生物信息学

高通量技术

吴凌云

中国科学院数学与系统科学研究院

Microarray

- "**Microarray**" has become a general term, there are many types now
  - DNA microarrays
  - Protein microarrays
  - Transfection microarrays
  - Tissue microarray
  - …
- We'll be discussing **cDNA** microarrays

DNA Microarray

- A grid of DNA spots (probes) on a substrate used to detect complementary sequences
- The DNA spots can be deposited by
  - piezoelectric (ink jet style)
  - Pen
  - **Photolithography**光刻法(Affymetrix)
- The substrate can be plastic, glass, **silicon** (Affymetrix)
- RNA/DNA of interest is labelled & hybridizes with the array
- Hybridization with probes is detected optically.

Types of DNA Microarrays

- What is measured depends on the chip design and the laboratory protocol:
  - Expression
    - Measure mRNA expression levels (usually polyadenylated mRNA)
  - Re-sequencing
    - Detect changes in genomic regions of interest
  - Tiling
    - Tiles probes over an entire genome for various applications (novel transcripts, ChIP, epigenetic modifications)
  - SNP
    - Detect which known SNPs are in the tested DNA
  - ?...

Expression Arrays

- Gene Expression
- mRNA levels in a cell
- mRNA levels averaged over a population of cells in a sample
- relative mRNA levels averaged over populations of cells in multiple samples
- relative mRNA hybridization readings averaged over populations of cells in multiple samples
- some relative mRNA hybridization readings averaged over populations of cells in multiple samples

Why “some”

- **“some”**
  - In a comparison of Affymetrix vs spotted arrays, 10% of probesets yielded very different results.
  - “In the small number of cases in which platforms yielded discrepant results, qRT-PCR generally did not confirm either set of data, suggesting that sequence-specific effects may make expression predictions difficult to make using any technique.”*
  - *It appears that <u>some</u> transcripts just can’t be detected accurately by these techniques*.

* Independence and reproducibility across microarray platforms., Quackenbush et al. Nat Methods. 2005 May;2(5):337-44

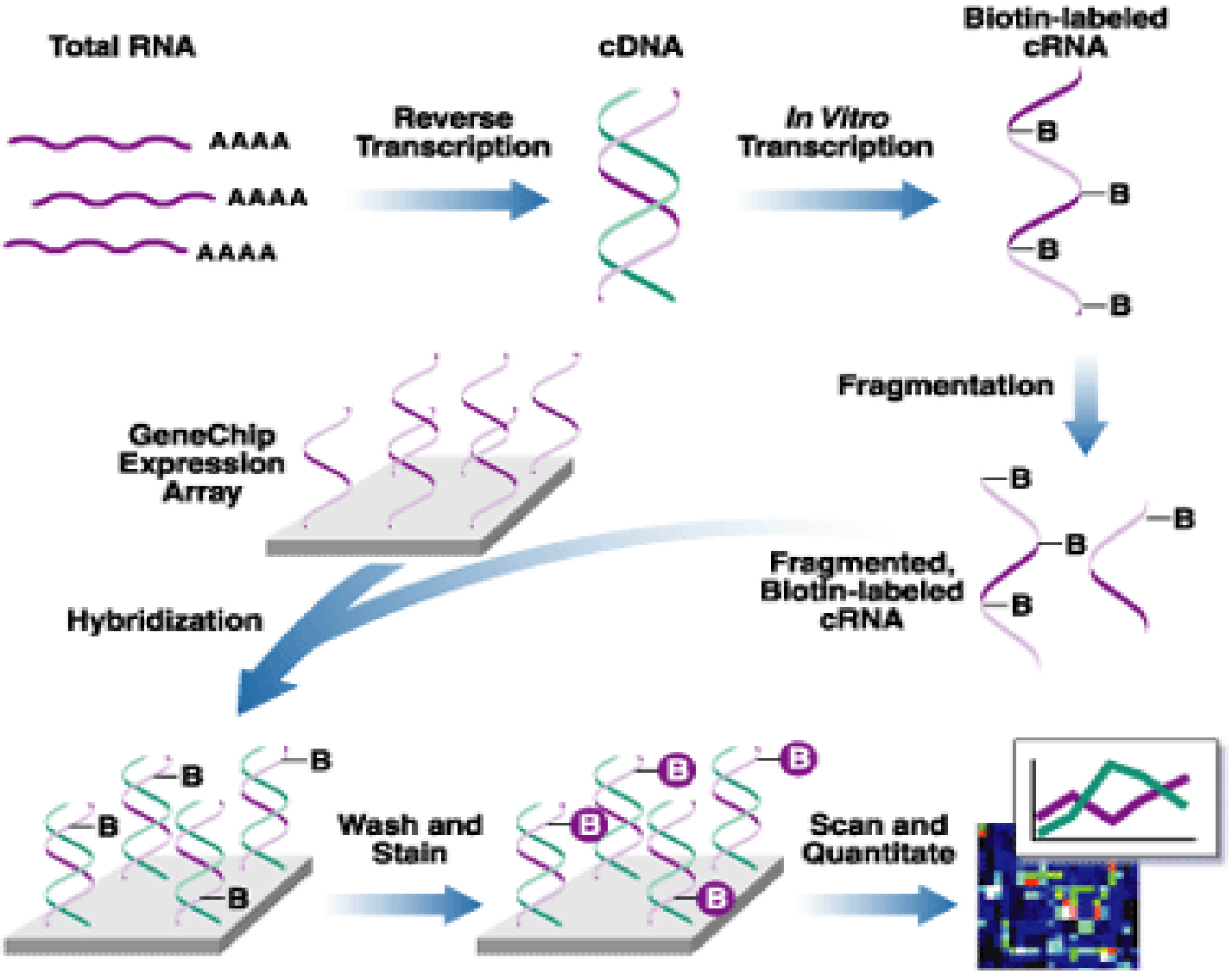
Why “multiple samples”

- **“multiple samples”**
  - We can only really depend on between-sample fold change for Microarrays not absolute values or within sample comparisons (>1.3-2.0 fold change, in general)

Central "*Assumption*" of Gene Expression Microarrays

- The level of a given mRNA is positively correlated with the expression of the associated protein.
  - Higher mRNA levels mean higher protein expression, lower mRNA means lower protein expression
- Other factors:
  - Protein degradation, mRNA degradation, polyadenylation, codon preference, translation rates, alternative splicing, translation lag…
- *This is relatively obvious, but worth emphasizing*

Affymetrix Expression Arrays

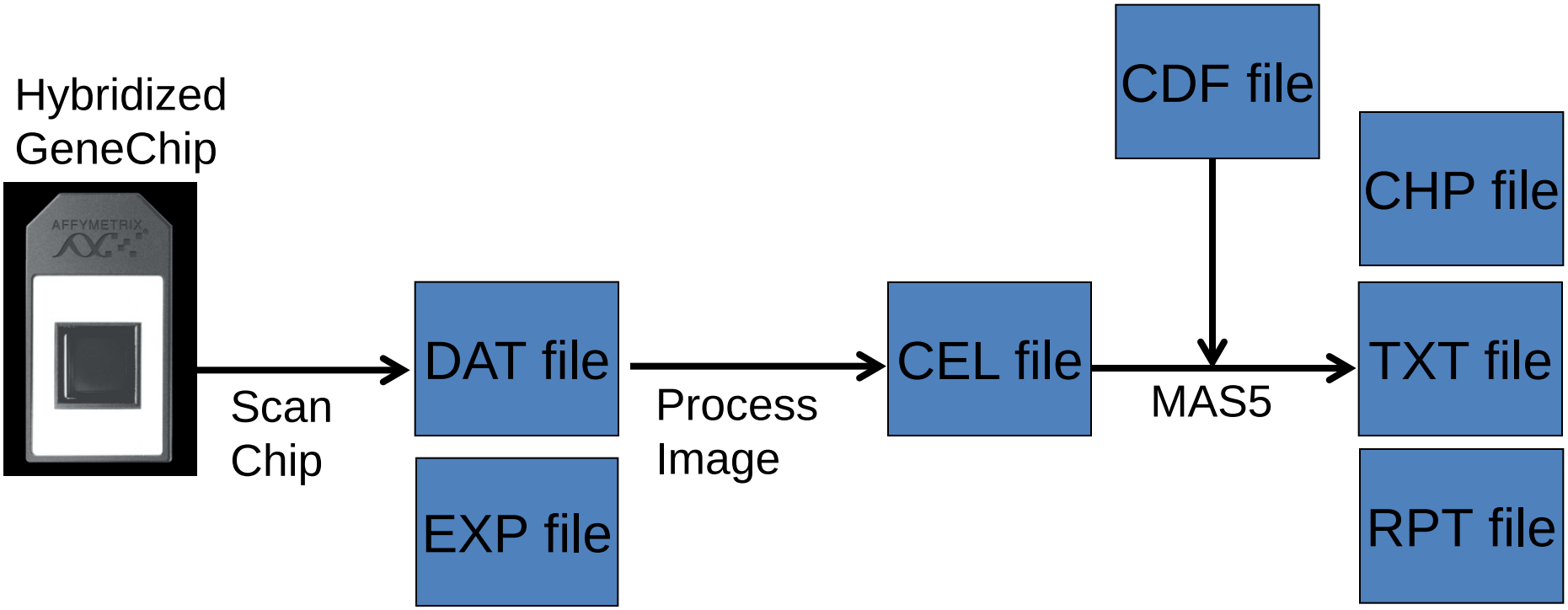


http://www.affymetrix.com/technology/ge_analysis/index.affx

Affymetrix File Types

- DAT file:
  - Raw (TIFF) optical image of the hybridized chip
- CDF File (Chip Description File):
  - Provided by Affy, describes layout of chip
- CEL File:
  - Processed DAT file (intensity/position values)
- CHP File:
  - Experiment results created from CEL and CDF files
- TXT File:
  - Probeset expression values with annotation (CHP file in text format)
- EXP File
  - Small text file of Experiment details (time, name, etc)
- RPT File
  - Generated by Affy software, report of QC info

Affymetrix Data Flow



Terminology

- A chip consists of a number of **probesets.**
- Probesets are intended to measure expression for a specific mRNA
- Each probeset is complementary to a **target sequence** which is derived from one or more mRNA sequences
- Probesets consist of 25mer **probe pairs** selected from the target sequence: one **Perfect Match (PM)** and one **Mismatch (MM)** for each chosen target position.
- Each chip has a corresponding **Chip Description File (CDF)** which (among other things) describes probe locations and probeset groupings on the chip.

Choosing probes

- How are taget sequences and probes chosen?
  - Target sequences are selected from the 3’ end of the transcript
  - Probes should be unique in genome (unless probesets are *intended* to cross hybridize)
  - Probes should not hybridize to other sequences in fragmented cDNA
  - Thermodynamic properties of probes
  - See Affymetrix docs for more details

http://www.affymetrix.com/support/technical/technotes/hgu133_p2_technote.pdf

Affymetrix Probeset Names

- Probeset identifiers beginning with AFFX are affy internal, not generally used for analysis
- Suffixes are meaningful, for example:
  - _at : hybridizes to unique antisense transcript for this chip
  - _s_at: all probes cross hybridize to a specified set of sequences
  - _a_at: all probes cross hybridize to a specified gene family
  - _x_at: at least some probes cross hybridize with other target sequences for this chip
  - _r_at: rules dropped *(my favorite!)*
  - and many more…
- See the Affymetrix document “Data Analysis Fundamentals” for details

Target Sequences and Probes

Example:

- 1415771_at:
  - **Description:** Mus musculus nucleolin mRNA, complete cds
  - **LocusLink:** AF318184.1 (NT sequence is 2412 bp long)
  - Target Sequence is 129 bp long

11 probe pairs tiling the target sequence

gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaa**gtcaaccatccaaaactctgtttgt**caaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaa**aactctgtttgtcaaaggtctgtct**gaggataccactgaagagaccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtca**aaggtctgtctgaggataccactga**agagaccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggt**ctgtctgaggataccactgaagaga**ccttaaaagaatcatttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggat**accactgaagagaccttaaaagaat**catttgagggctctgttcgtgcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaa**agaatcatttgagggctctgttcgt**gcaagaatagtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcat**ttgagggctctgttcgtgcaagaat**agtcactgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgag**ggctctgttcgtgcaagaatagtca**ctgatcgggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgagggctct**gttcgtgcaagaatagtcactgatc**gggaaactggttctt
gagaagtcaaccatccaaaactctgtttgtcaaaggtctgtctgaggataccactgaagagaccttaaaagaatcatttgagggctctgttcgtgcaagaata**gtcactgatcgggaaactggttctt**

Perfect Match and Mismatch

Target

tttccagacagactcctatggtgacttctctggaat

Perfect match

ctgtctgaggat**a**ccactgaagaga

ctgtctgaggat**t**ccactgaagaga

Mismatch

Probe pair

Affymetrix Chip Pseudo-image

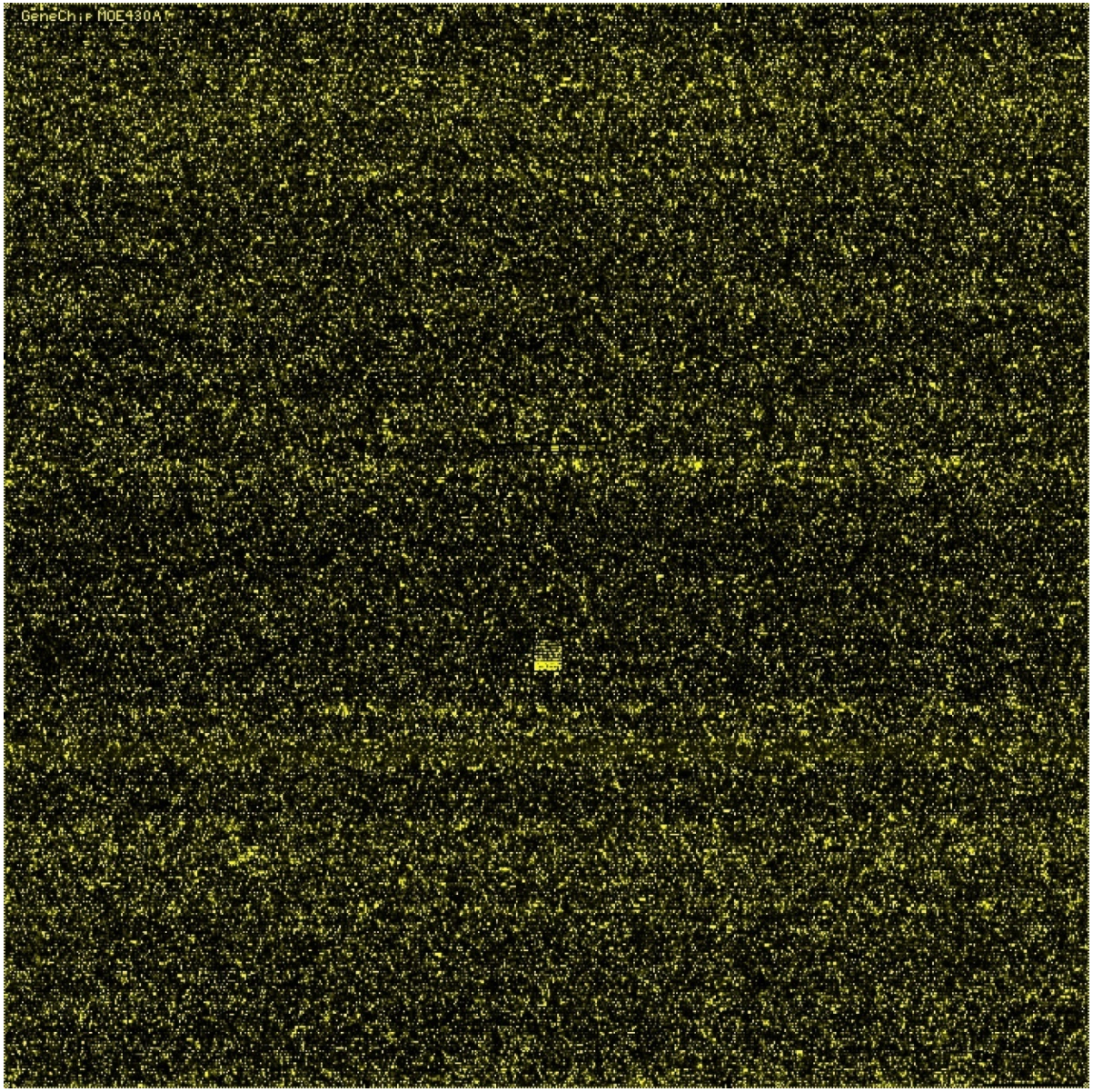


*image created using dChip software

1415771 at on MOE430A

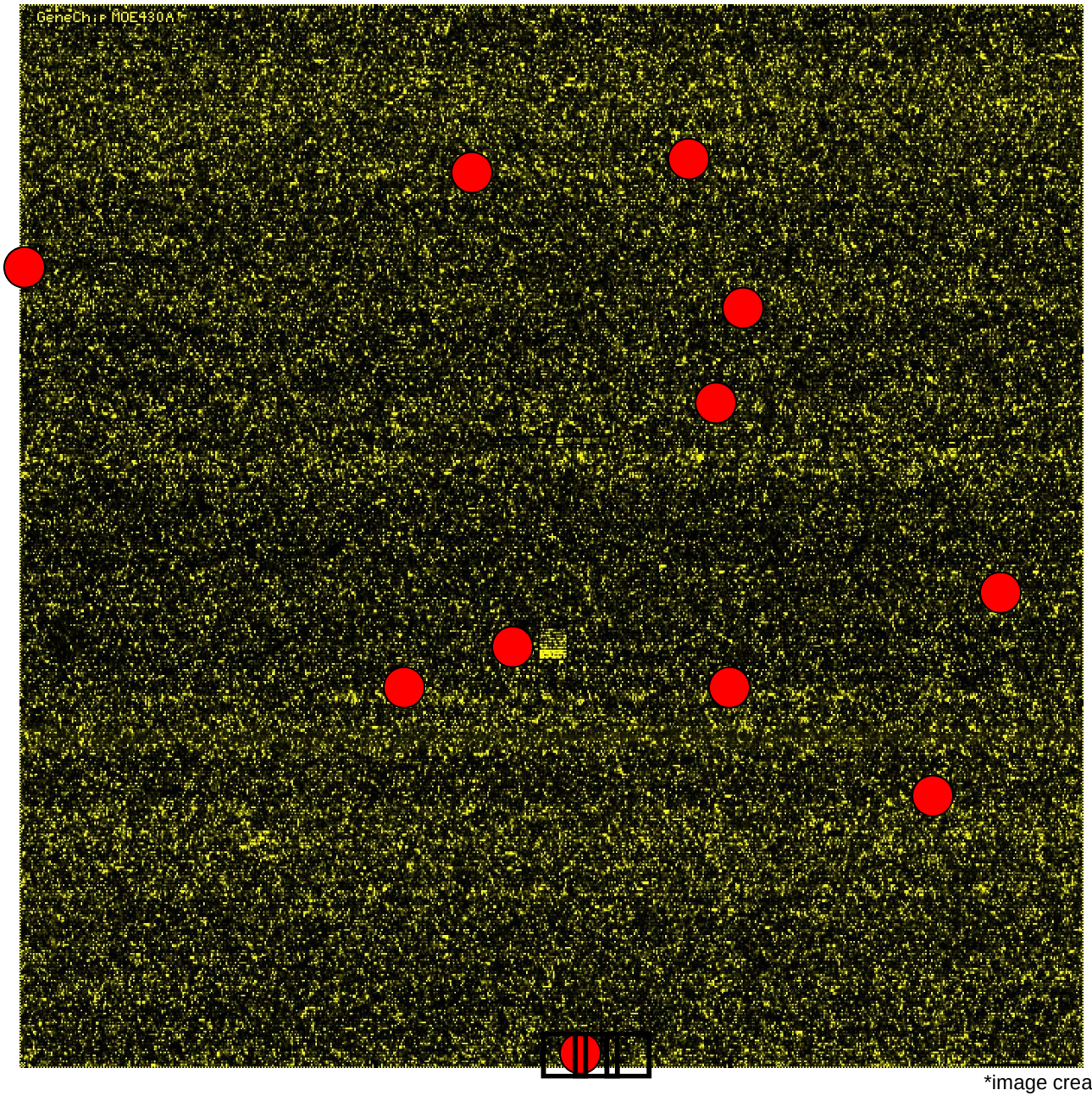
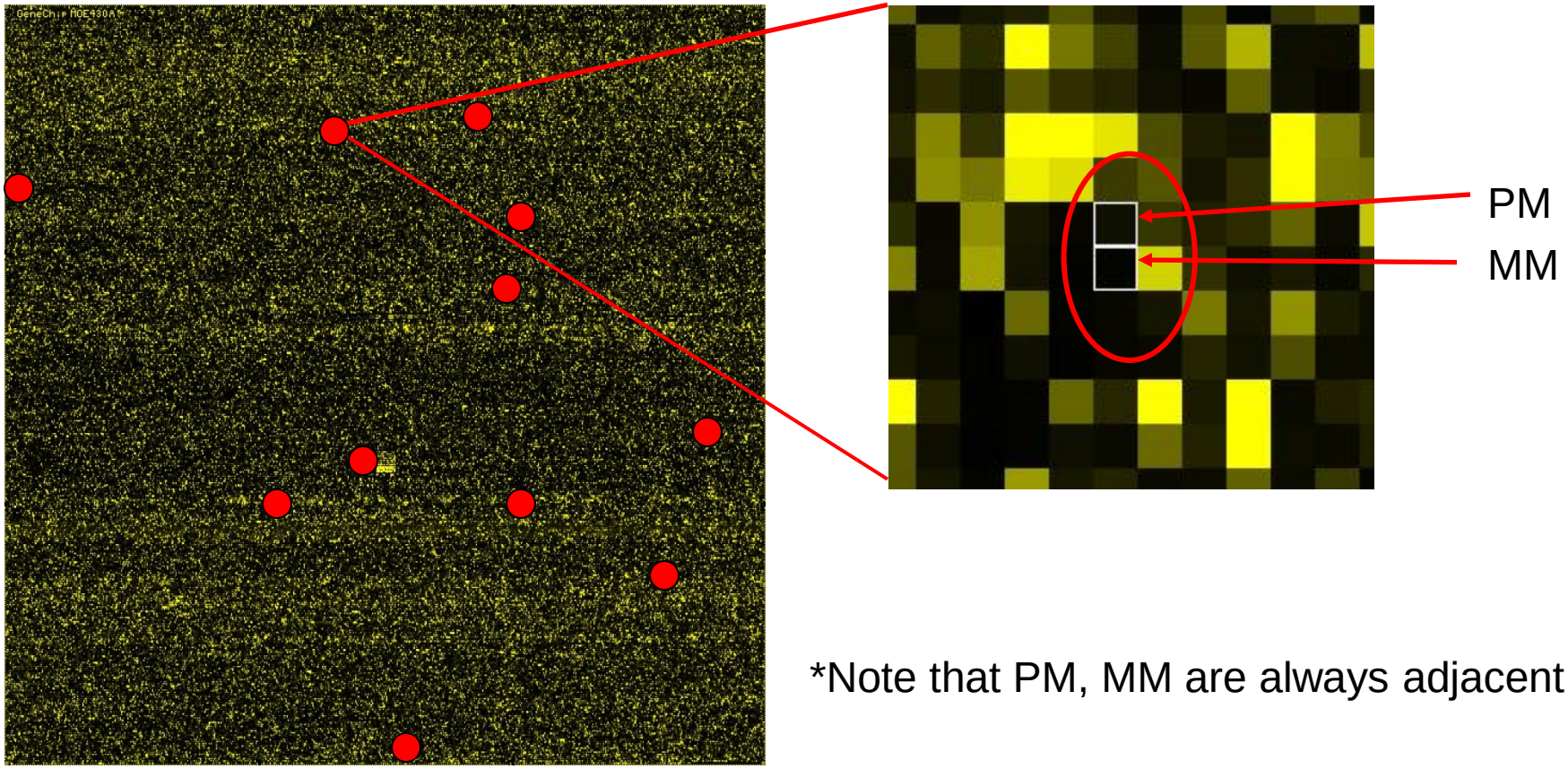


*image created using dChip software

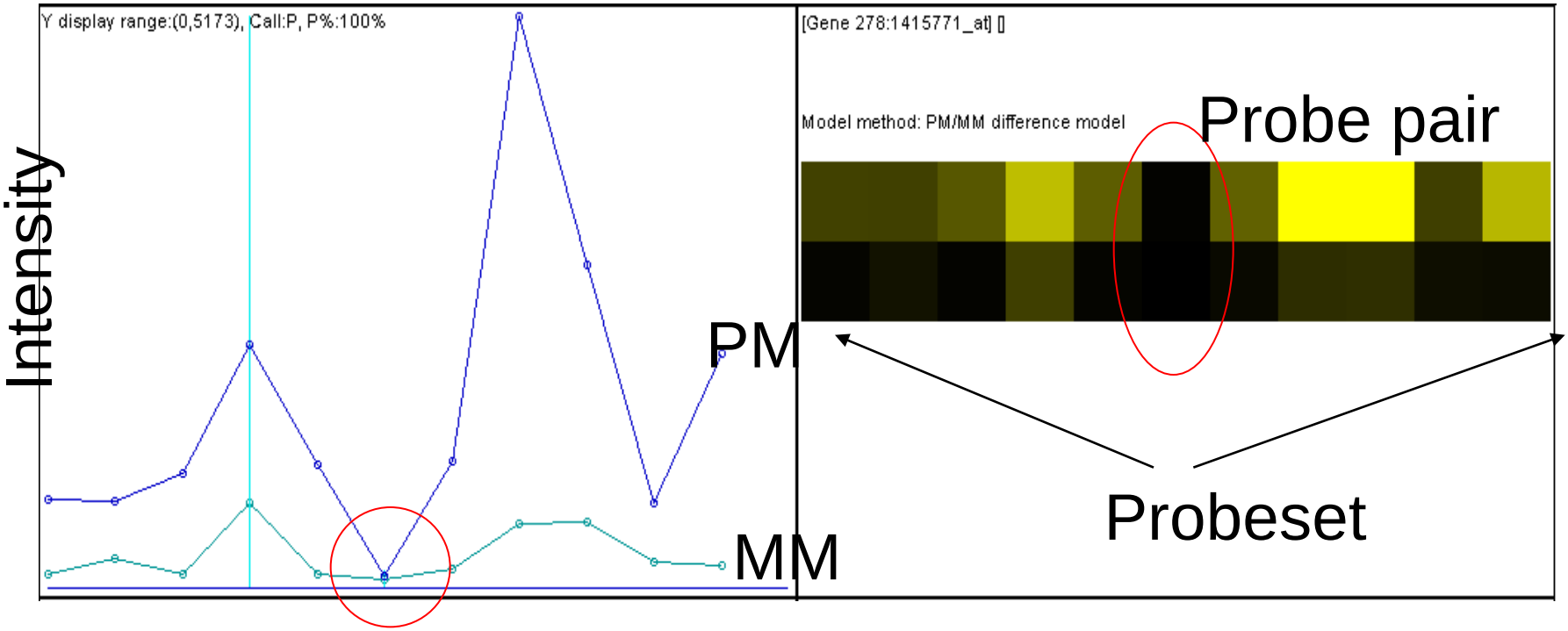
1415771_at on MOE430A



*Note that PM, MM are always adjacent

*image created using dChip software

1415771_at on MOE430A



Probe pair

*images created using dChip software

Intensity to Expression

- Now we have thousands of intensity values associated with probes, grouped into probesets.
- How do you transform intensity to expression values?
  - Algorithms
    - MAS5
      - Affymetrix proprietary method
    - RMA/GCRMA
      - Irizarry, Bolstad
    - ..many others
- Often called “normalization”

Common elements of different techniques

- All techniques do the following:
  - Background adjustment
  - Scaling
  - Aggregation
- The goal is to remove non-biological elements of the signal

MAS5

- Standard Affymetrix analysis, best documented in: http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf
- MAS5 results can't be *exactly* reproduced based on this document, though the `affy` package in Bioconductor comes close.
- MAS5 C++ source code released by Affy under GPL in 2005

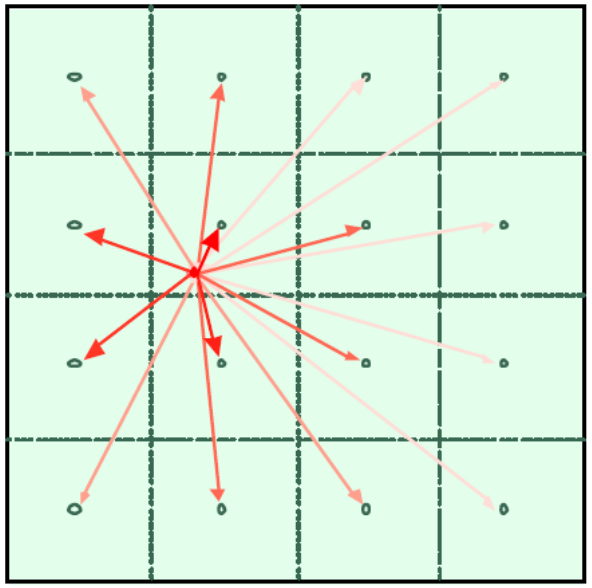
MAS5 Model

- Measured Value = N + P + S
  - N = Noise
  - P = Probe effects (non-specific hybridization)
  - S = Signal

MAS5: Background & Noise

Zone Values

- For purposes of calculating background values, the array is split up into K rectangular zones Z_k ($k = 1, \ldots, K$, default $K = 16$).
- Control cells and masked cells are not used in the calculation.
- The cells are ranked and the lowest 2% is chosen as the background b for that zone (bZ_k).
- The standard deviation of the lowest 2% cell intensities is calculated as an estimate of the background variability n for each zone (nZ_k).



$$b(x,y) = \frac{1}{\sum_{k=1}^{K} w_k(x,y)} \sum_{k=1}^{K} w_k(x,y)\, bZ_k$$

$$w_k(x,y) = \frac{1}{d_k^2(x,y) + smooth}$$

- From http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf

MAS5: Adjusted Intensity

$$A(x, y) = \max(I'(x, y) - b(x, y), NoiseFrac*n(x, y))$$
$$\text{where } I'(x, y) = \max(I'(x, y), 0.5)$$

A = Intensity minus background, the final value should be > noise.

A: adjusted intensity

I: measured intensity

b: background

NoiseFrac: default 0.5 (another fudge factor)

And the value should always be >=0.5 (log issues)

(fudge factor)

- From http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf

MAS5: Ideal Mismatch

Because Sometimes MM > PM

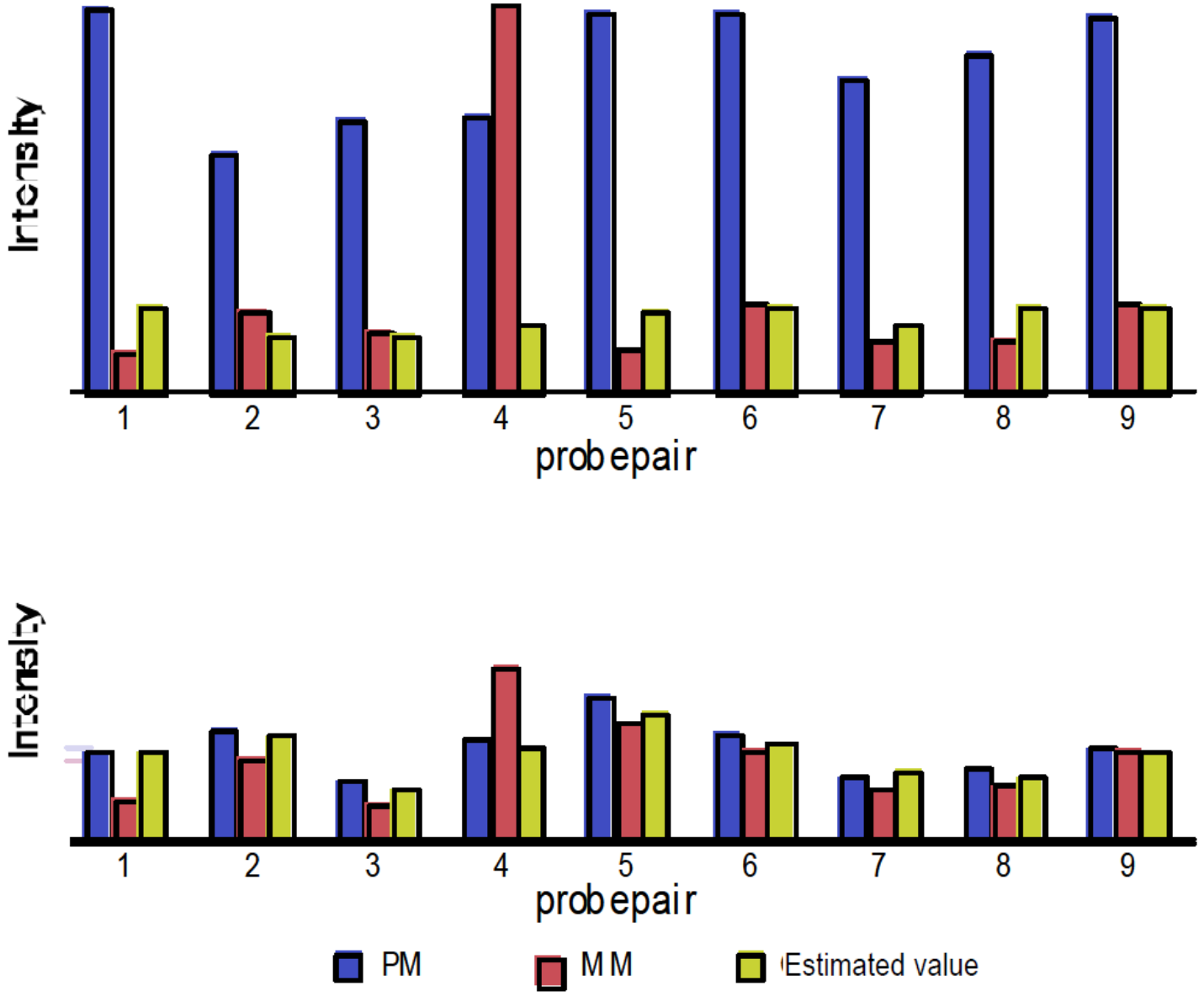
$$SB_i = T_{bi}\left(\log_2(PM_{i,j}) - \log_2(MM_{i,j}) : j = 1,\ldots,n_i\right)$$

$$IM_{i,j} = \begin{cases} MM_{i,j}, & MM_{i,j} < PM_{i,j} \\ \dfrac{PM_{i,j}}{2^{(SB_i)}}, & MM_{i,j} \geq PM_{i,j} \text{ and } SB_i > contrast\tau \\ \dfrac{PM_{i,j}}{2^{\left(\frac{contrast\,\tau}{1+\left(\frac{contrast\tau - SB_i}{scale\tau}\right)}\right)}}, & MM_{i,j} \geq PM_{i,j} \text{ and } SB_i \leq contrast\tau \end{cases}$$

default $contrast\tau$=0.03

default $scale\tau = 10$

- From http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf


Intensity
1
2
3
4
5
6
7
8
9
probepair
Intensity
1
2
3
4
5
6
7
8
9
probepair
PM
MM
Estimated value

MAS5: Signal

Value for each probe:

$$V_{i,j} = \max(PM_{i,j} - IM_{i,j}, d) \quad \text{default } \delta = 2^{(-20)}$$

$$PV_{i,j} = \log_2(V_{i,j}),\ j = 1, \ldots,\ n_i$$

Modified mean of probe values:

$$SignalLogValue_i = T_{bi}(PV_{i,1}, \ldots, PV_{i,n_i})$$

Scaling Factor

(Sc default 500)

$$sf = \frac{Sc}{TrimMean\left(2^{SignalLogValue_i}, 0.02, 0.98\right)}$$

Signal

(nf=1)

$$ReportedValue(i) = nf * sf * 2^{(SignalLogValue_i)}$$

T_{bi} = Tukey Biweight (mean estimate, resistant to outliers)

TrimMean = Mean less top and bottom 2%

- From http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf

MAS5: p-value and calls

- First calculate discriminant for each probe pair:

  R=(PM-MM)/(PM+MM)

- Wilcoxon one sided ranked test used to compare R vs tau value and determine p-value
- Present/Marginal/Absent calls are thresholded from p-value above and
  - **Present** =< alpha1
  - alpha1 < **Marginal** < alpha2
  - Alpha2 <= **Absent**
- Default: alpha1=0.04, alpha2=0.06, tau=0.015

MAS5: Summary

- Good
  - Usable with single chips (though replicated preferable)
  - Gives a p-value for expression data
- Bad:
  - Lots of fudge factors in the algorithm
  - Not *exactly* reproducible based upon documentation (source now available)
- Misc
  - Most commonly used processing method for Affy chips
  - Highly dependent on Mismatch probes

RMA

- Robust Multichip Analysis
- Used with groups of chips (>3), more chips are better
- Assumes all chips have same background, distribution of values: do they?
- Does not use the MM probes as (PM-MM*) leads to high variance
  - *This means that **half** the probes on the chip are excluded, yet it still gives good results!*
- Ignoring MM decreases accuracy, increases precision (reproducibility).

RMA Model

$$y_{ij} = m + \alpha_i + \beta_j + \varepsilon_{ij}$$

where $y_{ij} = \log_2 \mathrm{N}\left(\mathrm{B}\left(PM_{ij}\right)\right)$

α_i is a probe-effect i= 1,…,I

β_j is chip-effect ($m + \beta_j$ is log2 gene expression on array *j)* j=1,…,J

From a presentation by Ben Bolstad

http://bioinformatics.ca/workshop_pages/genomics/lectures2004/16

RMA Background

$$\mathrm{E}\left(S \mid O = o\right) = a + b\,\frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{o-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) - \Phi\left(\frac{o-a}{b}\right) - 1}$$

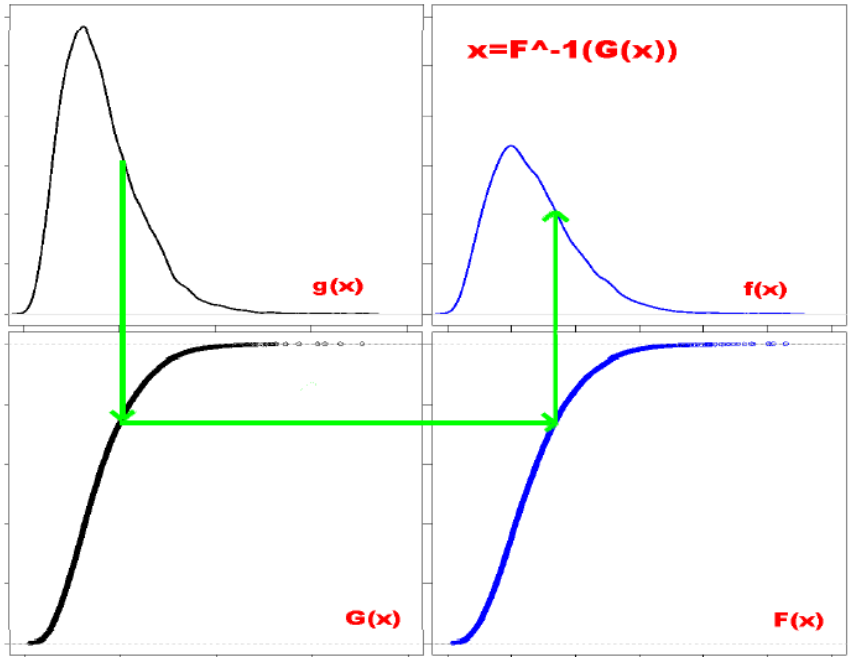
$$a = o - \mu - \sigma^2\alpha\,,\; b = \sigma$$

This provides background correction

From a presentation by Ben Bolstad
http://bioinformatics.ca/workshop_pages/genomics/lectures2004/16

RMA: Quantile Normalization & Scaling

- Fit all the chips to the same distribution
- Scale the chips so that they have the same mean.



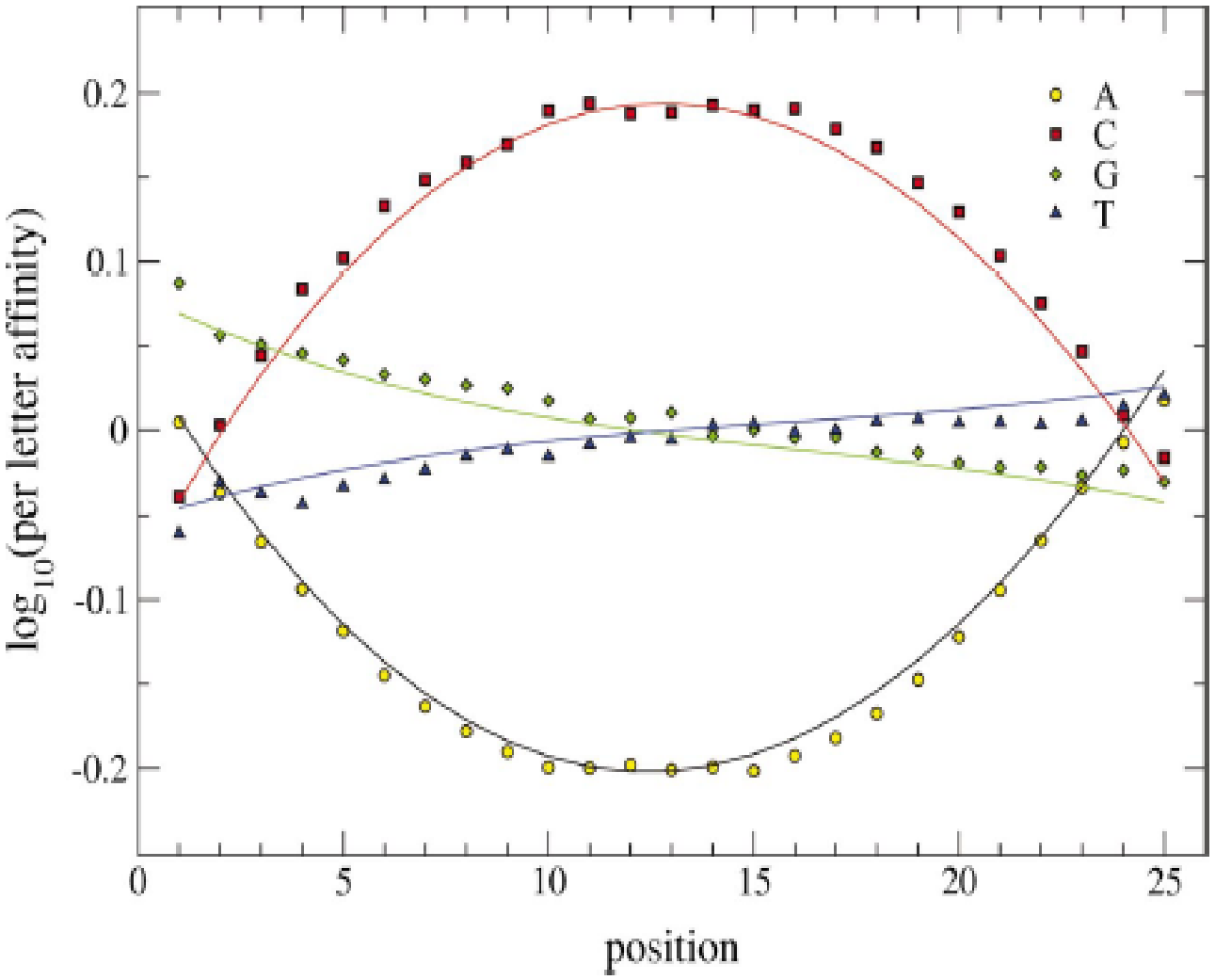
From a presentation by Ben Bolstad
http://bioinformatics.ca/workshop_pages/genomics/lectures2004/16

RMA: Estimate Expression

- Assumption that these log transformed, background corrected expression values follow a linear model,
- Linear Model is estimated by using a “median polish” algorithm
- Generates a model based on chip, probe and a constant

GCRMA: Background Adjustment

Sequence specificity of brightness in the PM probes.



PHYSICAL REVIEW E **68**, 011906 ~2003!

(GC)RMA: Summary

- Good:
  - Results are $\log_2$
  - GCRMA: Adjusts for probe sequence effects
  - Rigidly model based: defines model then tries to fit experimental data to the model. Fewer fudge factors than MAS5
- Bad
  - Does not provide “calls” as MAS5 does
- Misc
  - The input is a group of samples that have same distribution of intensities.
  - Requires multiple samples

Exon Array

3' Arrays

1 gene --- 1 or 2 probesets

Probes from 600 bps near 3' end

Probeset has 11 PM, 11 MM probes

54,000 probesets

Average16 probes per RefSeq gene

Exon Arrays

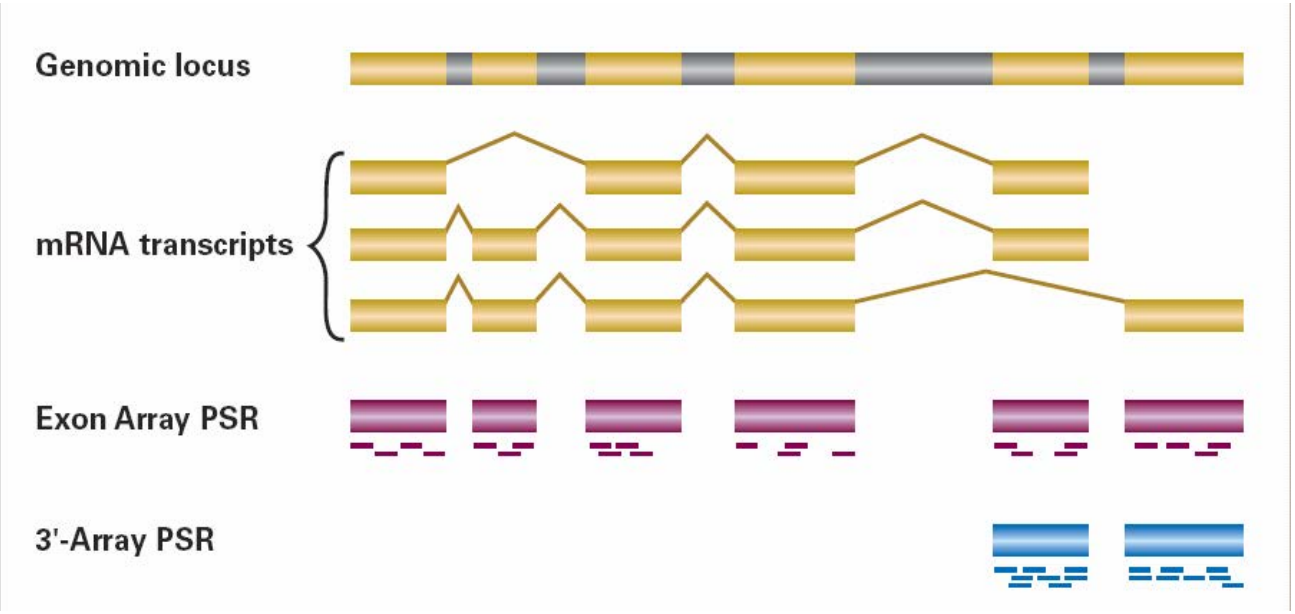
1 gene --- many probesets

Probes from each putative exon

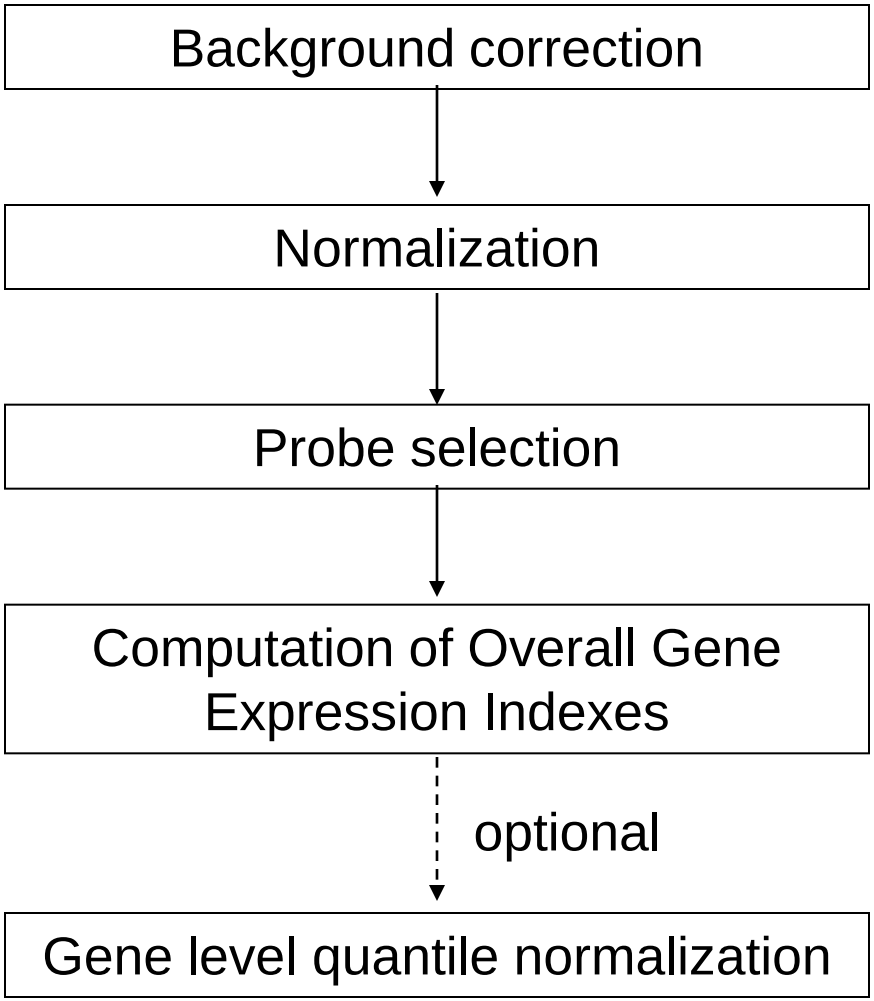
Probeset has 4 PM probes

1.4 Million probesets, 6 M features

Average147 probes per RefSeq gene

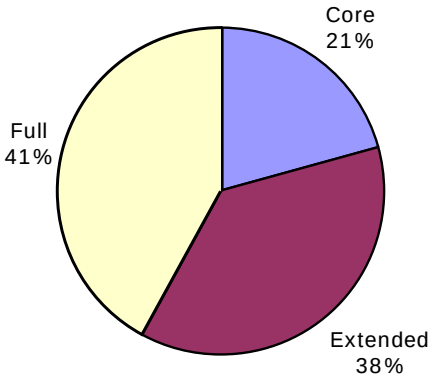


Workflow


Background correction
Normalization
Probe selection
Computation of Overall Gene Expression Indexes
optional
Gene level quantile normalization

Exon Array Probesets Classified by Annotational Confidence

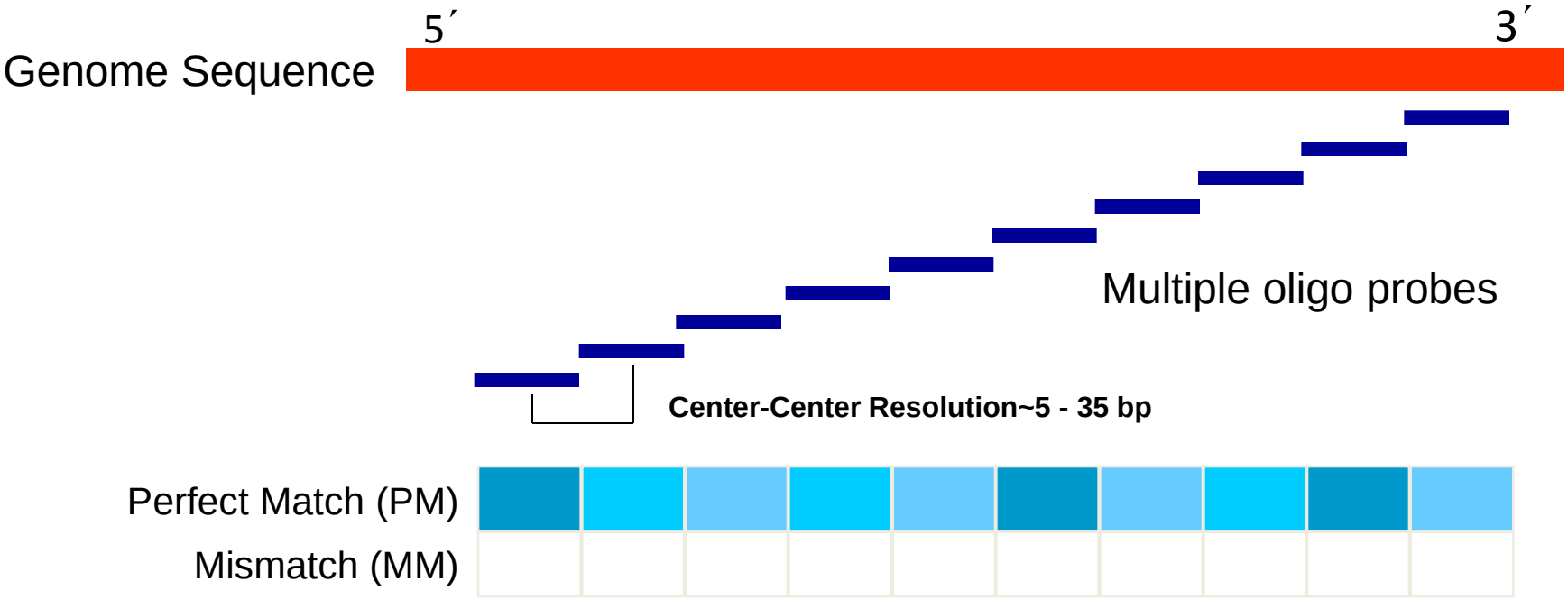
- ***Core probesets*** target exons supported by RefSeq mRNAs.
- ***Extended probesets*** target exons supported by ESTs or partial mRNAs.
- ***Full probesets*** target exons supported purely by computational predictions.



Probe Selection for Gene-Level Expression

- Most full and extended probes are not suitable for estimating gene-level expression
  - Probes may target false exon predictions
- Even some core probes may not be suitable
  - Bad probes with low affinity, or cross-hybridize
  - Probes targeting differentially spliced exons
- Probe selection
  - Selecting a suitably large subset of good probes targeting constitutively spliced regions of the gene
  - Use only to selected probes to estimate gene expression

Tiling Array



- Resolution (spacing between probes):

  35 bp spacing = 10 bp GAP between adjacent probes (whole human genome array-14)

  20 bp spacing = 5 bp OVERLAP between adjacent probes (ENCODE array-1)

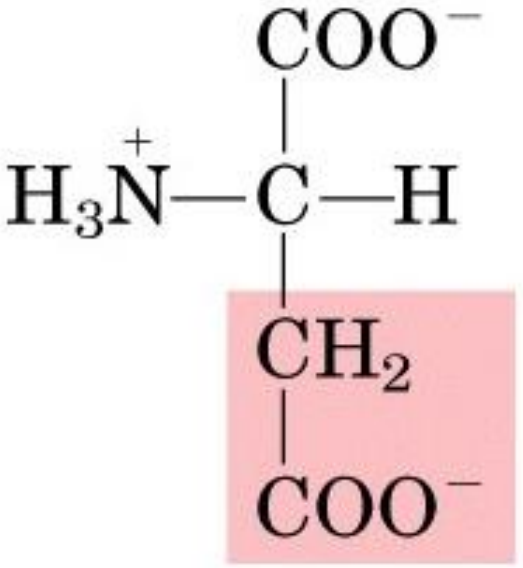
  5 bp spacing = 20 bp OVERLAP between adjacent probes (30% of human genome array-98)

- Probes can be made as sense or anti-sense;
- Labeling assay can be ‘strand-specific’ or not.

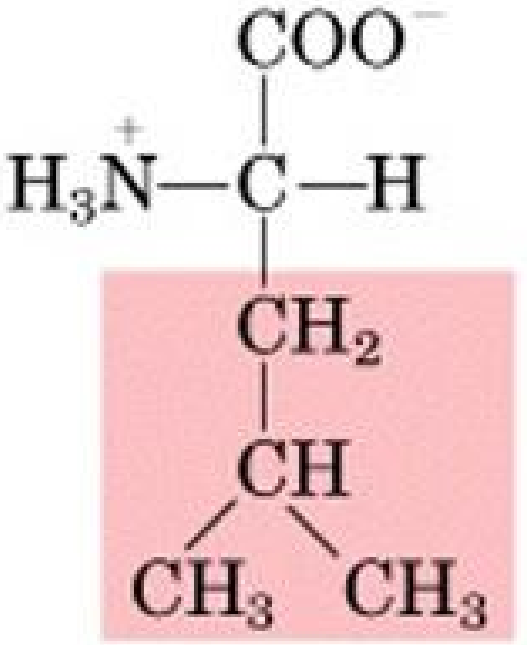
Applications of Tiling Array

- Small (including mi-RNA) and long RNA profiling/transcript discovery
- ChIP-chip:
  - Transcription Factors
  - Histone Modifications
  - DNA Methylation

Masses of Amino Acid Residues

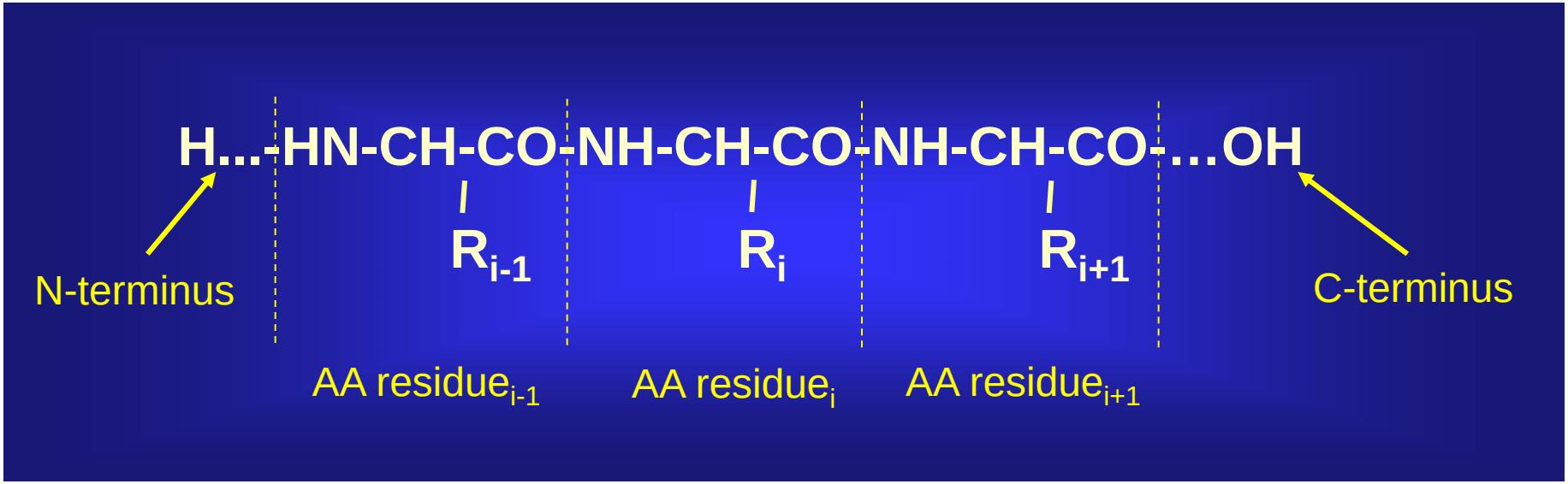


Aspartate

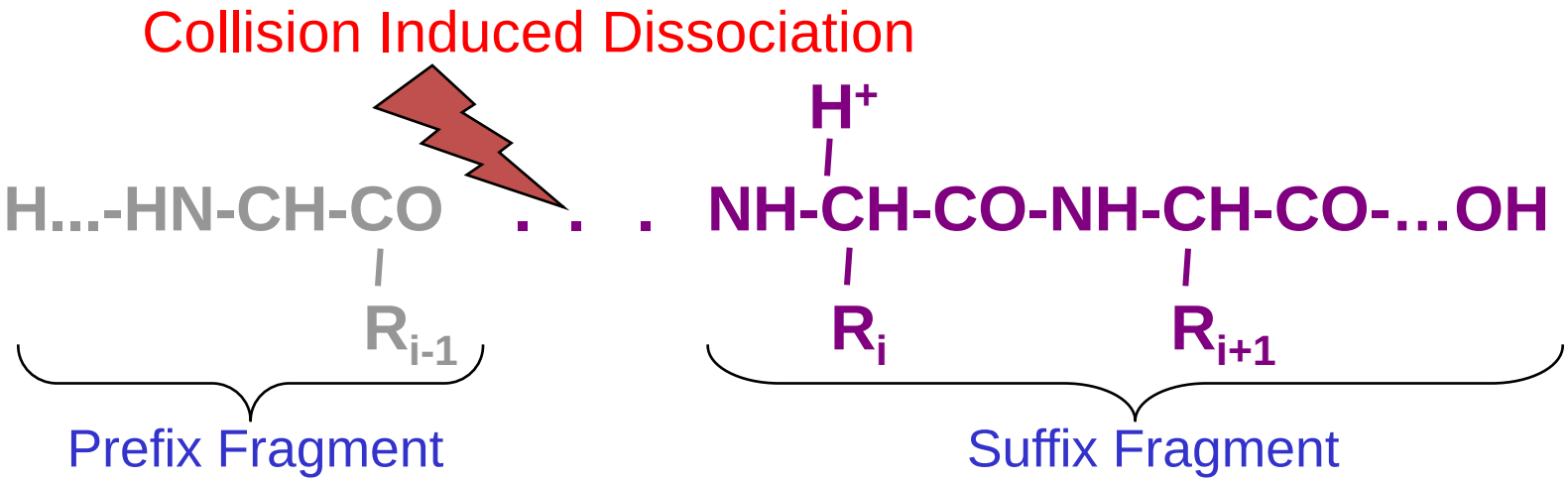


Leucine

Protein Backbone



Peptide Fragmentation



- Peptides tend to fragment along the backbone.
- Fragments can also loose neutral chemical groups like NH_3 and H_2O.

Breaking Protein into Peptides and Peptides into Fragment Ions

- Proteases, e.g. trypsin, break protein into *peptides*.
- A Tandem Mass Spectrometer further breaks the peptides down into *fragment ions* and measures the mass of each piece.
- Mass Spectrometer accelerates the fragmented ions; heavier ions accelerate slower than lighter ones.
- Mass Spectrometer measure *mass/charge* ratio of an ion.

Mass Spectrometry

Matrix-Assisted Laser Desorption/Ionization (MALDI)

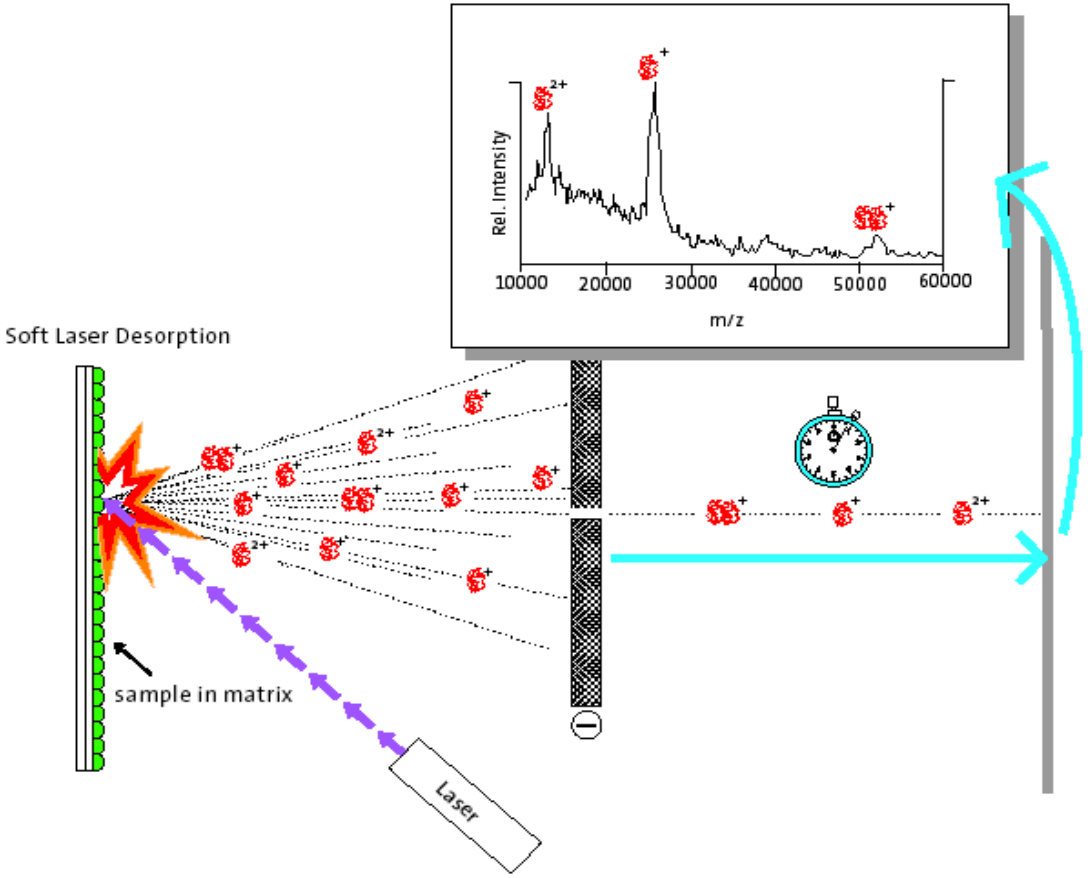
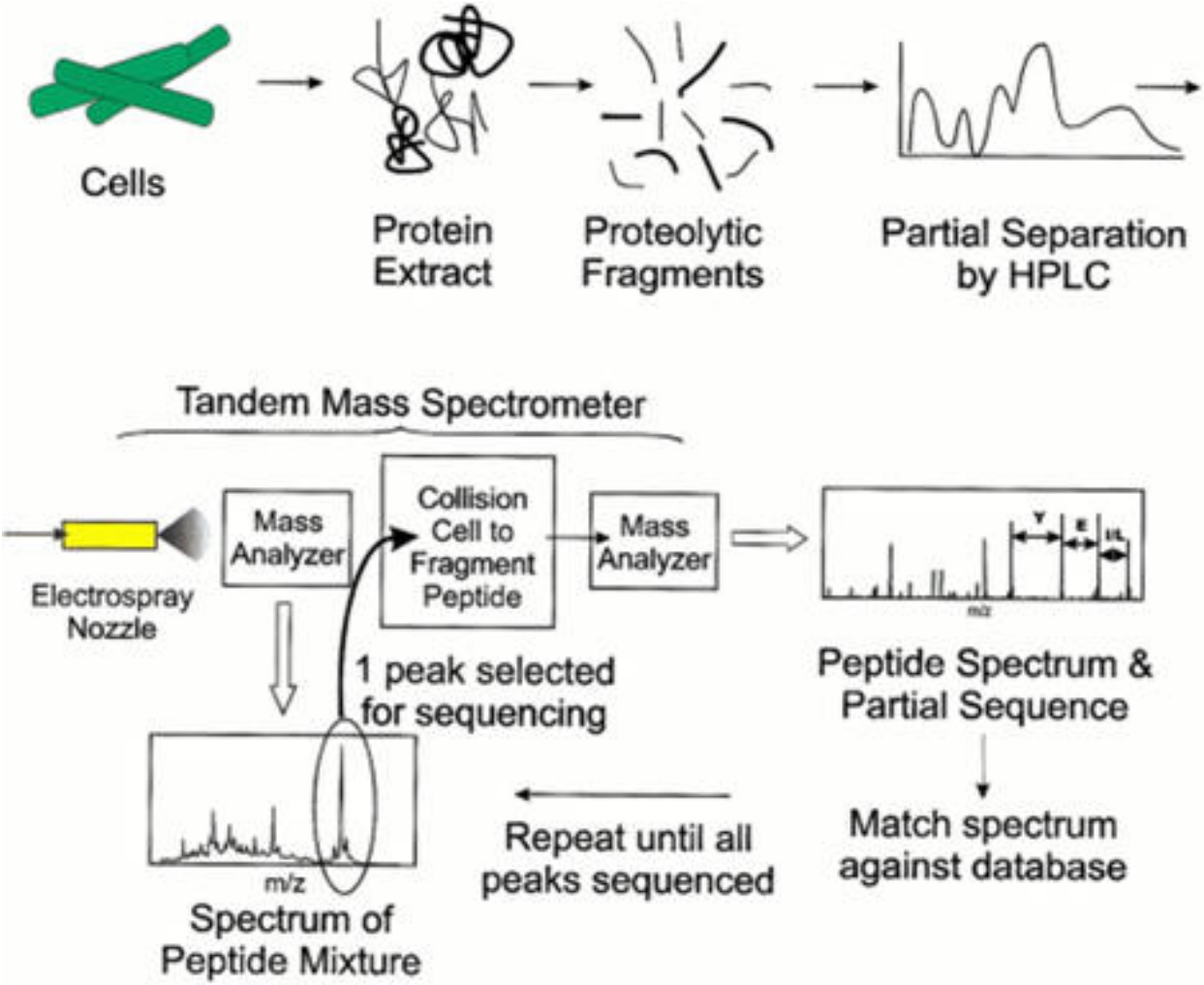


Figure 2. The soft laser desorption process.

From lectures by Vineet Bafna (UCSD)

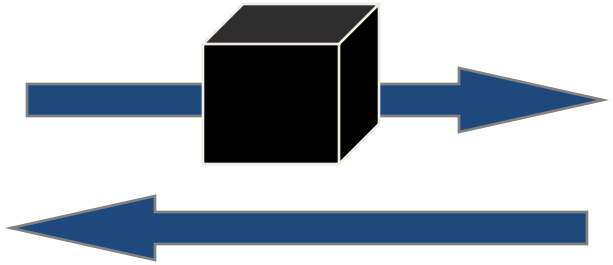
Tandem Mass-Spectrometry

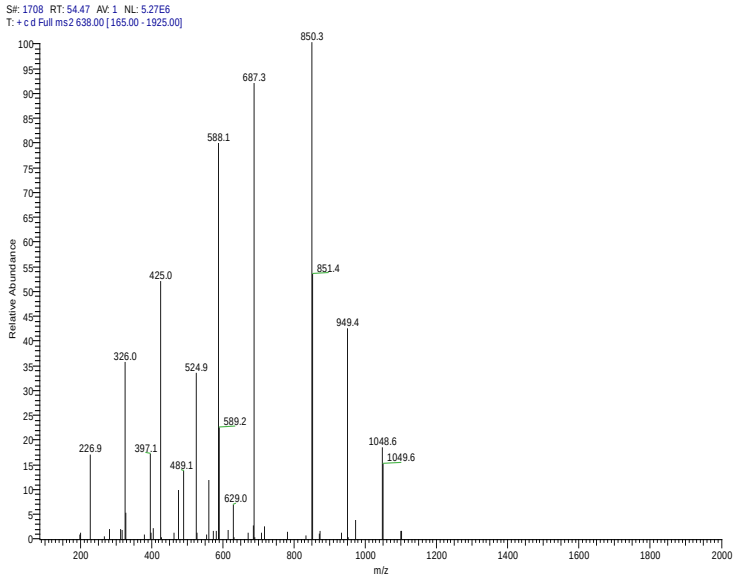


Protein Identification by Tandem Mass Spectrometry

Sequence

MS/MS instrument





Database search

- **Sequest**

de Novo interpretation

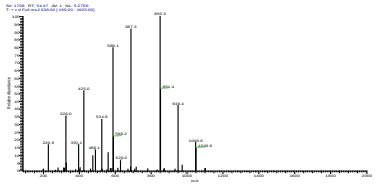
- **Sherenga**

Tandem Mass Spectrum

- Tandem Mass Spectrometry (MS/MS): mainly generates partial N- and C-terminal peptides
- Spectrum consists of different ion types because peptides can be broken in several places.
- Chemical noise often complicates the spectrum.
- Represented in 2-D: mass/charge axis vs. intensity axis

De Novo vs. Database Search

Database Search



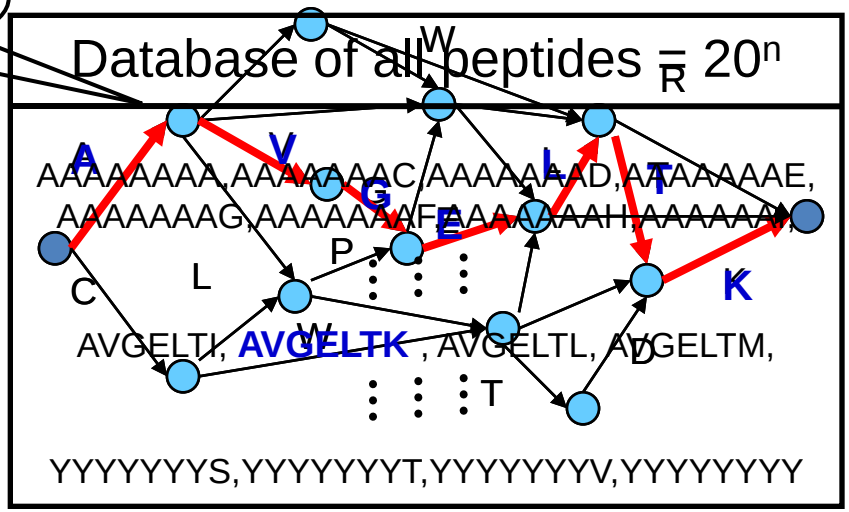
De Novo

Mass, Score

Database of known peptides

MDERHILNM, KLQWVCSDL, PTYWASDL, ENQIKRSACVM, TLACHGGEM, NGALPQWRT, HLLERTKMNVV, GGPASSDA, GGLITGMQSD, MQPLMNWE, AALKMNVRT, **AVGELTK**, HEWAILF, GHNLWAMNAC, GVFGSVLRA, EKLNKAATYIN..

Database of all peptides = 20^n

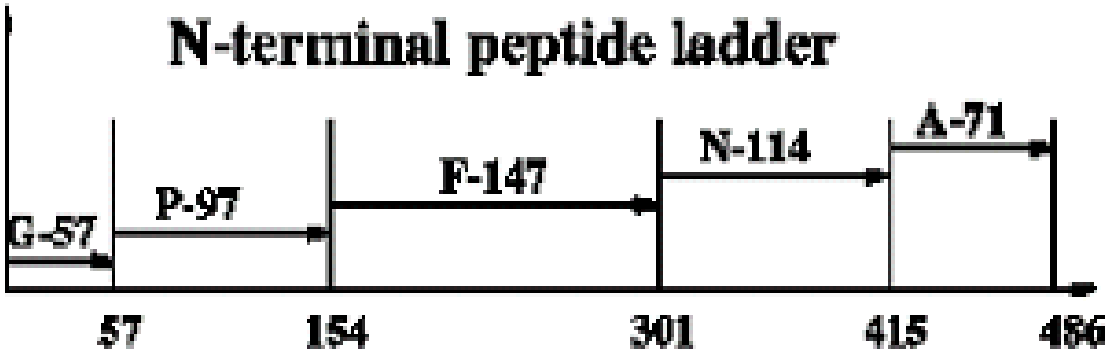


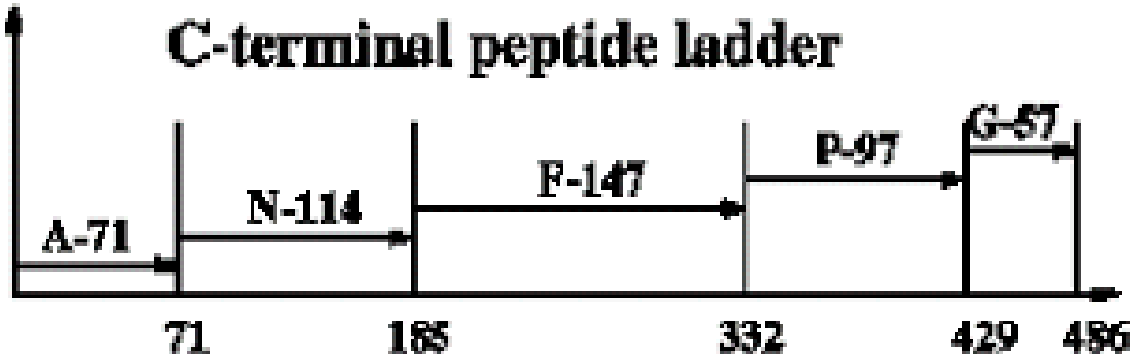
AVGELTK

De Novo vs. Database Search: A Paradox

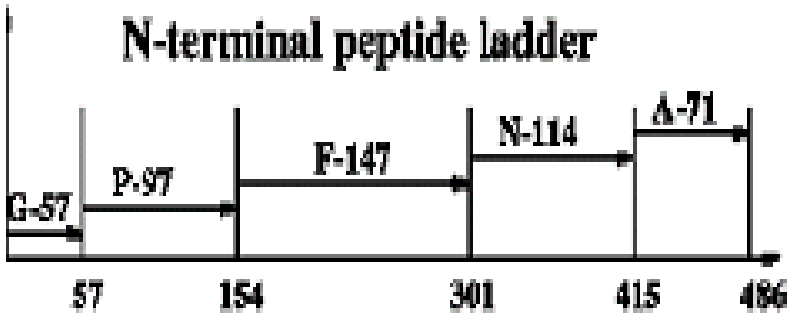
- The database of all peptides is huge ≈ $O(20^n)$.
- The database of all known peptides is much smaller ≈ $O(10^8)$.
- However, *de novo* algorithms can be much *faster*, even though their search space is much *larger!*
- A database search scans all peptides in the ***database of all known peptides*** search space to find best one.
- De novo eliminates the need to scan ***database of all peptides*** by modeling the problem as a graph search.

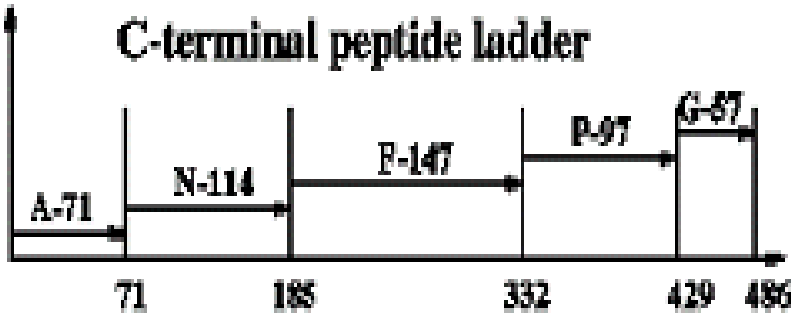
Theoretical Spectrum

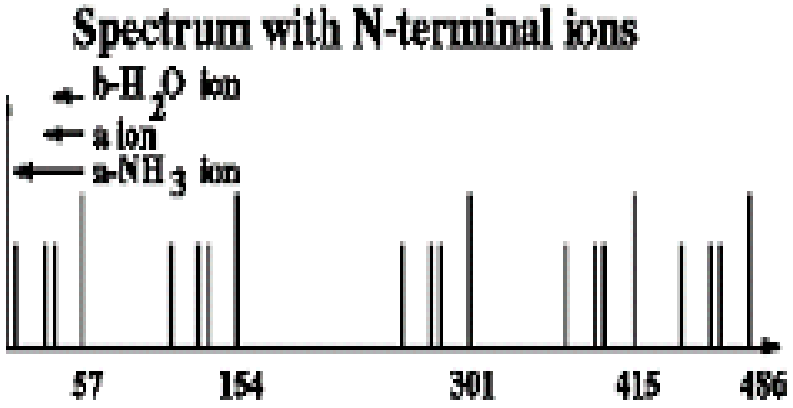


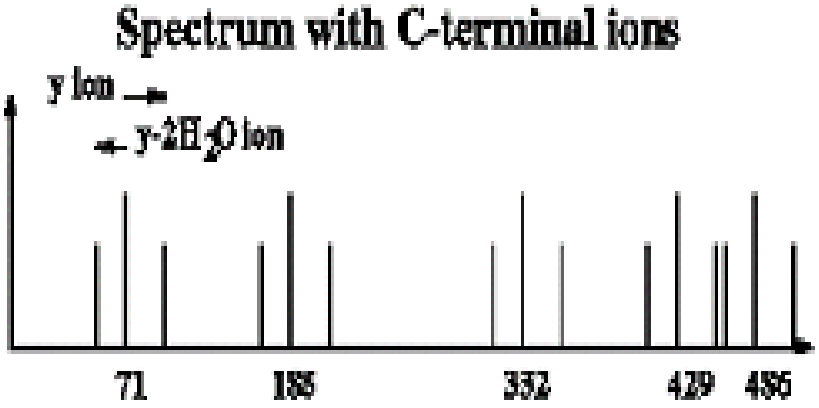


Theoretical Spectrum (cont'd)

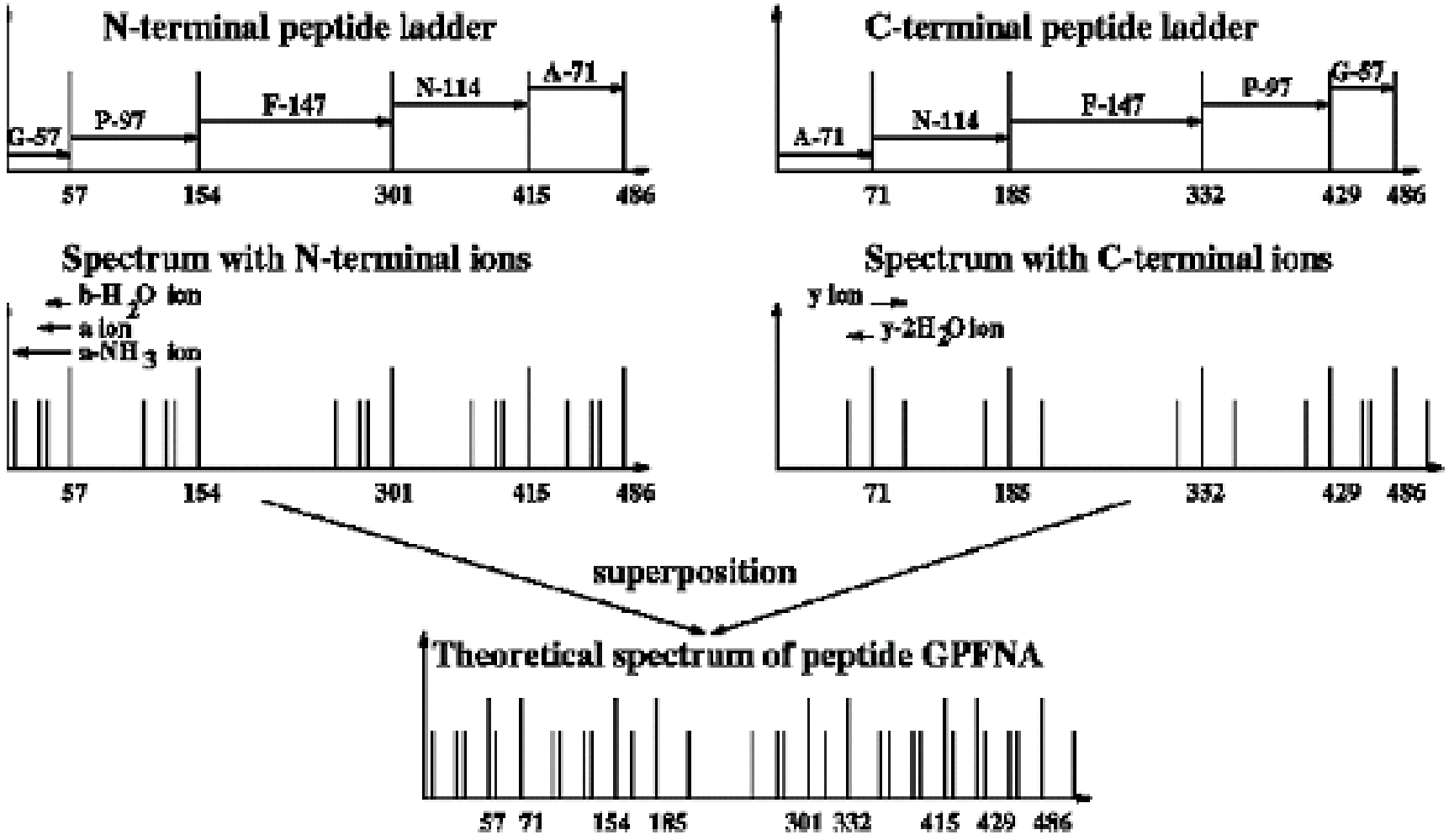




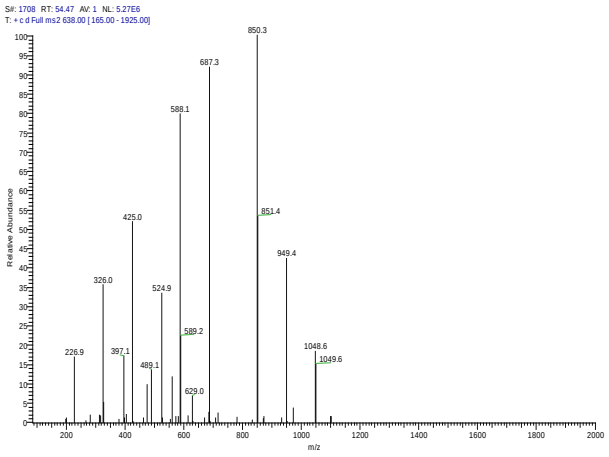


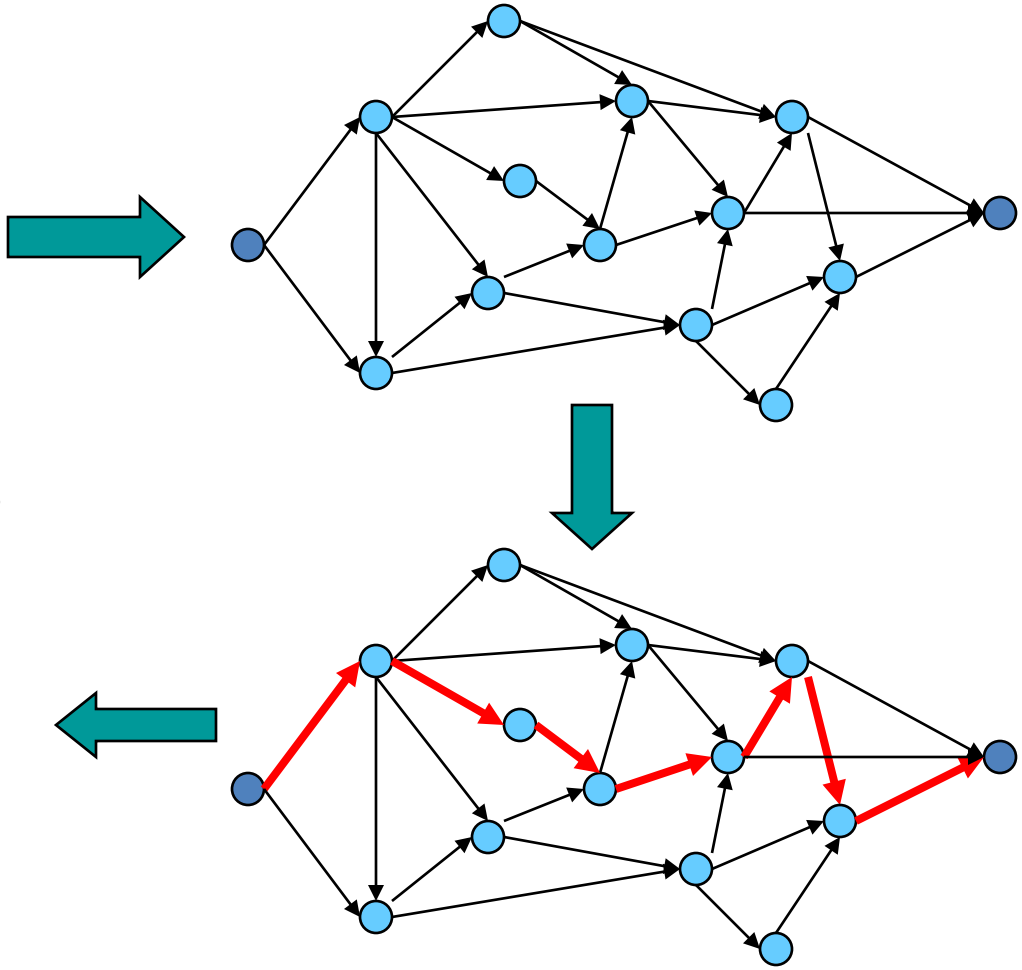


Theoretical Spectrum (cont'd)



De novo Peptide Sequencing





Sequence

Why Not Sequence De Novo?

- *De novo* sequencing is still not very accurate!

Algorithm	Amino Acid Accuracy	Whole Peptide Accuracy
Lutefisk (Taylor and Johnson, 1997).	0.566	0.189
SHERENGA (Dancik et. al., 1999).	0.690	0.289
Peaks (Ma et al., 2003).	0.673	0.246
PepNovo (Frank and Pevzner, 2005).	**0.727**	**0.296**

- Less than 30% of the peptides sequenced were completely correct!

Pros and Cons of *de novo* Sequencing

- Advantage:
  - Gets the sequences that are not necessarily in the database.
  - An additional similarity search step using these sequences may identify the related proteins in the database.
- Disadvantage:
  - Requires higher quality data.
  - Often contains errors.

MS/MS Database Search

Database search in mass-spectrometry has been very successful in identification of **already known** proteins.

Experimental spectrum can be compared with theoretical spectra of database peptides to find the best fit.

SEQUEST (Yates et al., 1995)

But reliable algorithms for identification of modified peptides is a much more difficult problem.

Limitations of Proteomics

- Experimental limitations:
  - Large-scale protein analysis difficult because:
  - Proteins are fragile
  - They can exist in multiple isoforms
  - There is no protein equivalent of PCR for amplification of a small sample

Limitations of Proteomics

- Data Analysis Limitations:
  - Data contains a lot of noise that is difficult to separate from actual signal. This results in wastage of computing resources on searching for unlikely spectra.
  - Database searches for matching spectra only give scores, leaving manual intervention necessary for eliminating false positives