

Wrapping Up

Biol4230

Bill Pearson wrp@virginia.edu

Tues, May 2, 2016

4-2818 Jordan 6-057

What we didn't cover

- variation – SNPs, structural variation
- function prediction – when do homologs have the same function?
- mapping sequencing reads
 - SAM/BAM files to genomes
 - to exomes / transcripts
- peak finding (for ChIP seq, epigenetic marks)
- clustering algorithms (k-means, etc)
 - finding sets of genes with similar expression/response patterns
- machine learning strategies
 - neural nets, SVMs, PCA

fasta.bioch.virginia.edu/Biol4230

1

Biol4230 – what we did cover:

- similarity searching for homology
 - homology is inferred from excess similarity
 - significant implies homologous
 - not-significant DOES NOT imply non-homologous
 - scoring matrices affect search sensitivity, alignment boundaries
 - scoring matrices have an evolutionary model
 - Position-Specific Scoring Matrices (PSSMs) and HMMs build a "central" model for a protein domain family
 - SP (sum of pairs) multiple sequence alignment is not biological
 - (rigorous) multiple sequence alignment is exponentially hard

fasta.bioch.virginia.edu/Biol4230

2

Biol4230 – what we did cover

- phylogenetics and tree reconstruction
 - types of trees (rooted, unrooted), always binary
 - numbers of trees (very large)
 - optimality criteria : always minimizing something
 - distance : observed vs "measured" similarity on tree
 - parsimony : minimal amount of change
 - maximum likelihood : tree that best fits the data (and model)
 - evaluating tree accuracy
 - simulations (with different models)
 - experimental evolution
 - bootstrapping
 - identifying positive and negative selection: dN/dS

fasta.bioch.virginia.edu/Biol4230

3

Biol4230 – what we did cover

- The human genome project
 - building complete "genomes" from pieces
- RNA expression analysis
 - RNA abundance vs protein abundance
 - the RNA abundance problem – many orders of magnitude between lowest and highest
 - looking for differential expression – how to normalize?
 - correcting for multiple tests (FDR)
 - looking for sets of co-regulated genes:
 - over-representation analysis (GO terms)

fasta.bioch.virginia.edu/Biol4230

4

Biol4230 – what we did cover

- Identifying functional sites
 - not homologous
 - short, not well conserved
 - not significant (in the entire genome context)
 - represent with PWM (position weight matrix, PSSM)
 - estimation with missing data (alignment/PWM)
 - predicting binding from protein structure

fasta.bioch.virginia.edu/Biol4230

5

Bioinformatics – the big picture

- Lots and lots and lots of data
 - is it "clean" enough?
 - do discrepancies in the data reflect biology, or technology
 - what inferences/conclusions are reliable?
 - $E() < 10^{-6}$ implies homology
 - what assumptions have been made?
 - multiple sequence alignment requires homology
 - GO experimental terms are "better" than BLAST results
 - the database is complete
 - the protein predictions are accurate

fasta.bioch.virginia.edu/Biol4230

6

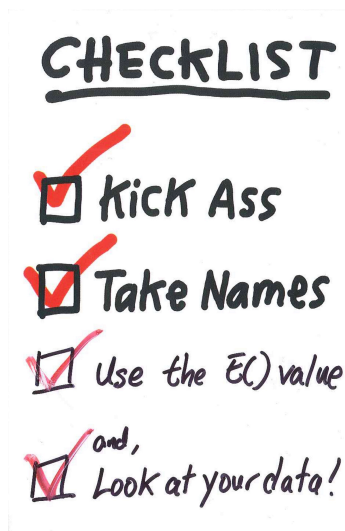
Bioinformatics – the big picture

- Why could this result be wrong?
- Does it make sense
 - can 100% identical sequences have different functions?
- What is the control?
 - what kinds of errors does the control detect?
 - what kinds of errors does it miss?

fasta.bioch.virginia.edu/Biol4230

7

Bioinformatics – the big picture



fasta.bioch.virginia.edu/Biol4230

8