Wintersemester 2014/2015, AST

Revision questions, part 2

Cluster analysis

1. What is the euclidean distance measure? What is the Manhattan distance measure? In two dimensions, three dimensions, and n-dimensions?
2. What problems can you encounter when using k-means clustering?
3. What is the difference between agglomerative and divisive hierarchical clustering?
4. What is a "linkage method" used for in agglomerative hierarchical clustering?
5. What problems can you encounter when using agglomerative hierarchical clustering? Think about potential problems associated with the linkage methods?

Protein domains

1. Name some of the different ways in which protein domains have been defined.
2. Why might evolution lead to modular proteins composed of domains? Would this process prefer domains to be contiguous or discontinuous in a protein chain?
3. How could one find protein domains given only a) the sequence or b) sequence and structure of a protein?

Protein function prediction

1. Which data could we use to determine the function of a protein without going into the lab?
2. How are the known functions of proteins stored in databases such as the PDB?
3. How could two very different protein sequences still have the same function?
4. How can two proteins with very similar sequences have different functions?
5. I have solved the structure of a protein and would like to predict its function. There is a protein that is very similar to it at the sequence and structural level. Why could it still be possible that my protein has a different function?

Protein fold recognition weak similarities

1. Why might a vague idea of a protein structure (aka a bad model) still be useful?
2. When aligning a sequence to a structure, how could one improve a sequence-based scoring matrix with information from the protein structure?