Literature assignment 2 **Due October 14th, 2011**

Article: Edgar, R. *Quality measures for protein alignment benchmarks* Nucleic Acids Res. 2010 April;38(7):214553.

In class we discussed basic multiple sequence alignment algorithms. Benchmarks have been created to assess how well different algorithms perform. This article compares several existing benchmarks and discusses what makes a good benchmark.

Read this article and briefly answer the following questions in your own words. You may read additional materials if you wish. If you do, you must cite your sources. **Please turn in your assignment in class or in person to Annette McLeod in MI 646 or electronically to comp-bio@cs.cmu.edu by 5pm on the due date.**

1. How is structural alignment information used to assess the quality of multiple alignments of amino acid sequences? How might it help determine if the sequence is over- or under-aligned?

2. BALIBASE reference alignments are provided in two versions, trimmed and untrimmed.

   (a) What is the difference between trimmed and untrimmed reference alignments?

   (b) How could untrimmed reference alignments be useful in benchmarking?

3. What are Edgar's concerns with benchmark alignments that include gaps?

4. Edgar discusses several measures of accuracy proposed by different benchmarks. In some, the accuracy score depends on the fraction of residue pairs whose alignment in the reference alignment is correctly reproduced while in others the accuracy depends on the whole column being aligned correctly. What are the pros and cons of using each of these methods? Which does Edgar prefer? Why?

5. Some measures of accuracy are based on the number of pairs or columns in the *reference alignment* which are correctly reproduced (for example, SPS and CS). In contrast, F_M is based on the fraction of pairs in the *test alignment* that are aligned the same way as they are in the reference alignment. Which of these approaches is preferable and why?

6. Reference alignments in SABMARK are constructed via pairwise alignments.

   (a) How is this done?

   (b) This approach leads to an issue with consistency. How does Edgar assess consistency in multiple alignments?

   (c) What do Edgar's results on the consistency of SABMARK suggest about progressive alignment methods?

7. Edgar evaluates multiple sequence alignment benchmarks in terms of coverage. What does he mean by coverage and why is it important?

8. Edgar states that the ideal multiple sequence alignment benchmark will contain sequences in the "twilight zone".

   (a) How does Edgar define the "twilight zone"?

   (b) Why is the "twilight zone" the ideal range of sequence identity for these benchmarks?

   (c) Give an example of an application where one might need to align a set of sequences in the "twilight zone"?

9. Edgar does not favor BALIBASE as a reliable benchmark.

   (a) How do his results support his conclusion?

   (b) What underlying characteristics of BALIBASE does Edgar identify as causing the observed problems?