changes, clustering & classification of genes & conditions/cell types, causality)
- Genomics-grade **measures** of RNA and protein and how we choose and integrate (SAGE, oligo-arrays, gene-arrays)
- Sources of random and systematic **errors** (reproducibilty 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

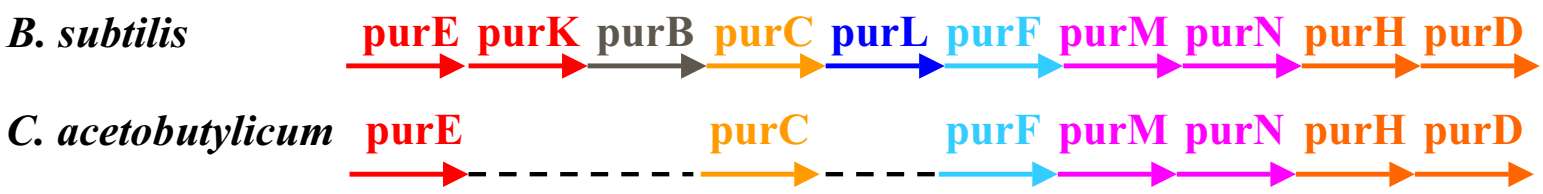
RNA measures are closer to genomic regulatory motifs & transcriptional control

Protein/metabolite measures are closer to Flux & growth phenotypes.

8 cross-checks for regulon quantitation

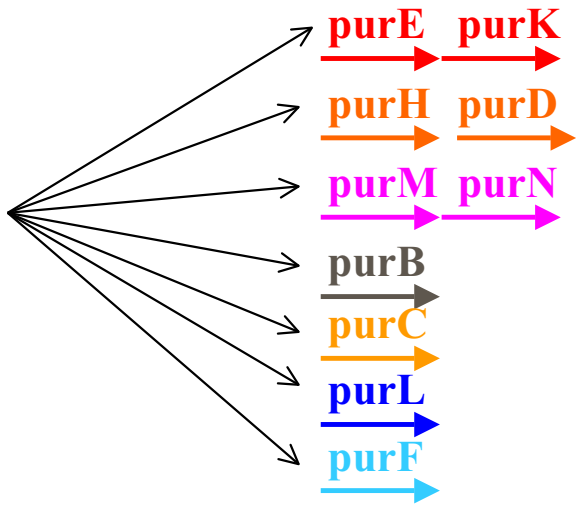


Check regulons from conserved operons (chromosomal proximity)



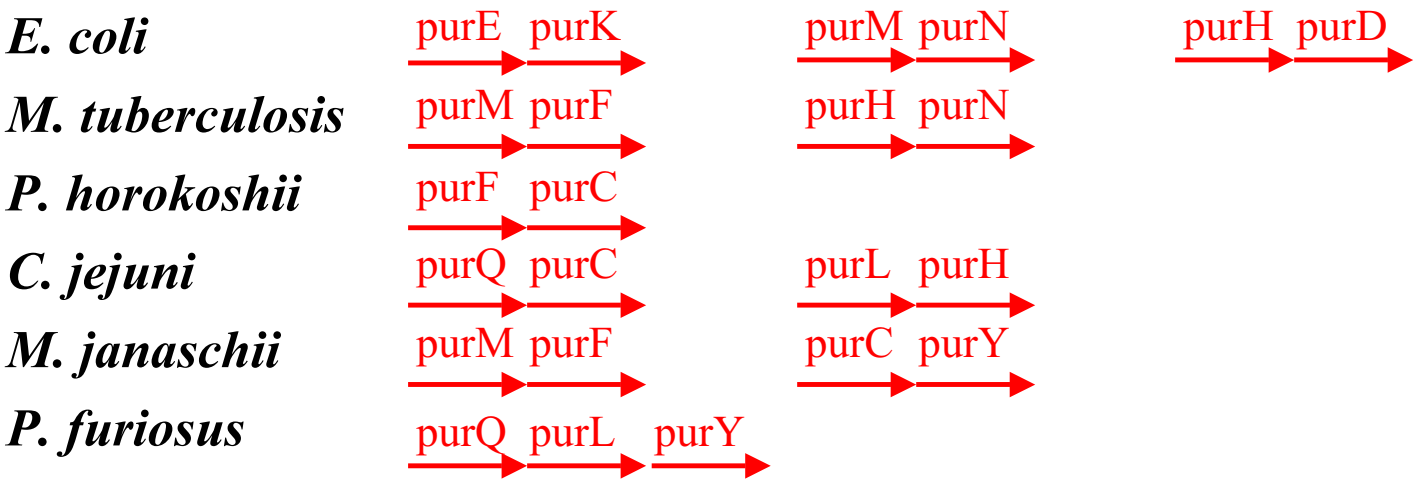
In ***E. coli***, each color above is a separate but coregulated operon:

E. coli PurR motif

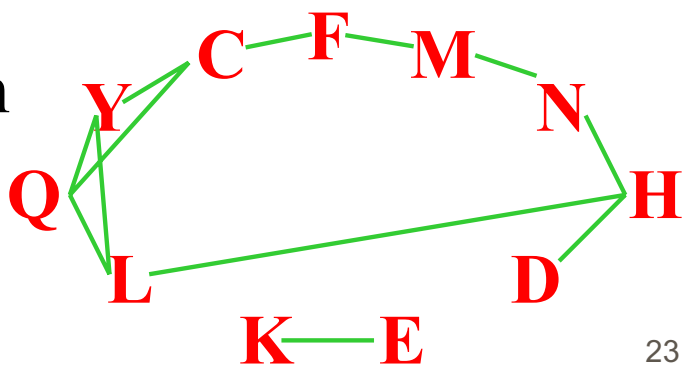


Predicting regulons and their cis-regulatory motifs by comparative genomics. Mcguire & Church, (2000) Nucleic Acids Research 28:4523-30.

Predicting the PurR regulon by piecing together smaller operons



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)

(Sub)cellular inhomogeneity

Dissected tissues have mixed cell types.

Cell-cycle differences in expression.

XIST RNA localized on inactive X-chromosome

(see figure)
(http://www.med.ic.ac.uk/dc/Brochure/XI/XI.html)

Fluorescent in situ hybridization (FISH)

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

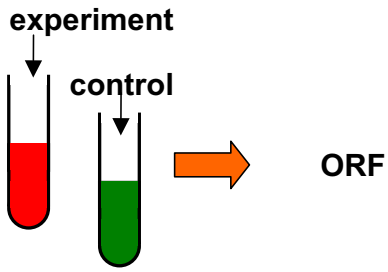
10 nm accuracy with far-field optics energy-transfer fluorescent beads nanocrystal quantum dots,closed-loop piezo-scanner (ref)
(http://www.pnas.org/cgi/content/full/97/17/9461)

RNA1: Today's story & goals

- 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** (reproducibilty 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

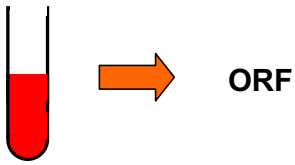
Steady-state population-average RNA quantitation methodology

Microarrays[1]
~1000 bp hybridization



- R/G ratios
- R, G values
- quality indicators

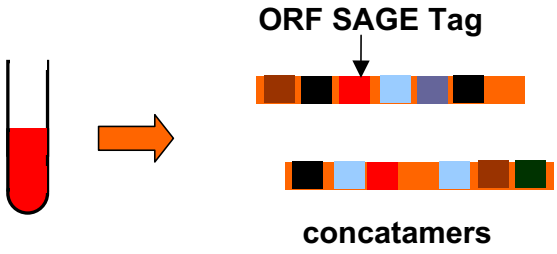
Affymetrix[2]
25-bp hybridization



- Averaged PM-MM
- "presence"

SAGE[3]
sequence counting

MPSS[4]



- Counts of SAGE

  14 to 22-mers sequence tags for each ORF

[1] see DeRisi, et.al., *Science* **278**:680-686 (1997)
[2] see Lockhart, et.al., *Nat Biotech* **14**:1675-1680 (1996)
[3] see Velculescu, et.al, Serial Analysis of Gene Expression, *Science* **270**:484-487 (1995)
4 Brenner et al,

Most RNAs < 1 molecule per cell.

See Yeast RNA

25-mer array in

Wodicka, Lockhart, et al. (1997)
Nature Biotech 15:1359-67

Reproducibility
confidence intervals
to find significant
deviations.

(ref) **(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=9415887&dopt=Abstract)**

Microarray data analyses

(web)**(http://linkage.rockefeller.edu/wli/microarray/soft.html)**

AFM
AMADA
Churchill
CLUSFAVOR
CLUSTER,
D-CHIP
GENE-CLUSTER
J-EXPRESS
PAGE
PLAID
SAM
SMA
SVDMAN
TREE-ARRANGE & TREEPS
VERA &SAM
XCLUSTER
ArrayTools
ARRAY-VIEWER
F-SCAN
P-SCAN
SCAN-ALYZE
GENEX
MAPS

Statistical models for repeated array data

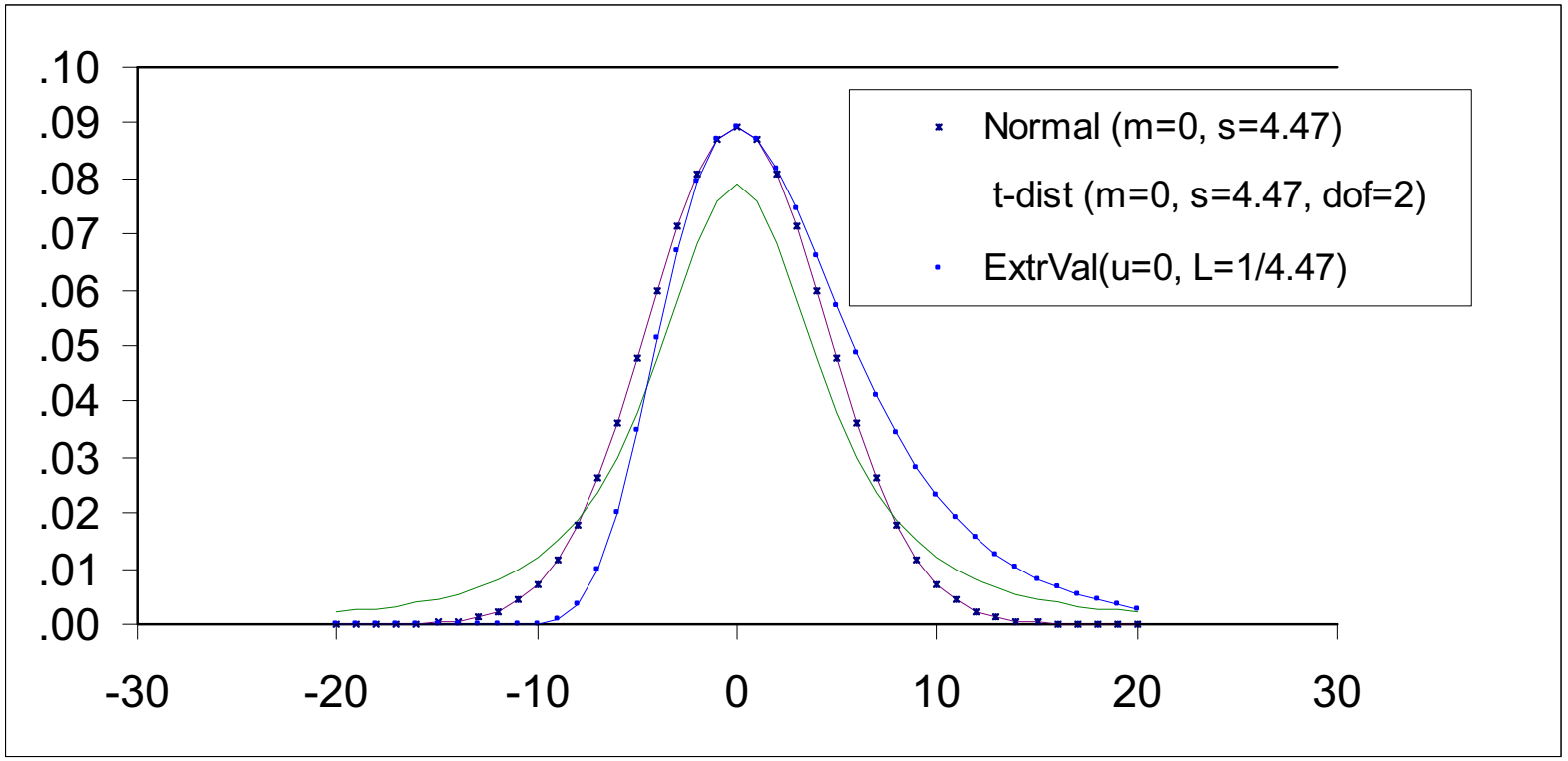
Tusher, Tibshirani and Chu (2001) Significance analysis of microarrays applied to the ionizing radiation response. PNAS 98(9):5116-21. **(http://www-stat.stanford.edu/~tibs/SAM/pnassam.pdf)**

Selinger, et al. (2000) RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. Nature Biotech. 18, 1262-7. **(http://arep.med.harvard.edu/pdf/Selinger00.pdf)**

Li & Wong (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol 2(8):0032 **(http://genomebiology.com/2001/2/8/research/0032/)**

Kuo et al. (2002) Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 18(3):405-12

“Significant” distributions



t-test $t = (Mean / SD) * \sqrt{N}$. 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.

(http://fonsg3.let.uva.nl/Service/Statistics/Signed_Rank_Test.html?values=)

Independent Experiments

See Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene Crx.
Livesay, et al. (2000) Current Biology

RNA quantitation

Is less than a 2-fold RNA-ratio ever important?

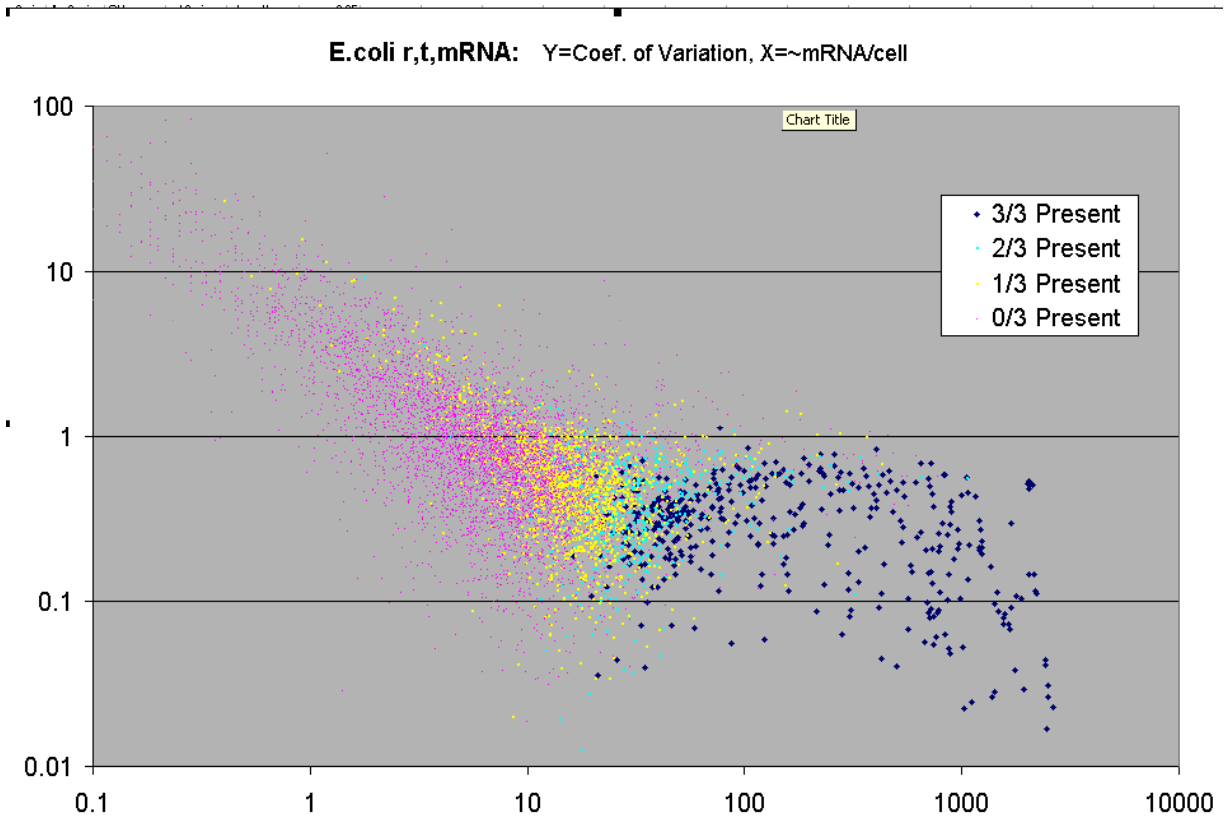
Yes; 1.5-fold in trisomies.

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?

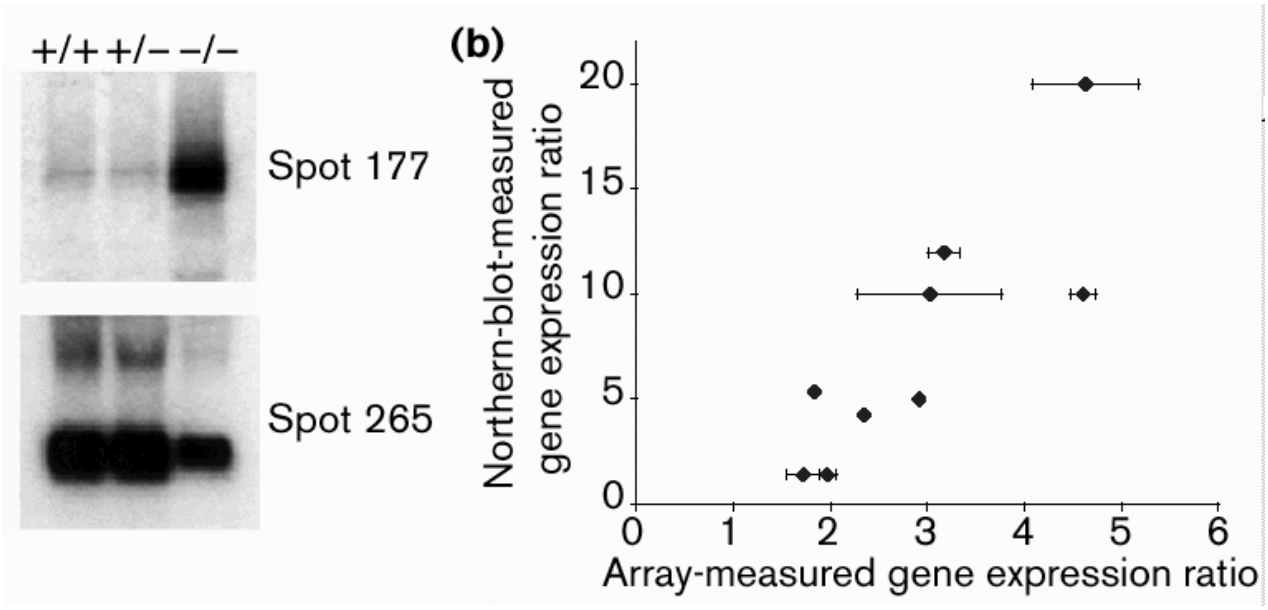
It makes trouble for later (meta) analyses.


E.coli r,t,mRNA: Y=Coef. of Variation, X=~mRNA/cell
Chart Title
3/3 Present
2/3 Present
1/3 Present
0/3 Present
100
10
1
0.1
0.01
0.1
1
10
100
1000
10000

(Whole genome) RNA quantitation methods

Method	Advantages
Genes immobilized labeled RNA	Chip manufacture
RNAs immobilized labeled genes- Northern gel blot	RNA sizes
QRT-PCR	Sensitivity 1e-10
Reporter constructs	No crosshybridization
Fluorescent In Situ Hybridization	Spatial relations
Tag counting (SAGE)	Gene discovery
Differential display & subtraction	"Selective" discovery

Microarray to Northern

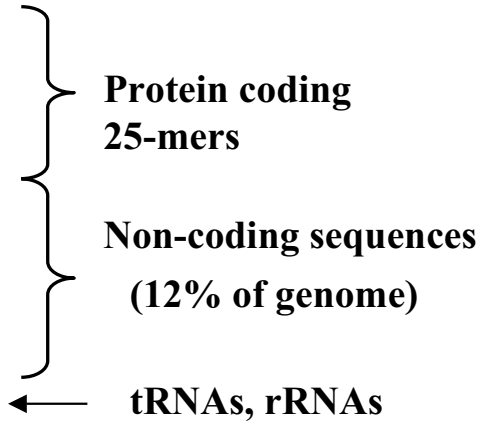


Genomic oligonucleotide microarrays

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

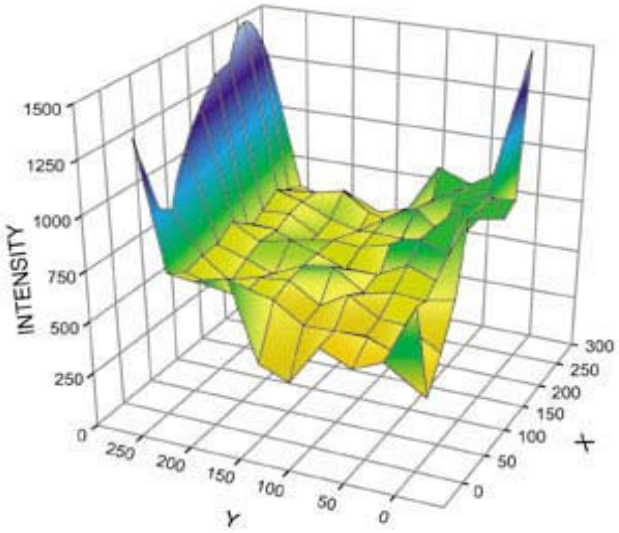


Affymetrix: Mei, Gentalen, Johansen, Lockhart(Novartis Inst)
HMS: Church, Bulyk, Cheung, Tavazoie, Petti, Selinger

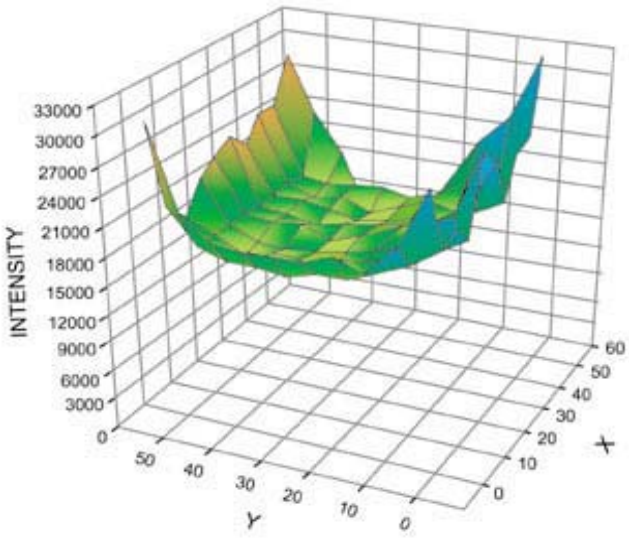
Random & Systematic Errors in RNA quantitation

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

Spatial Variation in Control Intensity



Experiment 1

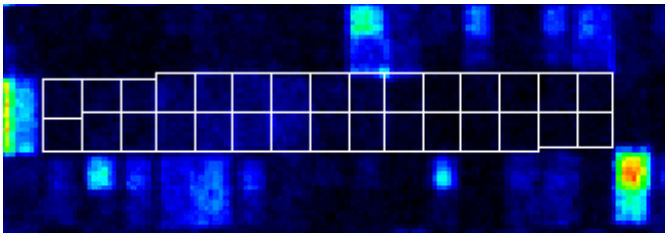


experiment 2

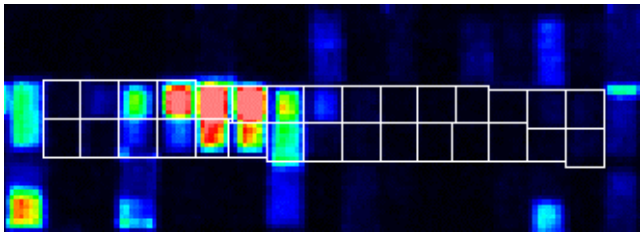
See Selinger et al

Detection of Antisense and Untranslated RNAs

Expression Chip

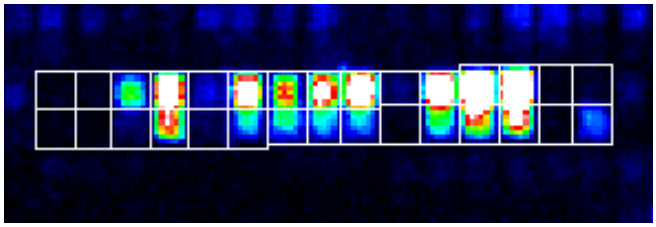


Reverse Complement Chip

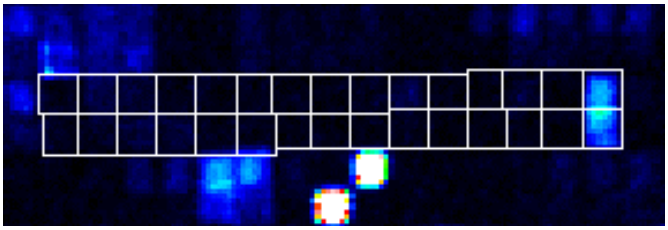


b0671 - ORF of unknown function, tiled in the opposite orientation

Crick Strand



Watson Strand (same chip)



“intergenic region 1725” - is actually a small untranslated RNA (*csrB*)

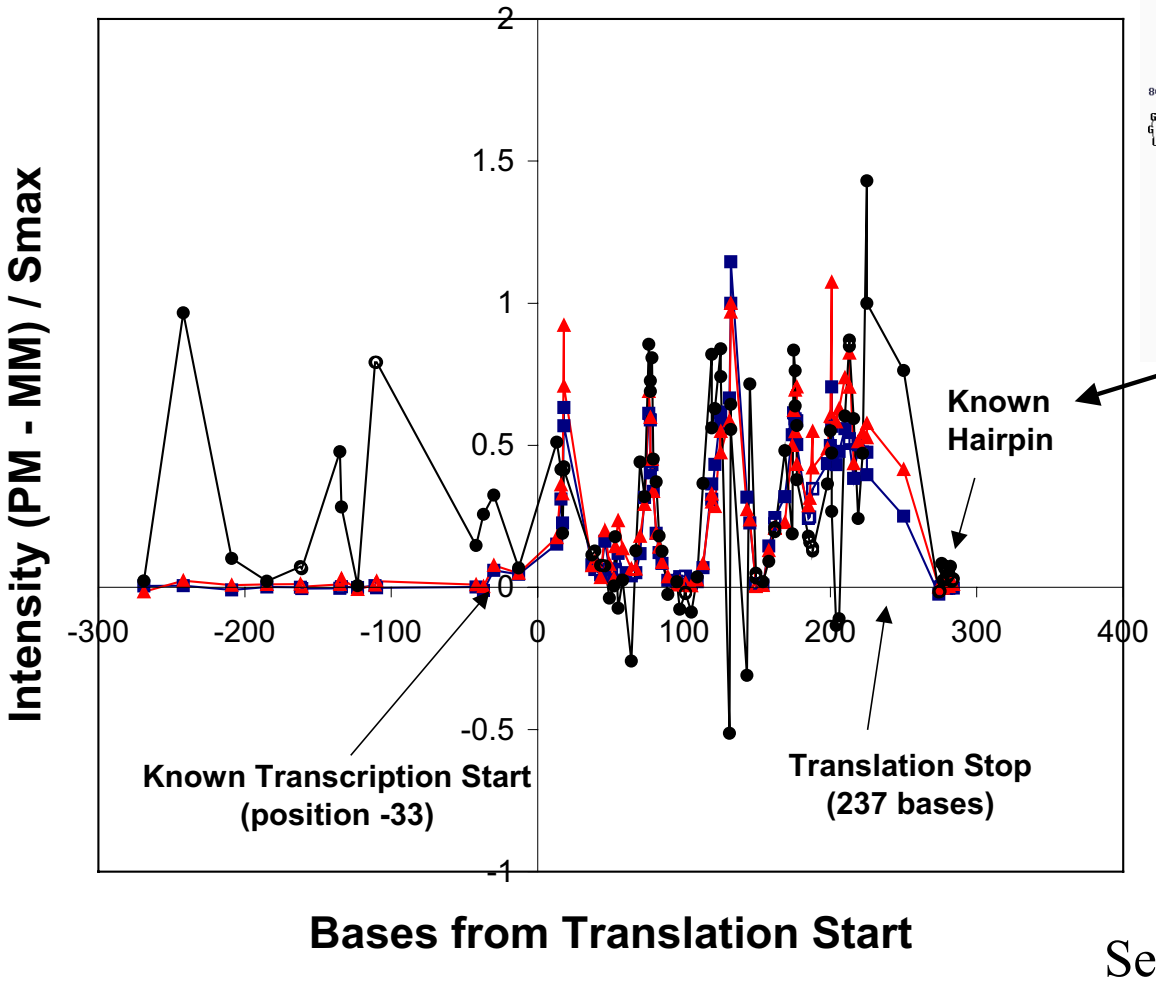
Mapping deviations from expected repeat ratios

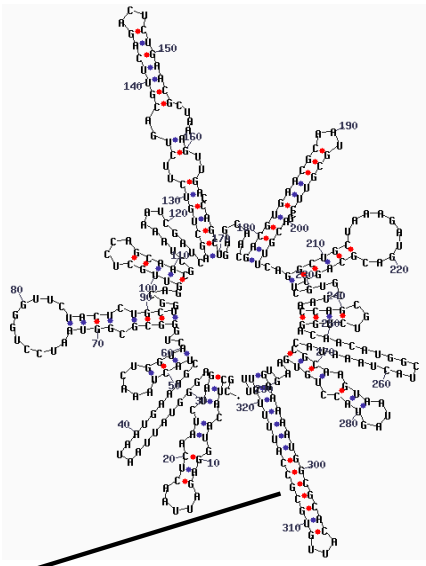
See Li & Wong

RNA1: Today's story & goals

- 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** (reproducibilty 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

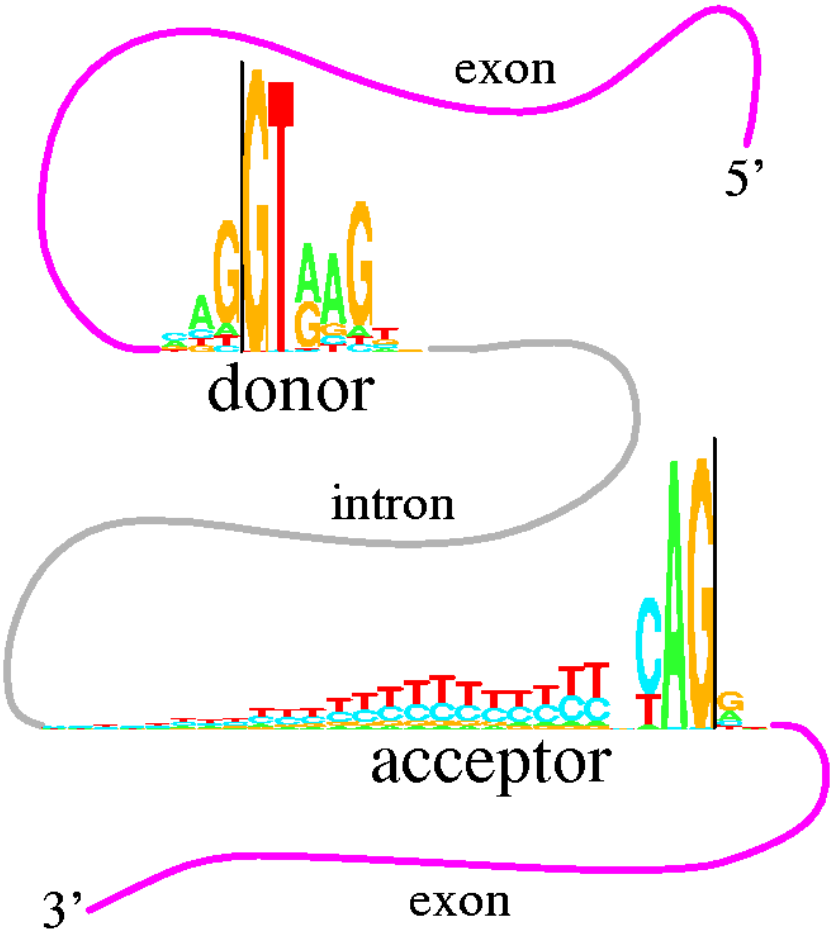




Log
Stationary
Genomic DNA

See Selinger et al

Predicting RNA-RNA interactions



Human RNA-splice junctions sequence matrix

http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html

Experimental annotation of the human genome using microarray technology.

See Shoemaker, et al. (2001) Nature 409:922-7.

(http://www.nature.com/cgitaf/DynaPage.taf?file=/nature/journal/v409/n6822/full/409922a0_fs.html&content_filetype=pdf)

RNA1: Today's story & goals

- 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** (reproducibilty 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

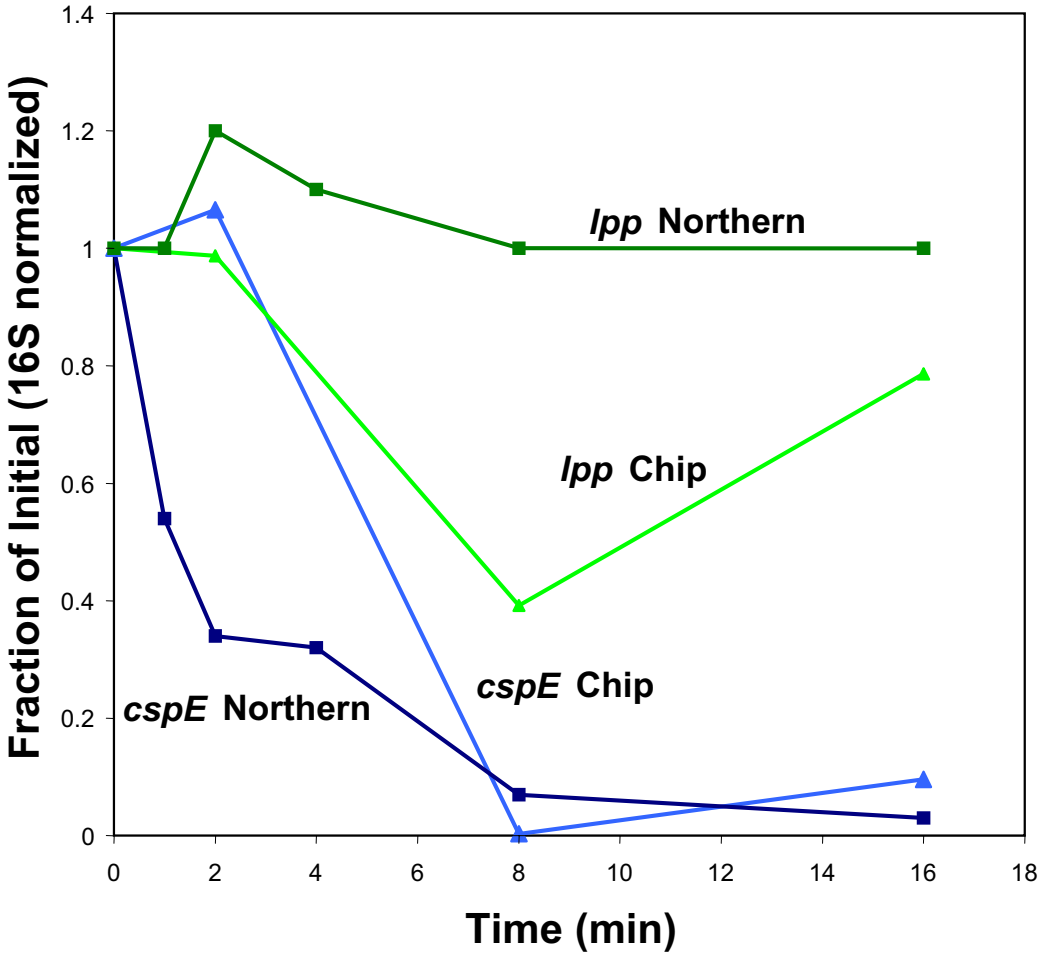
Time courses

•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)



cspE Chip
lpp Chip
cspE Northern
lpp Northern

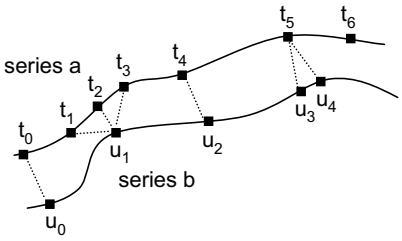
cspE	half life
chip	2.4 min
Northern	2.9 min

lpp	half life
chip	>20 min
Northern	>300 min

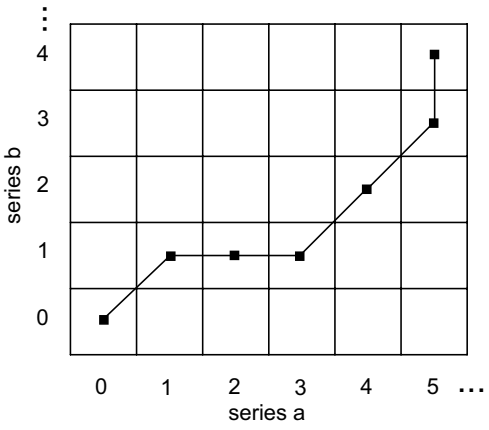
Chip metric = Smax

TimeWarp: pairs of expression series, discrete or interpolative

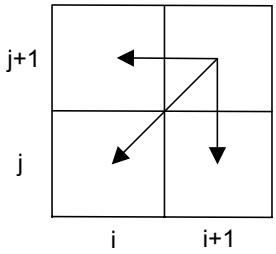
a



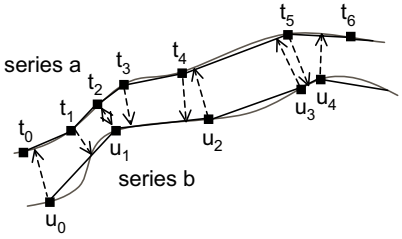
b



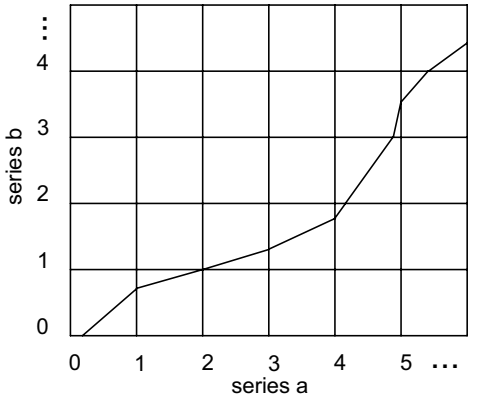
c



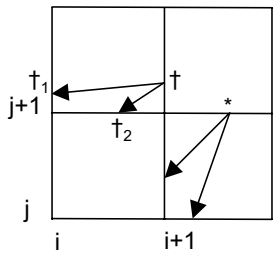
d



e



f



See Aach & Church

TimeWarp: cell-cycle experiments

TimeWarp: alignment example

RNA1: Today's story & goals

- 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** (reproducibilty 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