

CBB752 Final Quiz, Spring 2010

Name: KEY

200 points in total

1. Explain the difference between local and global optimization [8 points].

Global: optimizes over entire data range

Local: optimizes over a subset of the data space

2. What is the difference between a deterministic and stochastic model? Which term applies to Ordinary Differential Equation (ODE) models? [8 points]

- deterministic: same results for same initial parameters every time

- stochastic: probabilistic/outcomes not always the same

- ODEs = deterministic

3. In the context of mathematical modeling, what is an F test used for? [12 points]

F-test is used to determine if a particular parameter is worth including in a model

4. Assume that B and M are two different cell types (e.g., Naïve B cells (B) and memory B cells (M)). Write a brief description of the potential meaning for each parameter in the model: [20 points]

$$\frac{dB}{dt} = s + pB - cB - tB$$

$$\frac{dM}{dt} = tB - cM$$

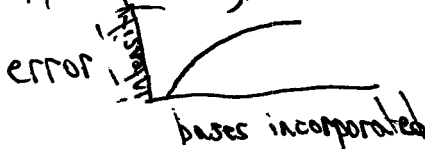
s: source of B cells

p: growth rate of B cells

c: death rate of B, M cells

t: rate at which B cells are converted into M cells

5. Describe how 454 sequencing (pyrosequencing) and Illumina sequencing (sequencing-by-synthesis) work. Describe the main type of error that occurs during 454 sequencing and why it occurs? Why is the read length produced by the Illumina platform limited? [20 points]

454 key points: pyrosequencing, flow 1 dNTP at a time, measure luminescence, picotiter plates beads
 error:  single nucleotide repeats/indels

Illumina key points: bridge PCR, flowcell/slide, reversible terminator chemistry, Flow all 4 dNTP types at once
 read length, limited by physics/location/compactness on flowcell. As seq. goes strands will bend/run together at 3' end

6. What are the advantages of single-molecule based sequencing compared to amplification based approaches? [10 points]

- No amplification (less bias)
- More quantitative

7. What are the two main confounding factors in DNA sequence assembly? In this context, describe the meaning of the N50 statistic. [10 points]

- repeats
- polymorphisms

N50: N50 is the longest length L such that 50% of the bases in a contig are at least length L . Statistic to assess assembly quality

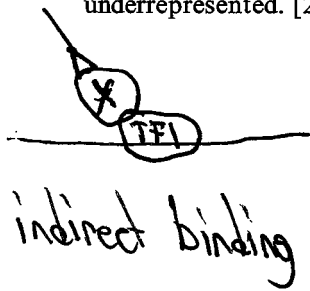
8. The majority of reads in a ChIP-Seq experiment are background [5 points]

9. When scoring ChIP-Seq data, what is the purpose of the peak shift? What statistical distribution is typically used to assess significance of a ChIP-Seq peak? [15 points]

- locate most probable location of transcription factor binding site on DNA

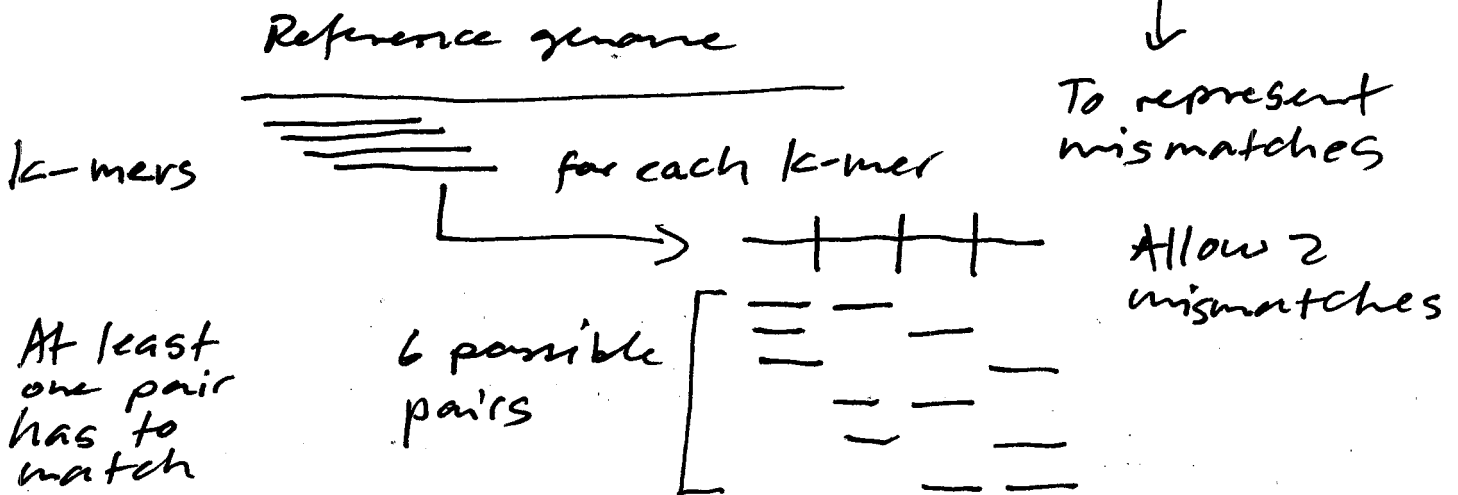
- negative binomial or Poisson

10. After running a ChIP-Seq experiment targeting factor X, you find that only 30% of the peaks contain the known motif for factor X. Assuming no experimental error, explain using *diagrams and text* the possible reasons why the motif may be underrepresented. [20 points]



Also accepted antibody specificity + peak calling

11. In detail using text and diagrams, explain how an index-based short-read mapping algorithm works. Why are spaced-seeds necessary when using this approach? Make sure to show the spaced seeds approach in your diagram. [20 points]



12. What is the main drawback of microarrays compared to RNA-Seq? Name three new applications using RNA-Seq that were not possible with tiling arrays. [10 points]

cross-hybridization

- RNA editing
- fusion transcripts
- alternative splicing (isoform detection)

13. Rank the following program in terms of sensitivity (1 = most sensitive, 6 = least sensitive) for finding sequence alignments: [12 points]

- 6 Bowtie
- 5 BLAST
- 4 Smith-Waterman
- 1 HMM
- 3 PSI-BLAST
- 2 Sequence Profiles

14. What are the advantages of using a Biplot to represent a data matrix? [10 points]

- Biplot aims to represent both the observations and the variables of a matrix on the same plot

15. Explain how PCA works and why it is relevant to biological data sets? What is the interpretation of the principal components and what are the loadings? Explain the role of PCA in eigenfaces analysis presented in class. [20 points]

- PCA is used to reduce the dimensionality of a data matrix
- PCA finds the directions of highest variance in a given data matrix by spectral analysis of its correlation/covariance matrix
- Many biological data sets have a high number of dimensions
- Principal components are orthogonal vectors pointing along the highest variance in the data matrix
- Loadings are the contributions of each variable to a principal component
- Eigenfaces: Each principal component shows a different source of variation in the image of human faces in the order of their contribution to total variance (hair color, ...)