Review questions

Zackery Field

January 28, 2014

Readings for this lecture, and some subsequent lectures

"'Antedisciplinary" Science

"Pre-review" Questions

1. What is meant by genome annotation? Attaching **structural** and **functional** information to genomic data.

2. What is the difference between a structural and a functional annotation of a genome? Functional concerns what the gene actually does. Structural seeks to identify specific regions in the genome without concern for their specific function.

3. What types of errors occur in a structural annotation of a genome? Commonly miss an exon, because of a poor gene model. Guesses that %25 Eukaryotic gene structural annotations are incorrect. Might miss a gene completely, false negative. Could catch a **pseudogene** (remnant of what was once a fully functioning gene), false positive. **Synteny** is the identification of equivalence across species of regions of a genome. **Duplication** is a natural process, and while it can be considered an error, it is more of a fortuitous accident.

4. What is meant by a gene model? Some general mapping of what is meant by a gene, and a data

5. How do errors in gene models affect a predicted function?

6. What do we mean by protein "function"?

7. How do we determine these functions experimentally?

8. How precise are these experimental techniques?
9. How are experimental techniques benchmarked? (Are they?)
10. How do we predict these functions using bioinformatics tools?
11. How good are our predictions?
12. How do we evaluate prediction methods?
13. Can we determine, for a single gene, whether the functional annotation is correct?
14. What is the typical functional annotation protocol? When you BLAST a large database with some protein dataset and look for proteins that have some significant likeness to the queried protein. Just see the directions before. **Homology-based annotation transfer protocol**
    (a) Run BLAST
    (b) Identify top hit
    (c) Check for significance of score
    (d) significance is some E-value (usually) <0.001
    (e) Some further criteria must be met, like informativeness of transferred annotation
    (f) transfer annotation
15. What are the fundamental assumptions of that protocol?

    Assumptions
    (a) Assumes that initial annotation is correct **%25 WRONG**
    (b) Sig. BLAST score means homology **Promiscuous domains could cause issue, but they are hidden by Seg software**
    (c) Homologs can be detected in the above process (short queries problematic)
    (d) **homology implies same function**

    What fraction of genes in a genome are "hypothetical genes" **30%**.
16. What are the sources of error in that protocol?

    There are some issues defined above in bold. BLAST may make a homology match, but this does not imply that there is a functional similarity. BLAST does not differentiate between paralogs and orthologs. paralogs often do not share the same function, but they can.
17. What is the estimated functional annotation error rate in the standard sequence databases?

18. What modifications to functional annotation protocols are needed to avoid these errors?

19. How can you personally detect possible errors in an existing functional annotation, using bioinformatics tools and analysis?

20. What is a pseudogene?

21. What is meant by horizontal gene transfer (HGT)?

22. What is meant by a proteinâĂŹs multi-domain architecture (MDA)?

23. How can you determine a MDA?

24. Describe a typical use of the BLAST webserver by a biologist.

25. What information does BLAST provide?

26. Describe a typical use of the Pfam webserver by a biologist.

27. What information does Pfam provide?

28. What is meant by a Pfam clan, and what other database(s) provide similar information?

29. What is meant by a missense mutation?

30. What is meant by a nonsense mutation?

31. What is meant by a conservative substitution?

32. List the two most conservative substitutions for isoleucine. (What was the basis for your answer?)

33. What is meant by the redundancy of the Genetic Code?

34. Describe the Central Dogma of Molecular Biology.

35. What information does protein 3D structure provide?

36. Name two experimental methods used to solve protein 3D structures?

37. How accurate are protein structures?

38. What are the two main classes of approach to predicting protein 3D structure?

39. What is the Gene Ontology? What are GO evidence codes?

40. What two bioinformatics databases provide hierarchies of structural domains, elucidating their structural similarities and related functions? (You should be able to describe how one of these databases organizes protein domains, and what types of relationships between domains are implied by that organization.)

41. What is meant by gene fusion and gene fission?

42. What is meant by a gene duplication event?

43. Name the two major protein sequence databases with the largest coverage (of both species and functions).

44. Name one protein sequence database that is manually curated.

45. Name one major bioinformatics database/webserver that enables biologists to evaluate the structural superposability of protein 3D structures.

46. What database is the repository of protein structures?