Assignment 1

for lecture "Bioinformatics III" WS 08/09

Return before lecture on Oct. 30, 2008 or by email to p.walter@bioinformatik.uni-saarland.de
This assignment will be discussed in the tutorial on Nov. 3, 2008, room15, building E1 3

1. Intro to Python

For the programming tasks of the assigments we recommend Python (or any other scripting language as, e.g., Perl). You can find a lot of documentation on Python online in the "documentation" section of www.python.org. The most important documents are the classic tutorial by Guido van Rossum and the "Python Library Reference". The task of this assigment is to introduce you to network measures and the handling of data files. In order to reduce the complexity for this assigment you will in part 1 deal with the random network only, the comparison to the scale free graph will take place next week. To avoid to code the same stuff over and over, today's task is split up into separate modules, which will be reused later. Please hand in your solution including the (important parts of the) source code on paper (typed or handwritten) or electronically (as a single printable PDF file — don't expect me to have the latest MSWord or OpenOffice).

The aim of the following exercise is to briefly introduce you to the programming aspects of Python. This part is optional i.e. you don't get any points for submitting the solutions. The python programming language can be overviewed as consisting of the following main points: Variables, Loops, Conditionals, Functions, Tuples, Lists, Dictionaries, Classes, Importing Modules, File I/O, Error Handling.

- Download the pdb file for SUCROSE-SPECIFIC PORIN (PDB ID 1a0s) from www.rcsb.org.
- Write a python script to open the file, read and write the contents to another file in lower case. The program should be able to obtain the new filename from the user. Hint: don't forget to use the import statement. Example: import os, sys, string
- How many lines begin with "ATOM"? Hint: look at the inbuilt function "startswith()".
- Print the lines that start with "ATOM". What is the frequency count of each amino acid?
- Make a dictionary with the 20 amino acid residues as the keys. Read the pdb file line by line. If the line starts with "ATOM" then use the function string.split() to convert the line (which is currently being read a string) into an array. Look at the 3rd element of the array. This should be the amino acid residue name. Hint: find out more about python dictionaries, keys and values.
- Increment the amino acid count in the dictionary consisting of residue names.
- Write a FUNCTION to iterate through the dictionary and find the total number of amino acid residues.
- The HETATM records present the atomic coordinate records for atoms within "non-standard" groups. These records are used for water molecules and atoms presented in HET groups. (http://www.bmsc.washington.edu/CrystaLinks/man/pdb/part_67.html). How many HETATM are HOH in 1a0s? Hint: read the pdb file, if line begins with HETATM, convert it into an array, look at the 3rd element of the array.

2. The random network (75 points)

a) Implement an algorithm for creating a random graph. Start from a given number of nodes and add one link after the other. Take care not to add the same link twice. Store the constructed network in a file.
Proposal: Create two classes in python with the following methods:

Graph:	Node:
def addNode(self, node):	def getVertices(self):
def addVertex(self, n1, n2):	def getDegree(self):
def isVertex(self, n1, n2):	def getId(self):
def getNode(self, id):	def setVertex(self, id):
def printHisto(self):	

The Graph class creates a new network and stores all the nodes (e.g. in a list) which are represented by instances of the Node class. Besides Graph includes methods for performing operations on the graph as well as for the output (e.g. the values for a histogram). A Node instance stores its degree, a set of connected nodes (=vertices) and an identifier (id).

b) Create some random graphs with the following number of nodes and vertices (each given in the form nodes,links):
500,100 500,200 500,400 500,800 500,1600 500,6400
Plot the histograms (x-axis: degree, y-axis: frequency) for these graphs. Which trend can you observe?

c) The degree distribution of random graphs follows the Poisson distribution:

$$P(k)=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$$

Consider a random graph with 1845 nodes and 6234 vertices. Determine λ from the number of nodes and links.

- Calculate the probability that any node has exactly 3 vertices.
- Calculate the probability that any node has between 1 and 6 vertices.

3. Existing networks (25 points)

Characterize (with a short explanatation) the following examples of networks into the following categories:
random, scale-free, hierarchic, clustered
Which of the following networks belong to one of these categories? Some of the examples might fit into more than one categories. If so, explain your choice.

- Evolutionary tree
- Global banking system
- World Wide Web
- German road system

Contact: Peter Walter, phone -3613, email: p.walter@bioinformatik.uni-saarland.de