Serum Glycan Biomarkers for Diagnosis and Prognosis

David M. Rocke
Division of Biostatistics
Department of Public Health Sciences
University of California, Davis

Biomarkers

- Tissue (gene/protein/miRNA)
  - Invasive
  - Only useful post-surgery/biopsy
  - Gold standard in some ways
- Imaging
  - Noninvasive
  - Limited in resolution
  - Limited molecular information

Minimally Invasive Biomarkers

- Blood, saliva, urine
- Gene Expression
  - May be expression of genes in cells far from the tumor/organ site/disease site
  - Excellent analytical methods now exist, though large number of probes and small samples sizes often mean that only large signals can be seen
- DNA, RNA, metabolites

Proteins

- Shotgun proteomics and 2D gels both have significant problems
- Targeted biomarker discovery using ELISA, protein affinity chips, Luminex very feasible
- In general, these methods cannot detect post-translational modifications because the adducts are stripped before analysis or because antibodies specific to the modified form do not exist

Metabolomics

- Mass spectrometry (e.g., LC/ToF-MS) can detect and measure relative amounts of small molecules much more easily than with proteins
- Lipids, saccharides, and others
- For example we can speciate 500 lipids including di- and tri-glycerides species
- For example, we can measure over a hundred compounds in the arachidonic acid pathway including prostaglandins, COX1, COX2, and LOX products, and CyP450 pathways eicosanoids.
- Can measure enzymes, substrate, and product

Glycosylation

- Many human proteins occur in glycosylated form (probably over 50%)
- Glycosylation affects the activity of proteins on cell surfaces and intra-and inter-cellularly.
- Some well known cancer biomarkers are glycosylated proteins
  - Mucins are upregulated in cancer, but also have altered glycosylation, especially in metastases.
  - MUC1 and MUC16 (CA-125)
  - PSA

Glycobiology

- Glycosylation is a process that can be affected by conditions with the cell
- The structure of the glycan depends on the sequence of action of a series of specific glycosyltransferases.
- The composition alone is often aberrant in cancer (high sialic acid, high mannose), as well as the structure.

Glycans

- Glycans are oligosaccharides adducted to proteins.
- Composed of monosaccharides such as glucose, hexose, mannose, sialic acid
- In general have a branching structure composed of monosaccharide units
- Focus on N-linked rather than O-linked as these are more abundant

Glycomics

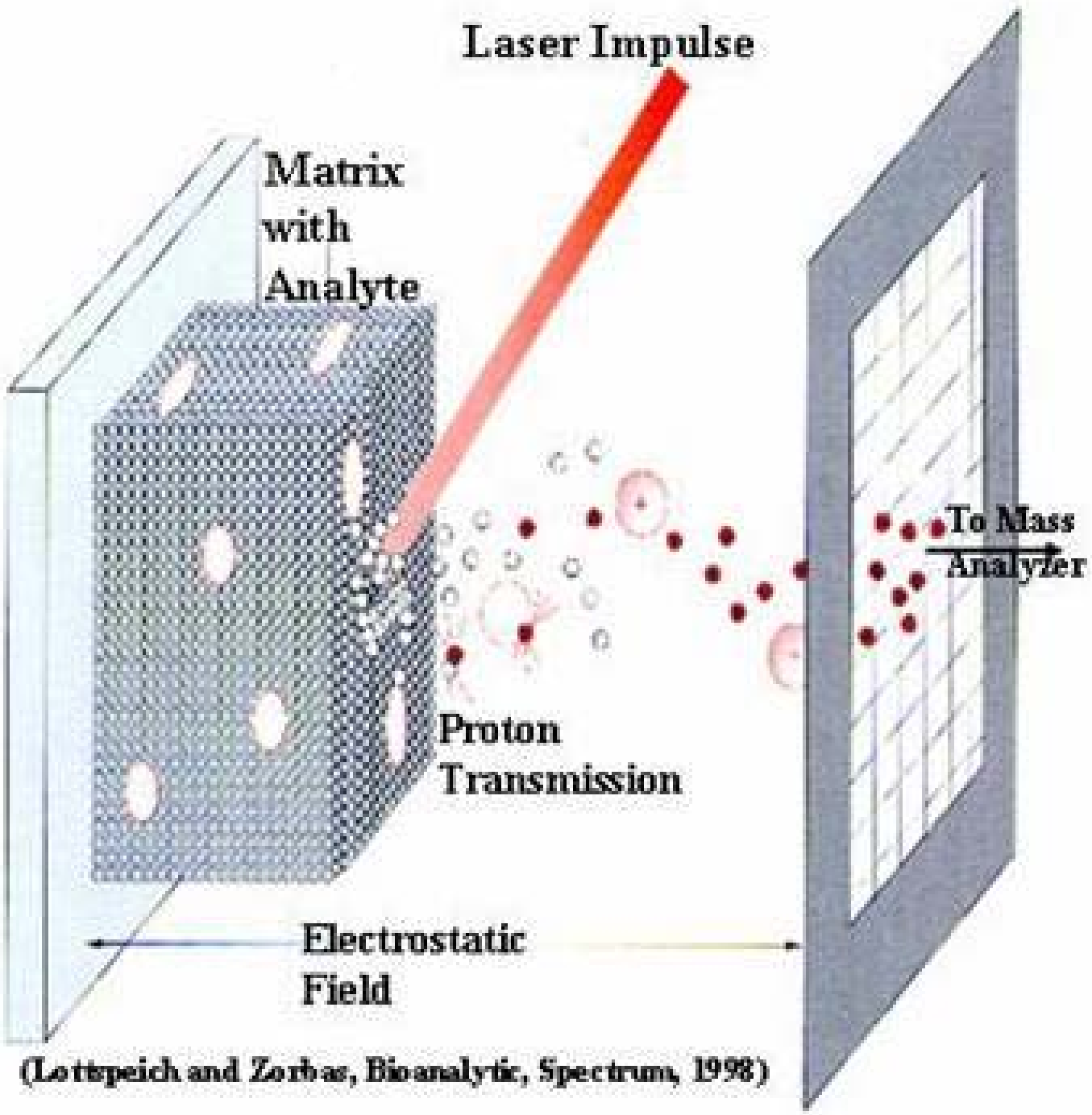
- Extract protein from serum, tissue or other biological samples.
- Isolate glycans removing most or all of the peptide component.
- Compare glycan profiles between phenotypic or experimental conditions.
- Identify differentially expressed glycans, ratios of glycans, or glycan signatures.
- Identify proteins containing these glycans to develop mechanistic understanding.

Instrumentation

- Substantial chemistry in isolation and separation
- MALDI ionization
- Glycans are relatively small, 500-2000 Daltons, much easier than proteins
- FT-ICR mass spec (Fourier Transform Ion Cyclotron Resonance
- Highly mass accurate if properly calibrated (1000 ± 0.001 Daltons)

MALDI

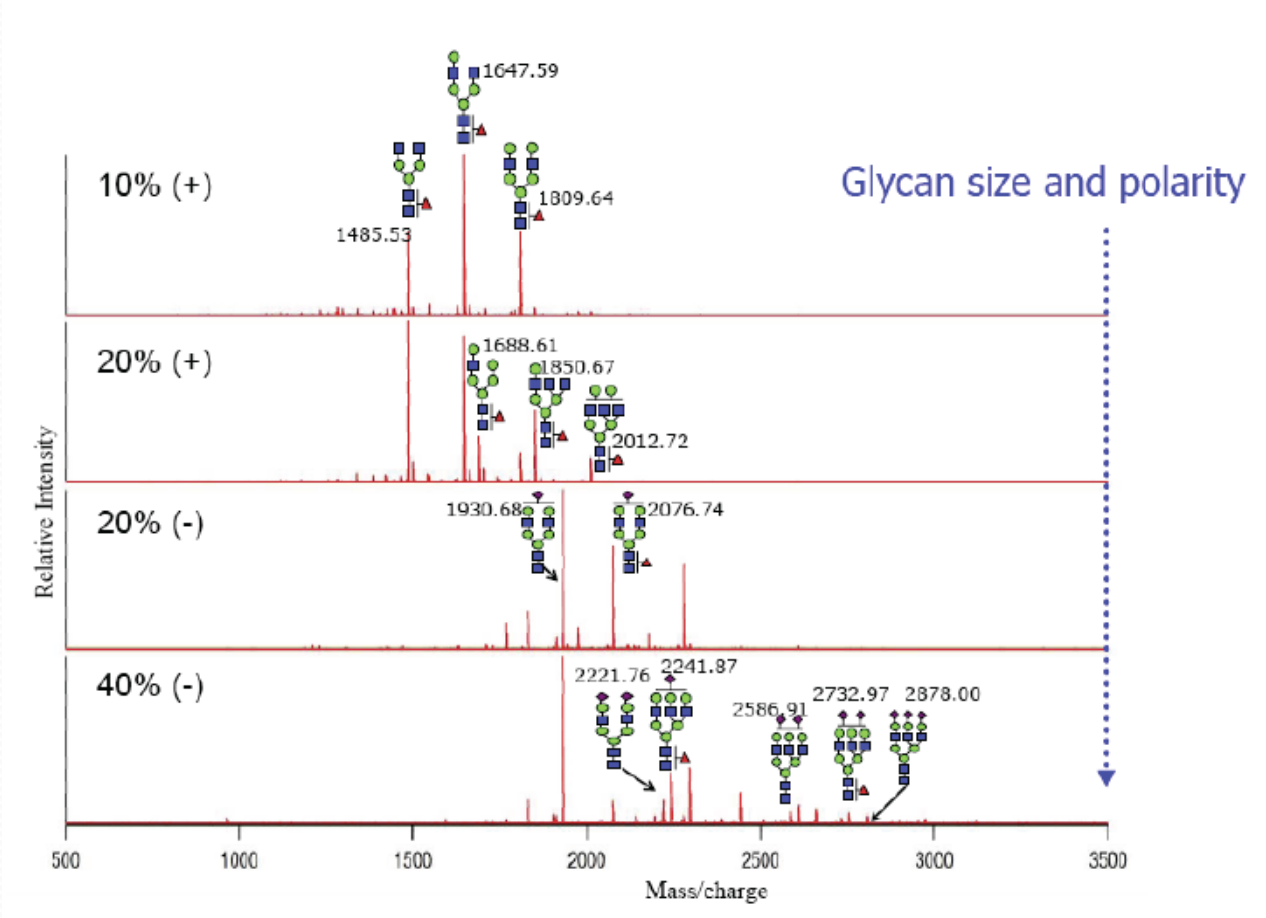
- Matrix Assisted Laser Desorption Ionization
- Sample is embedded in matrix on a slide
- Laser energy absorbed by matrix, ionizes and vaporizes matrix and sample.
- Can fragment compounds
- Can doubly or triply ionize compound—Look for peak at half or a third of the expected m/z
- Results of different shots can be very different—multiple shots for each sample


Laser Impulse
Matrix
with
Analyte
To Mass
Analyzer
Proton
Transmission
Electrostatic
Field
(Lottspeich and Zorbas, Bioanalytic, Spectrum, 1998)

FT-ICR MS

- Ions circulate in a magnetic field.
- Current is cyclical with a given compound having a rotation period determined by its mass.
- Fourier transform puts signals on a frequency basis instead of time.
- This is converted by mathematical formula to m/z
- Induces vary different behavior in the spectra than, for example, time-of-flight MS, in which ion current at a given time following ionization correlates directly to concentration of ions with a specific m/z.

MALDI-MS spectra of N-linked glycans from GOG Sample



Processing Spectra

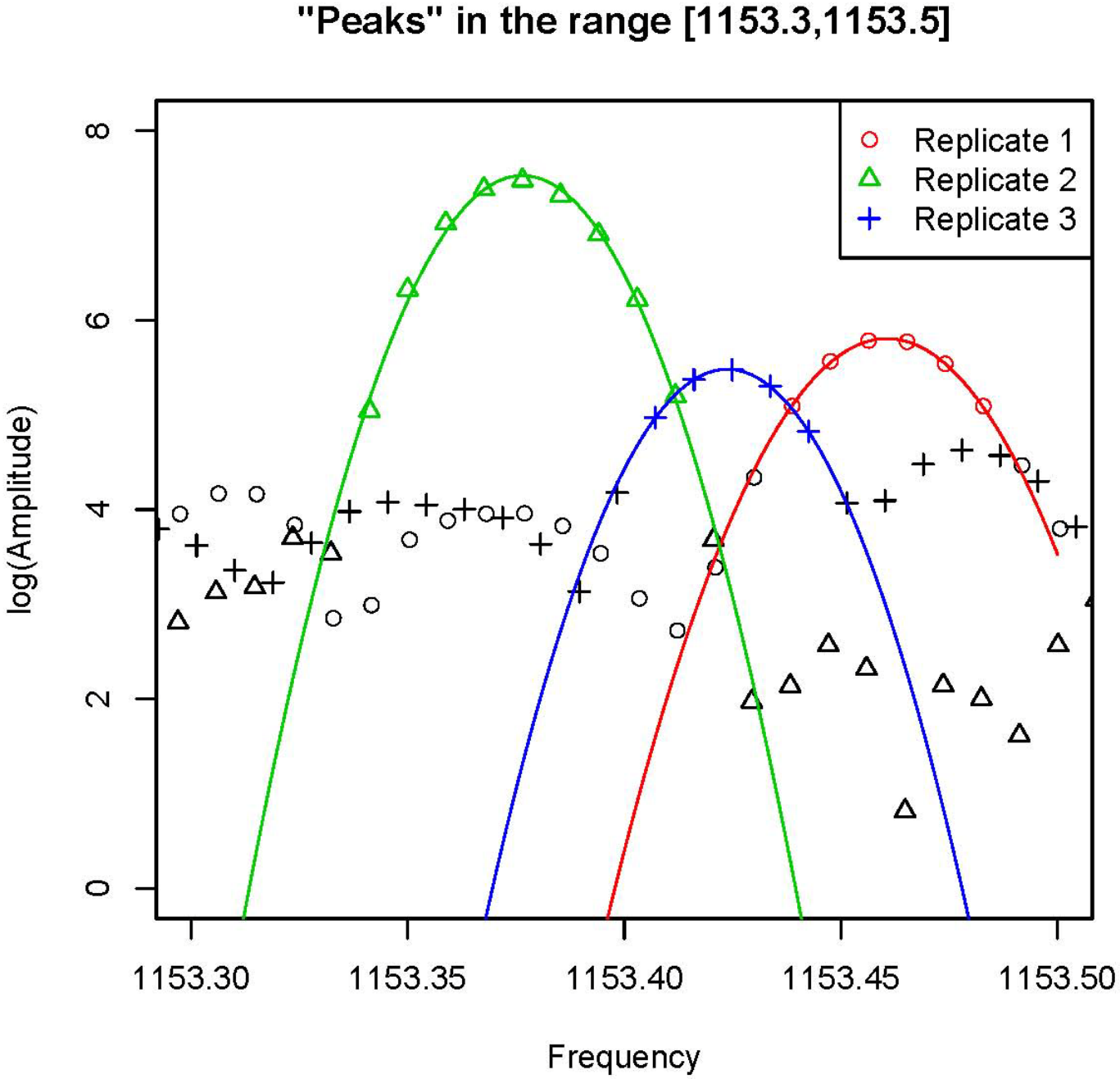
- Baseline estimation
- Peak identification
- Mass calibration
- Data transformation
- Normalizing across spectra
- Peak quantitation
- Identification of isotope and shoulder peaks
- FTICRMS R package

Baseline Correction

- In regions of the spectrum where no analyte is present, the signal should fluctuate around a baseline, which should be set to zero
- In peak regions, the estimated baseline forms the base for peak heights or area.
- For some technologies, the baseline tends to be relatively flat, but for others, such as NMR and FT-ICR MS, the baseline is curved or wavy.
- It is important to get the baseline right to find peaks and accurately measure them.

Peak Identification

- In some cases, a peak is just a point on the spectrum where the points to the left and right have lower magnitudes.
- But this can result in false multiple peaks from jagged signal.
- For FT-ICR MS, the peaks look quadratic on the log scale which allows accurate identification.


"Peaks" in the range [1153.3,1153.5]
Replicate 1
Replicate 2
Replicate 3
log(Amplitude)
0
2
4
6
8
1153.30
1153.35
1153.40
1153.45
1153.50
Frequency

Mass Calibration

- This is supposedly done by in-machine software using known large peaks
- In principle, the mass accuracy at 1000 Daltons should be 0.001 Daltons or better
- In practice, as shown on the previous slide, the same compound can be as much as 0.1 Daltons apart
- Because the theoretical mass accuracy of FT-ICR is so high, this mass calibration process can be done very effectively.
- The accuracy of ToF MS and NMR is much worse.

Data Transformation

- Data from MS and most other technologies needs to be transformed, usually to a log scale to make analysis effective.
- Care needs to be taken at the low end .
  - Logs of zero and negative numbers are not defined
  - Signal fluctuating around a zero baseline will often be negative.
- Shifted log (add a constant), generalized log, and other methods can be used.

Glycan Mass Identification

- The repertoire of monosaccharides occuring in glycans is known and relatively small.
- Given the known structural constraints, all of the glycans less than 2000 Daltons can be computed and a list of masses generated.
- Analysis can be limited to glycans, increasing statistical power.
- Other peaks are possibly glycan fragments, remaining peptides, or other components not eliminated by the isolation and separation

Statistical Analysis

- Use only peaks that are present in a minimum number of samples, impute peaks for samples in which peaks are not detected
- Appropriate statistical analysis per peak depending on the design (e.g., one-way ANOVA, two-way ANOVA, linear regression).
- Correct for multiple comparisons using Benjamini Hochberg False Discovery Rate control methods
- Determine signatures by a variety of methods: logistic regression, PAM, SVM

Selection Effects and False Discovery

- Even with the FDR adjustment, as in any biomarker study with many candidates, some results may be false positives.
- The degree of predictability within samples can be assessed by cross validation
- This needs to encompass all aspects of the discovery process ab initio for each cross validation run
- Then computed biomarkers can be tested on blinded evaluation sets

Biomarker Studies

- Diagnosis: patients with cancer vs. healthy controls
- Differential Diagnosis: patients with a potential problem that is cancer vs. a benign condition.
- Grading and staging
- Prognosis
- Treatment selection
- Treatment evaluation

Diagnostic Studies

- Samples from GOG ovarian epithelial adenocardinoma
- Normal controls identified only by age
- 48 samples of each group, frequency balanced by age, sent to the mass spectrometry lab in 8 batches of 12.
- Blinded
- Each batch had 6 OC and 6 controls
- Randomized selection from sample set, randomized order within and between groups
- Such studies are subject to many difficult bias and confounding problems

Bias and Confounding

- Since the patients and controls are not selected from the same population necessarily, other factors could explain any differences
- Age, sex (if lung, colon, etc.), diet, genetics can all possibly affect any biomarker.
- Only large amounts of data over many studies in many sites can be convincing.

Results

- For the GOG study, 1937 peaks tested
  - 505 significant for diagnosis after controlling for batch and age
- In an earlier study of breast, prostate, and ovarian, markers were found for each cancer, some in common, and some unique
- Also, an earlier study in ovarian cancer showed high predictive ability.
- The latter two are published (*J. Proteome Research, Bioinformatics, Analytica Chimica Acta, BMC Bioinformatics*), the first in preparation.

Differential Diagnosis

- Patients present with suspicious ovarian mass.
  - Serum is analyzed for glycans.
  - Diagnosis: ovarian cancer (stage, grade, type), POS, other non malignant
- Patients present with elevated PSA
  - Prostate cancer or BPA?
- These studies have fewer problems with bias and internal validity
- EDRN PRoBE design
- Ongoing studies at UC Davis

Acknowledgments

- Lebrilla mass spec lab:
  - Carlito B. Lebrilla
  - Hyun Joo An
  - Crystal Kirmiz
  - Scott Kronewitter
  - Maria Lorna de Leoz
- Rocke lab
  - Donald A. Barkauskas
  - Hao Liu
  - John S. Tillinghast
  - Yuanxin Xi
  - Jingjing Ye
- David Woodruff (algorithms)
- Clinical Collaborators:
  - Ralph de Vere White (prostate)
  - Helen Chew (breast)
  - Gary S. Leiserowitz (ovarian)
  - Jay V. Solnick (gastric)
  - Susanne Miyamoto (glycobiology)
- Funding for this work from NCI, NHGRI, NIAID, DOE, AFOSR, and OCRF is gratefully acknowledged.