Representing Errors and Uncertainty in Plasma Proteomics

David J. States, M.D., Ph.D.
University of Michigan
Bioinformatics Program
Proteomics Alliance for Cancer

Genomics vs. Proteomics

Genome sequence

- One copy of each gene per genome
  - *Duplicated genes*
  - *Diploid or higher ploidy*
- No tissue variation
  - *Somatic cell recombination*
- Modification not an issue
  - *Methylation*
- Few sample handling issues
  - DNA is DNA is DNA
  - Very stable

Proteome

- Wide range of expression levels
  - 10-12 orders of magnitude
  - Quantification is a goal
- Tissue specific variation
  - Plasma proteome
- Post translational processing
  - Many known modifications
  - Potentially novel chemistry
- Sample processing matters
  - Serum vs. plasma
  - Many protocols

Errors and Uncertainty in Genomics

- GenBank before there were errors
  - "My lab would never submit an erroneous sequence..."
  - Krawetz survey – 1 in 300 nt later revised or retracted
- Developing a framework
  - Identifying error processes
  - Computational representation
  - Quantitative error analysis
- Setting standards
  - Data use => accuracy requirements
  - Cost vs. accuracy analysis
- Validating lab performance

A Framework for Errors in Genome Sequencing

Genomic DNA

↓

Subclone

↓

Sanger sequencing

↓

Sequence assembly

- Polymorphism
- Degradation
  - Sheering, cross linking, oxidation, depurination
- Polymerase fidelity
  - cDNA
- Clonal instablity
  - Repeats, poison sequences, rearrangements
- Lane tracking
  - pre-capillary
- Base calling
- Assembly errors

Classes of Genome Sequencing Errors

- Single base substitutions
  - Base calling errors
  - Resolution
    - Quality measures in basecalling
    - Redundant data (10X coverage)
- Small insertions and deletions
  - Gel compressions
  - Resolution
    - Improved gel technology
    - Redundant data (forward and reverse reads)
- Large rearrangements
  - Assembly errors
  - Resolution
    - Redundant data (whole genome shotgun)

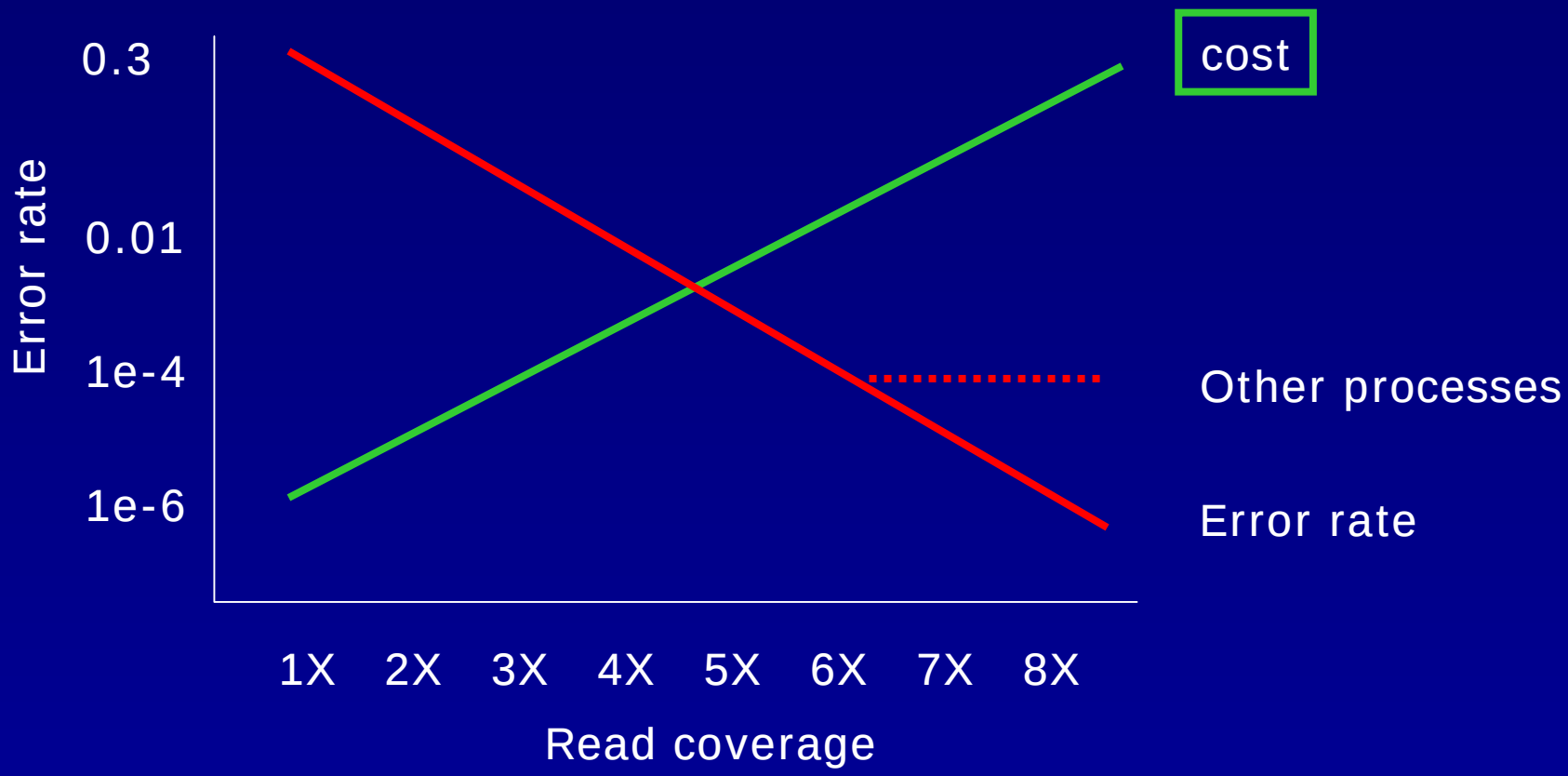
Representing Sequence Errors

- Dominant error processes
  - Single base substitution
  - Small insertions and deletions
- Representation
  - Hidden Markov Model (PHRED)
    - Explicit substitution probabilities
    - Explicit insertion/deletion probability
    - Same representation carries through from reads to assembled sequence
- Sequence alignment
  - N^2 dynamic programming
  - Same as string to string alignment

Setting Standards

- Applications drive accuracy requirements
  - Homolog identification
    - Very error tolerant (evolution as an error process)
  - **Translation of nt to aa sequence**
    - Frameshift and nonsense errors rare in an ORF
    - 1 kb reading frame => 1/10kb error rate
  - Polymorphism identification
    - Error rate substantially below the polymorphism rate
- "Bermuda Quality"
  - Standard
    - No more than one substitution per 10 kb
    - No unclosed gaps
    - No more than 1% of the sequence in gaps
  - Arrived at by community consensus conference
  - Does not meet all needs

Cost vs. Accuracy in Genome Sequence Assembly



Quality Assurance Exercises

- CRADA Funding Mechanism
  - Contract with deliverables
  - Supervision by NIH staff
- Blind resequencing of test samples
- Until you have done a megabase of sequence, you can not really estimate error rates at 1/10kb
- 8 Labs => 3 Centers

Quantitative Assays of Gene Expression

- Multiple sources of error
  - Probe synthesis
  - Target spotting
  - **Affinity reaction**
  - Fluorescence process
    - Photobleaching, quenching, surface effects
  - Image processing
    - Spot detection
    - Background removal
    - Systematic errors across image
  - Data reduction
    - Statistical significance
    - Systematic and data dependent error processes

Challenges for Quantitative Error Analysis in Proteomics

- Many types of errors
- Many sources of errors
- Quantification as well as sequence
- Rapidly evolving technology
  - Difficult to compare labs (apples and oranges)
  - Difficult to gain experience

Classes and Sources of Errors in Proteomic Identification

- Missing data
  - Samples degraded, insoluble, adherent
  - Partial or incomplete proteolysis
  - Ionization is variable and not well understood
- Identification errors
  - Ambiguous matches (esp. 2D gel w/o MS)
  - Databases incomplete and error prone
  - Polymorphism
- Sequence determination
  - Degenerate masses
  - Missing data, fingerprints vs. partial vs. *de novo* sequence determination
  - Incorrect assumptions (gene models, alternative splicing, polymorphism, etc.)
- Polymorphism
- Structure of post translational modifications
  - May be complex and variable (e.g. glycosylation)
- Location of post translational modifications
  - May be ambiguous (e.g. there is a phosphate on the peptide but there are 2 serines, a threonine and tyrosine)

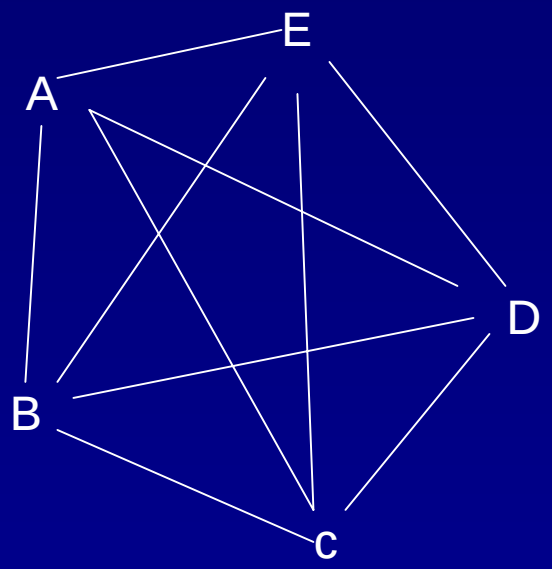
Errors in Quantification

- Sample handling
  - Acquistion, storage, etc.
- Affinity vs. resolution methods
  - Affinity based methods (arrays, etc.)
    - Easy to score
    - Uncertainty in affinity reagent and binding
  - Resolution
    - Potential for systematic errors
    - Harder to score
- Absolute vs. relative quantification
- Ambiguity in identification

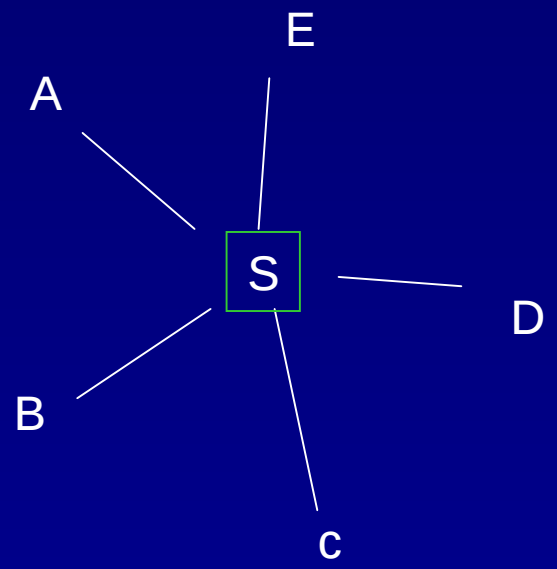
Computational Representation of Uncertainty

- Pilot phase
  - Track technology and protocol
  - Anticipate continued technological evolution
- Multiple technologies
  - Ambiguity in identifiers
  - Technique specific errors
- Multiple users and applications
  - Biomarkers to screen for unknown disease
  - Monitoring known disease
  - Investigation of disease processes
  - Research on biological processes

Need for Data Standards



N^2 problem
Changes at one lab affect all



N problem
Changes local to a lab

Standards in the Face of Multiple and Evolving Technologies

- XML
  - Widely used extensible standard
  - Rich representations are possible
  - Intrinsic support of hierarchical organization
- Upward compatibility is key
- Hierarchical organization
  - Common denominator representation at higher levels
  - Progressively more lab and technique specific representations at lower levels
- Problem: multiple hierarchies
  - Fractionation (2D gel, chromatography, …)
  - Identification (MS, MS/MS, affinity, …)

Databases and Publications

- Database entry
  - Very widely “read”
  - May be a low priority for the lab
  - Not subject to peer review and editing
- Quality is an issue

- Publication
  - Smaller number of readers
  - Key indicator of productivity and reputation
  - Peer reviewed and professionally edited

Moving Forward

- PPP as a lead problem in proteomics
  - Sample prep is comparatively well defined
  - Lots of prior data, labs and users
- Input from experimental groups
  - Anticipated as well as current platforms
  - Defining classes of errors
  - Quantitative error and reliability data
- Input from the user community
  - Anticipated applications
  - Defining data reliability requirements for specific applications