# Instructions

Allowed tools: Paper, pen and dictionary.

Write name and personal number on here:

……………………………………………….

Write the secret number given to you by on **EVERY** page and here:

……………………………………………….

If you want your result sent to you be email write your email address here:

………………………………………………

Maximum number of points on the exam is 100. Another 100 points is given at the online exam. The grades will be based 50% of the online and 50% of the written exam. To pass the course you need at least 40 points at each of the exams.

Question 1 at the exam can be skipped if the oral paper presentation is finished in time. This is true for all that presented the project on Sep 17.

Grades are accordingly to the following

* A 92-100%
* B 80-91%
* C 68-79%
* D 56-67%
* E 50-55%
* Fx 45-49%
* F 0-44%

Each question should be answered within the provided space, if you need more space write on the back of the paper.

You can answer in either English or Swedish

Arne can be reached at 070-6951045 for questions, and will pass by a few times during the exam.

Short questions (10 p per question):

1. Describe how ClustalW (and other progressive multiple sequence alignment methods) works (10p) ?

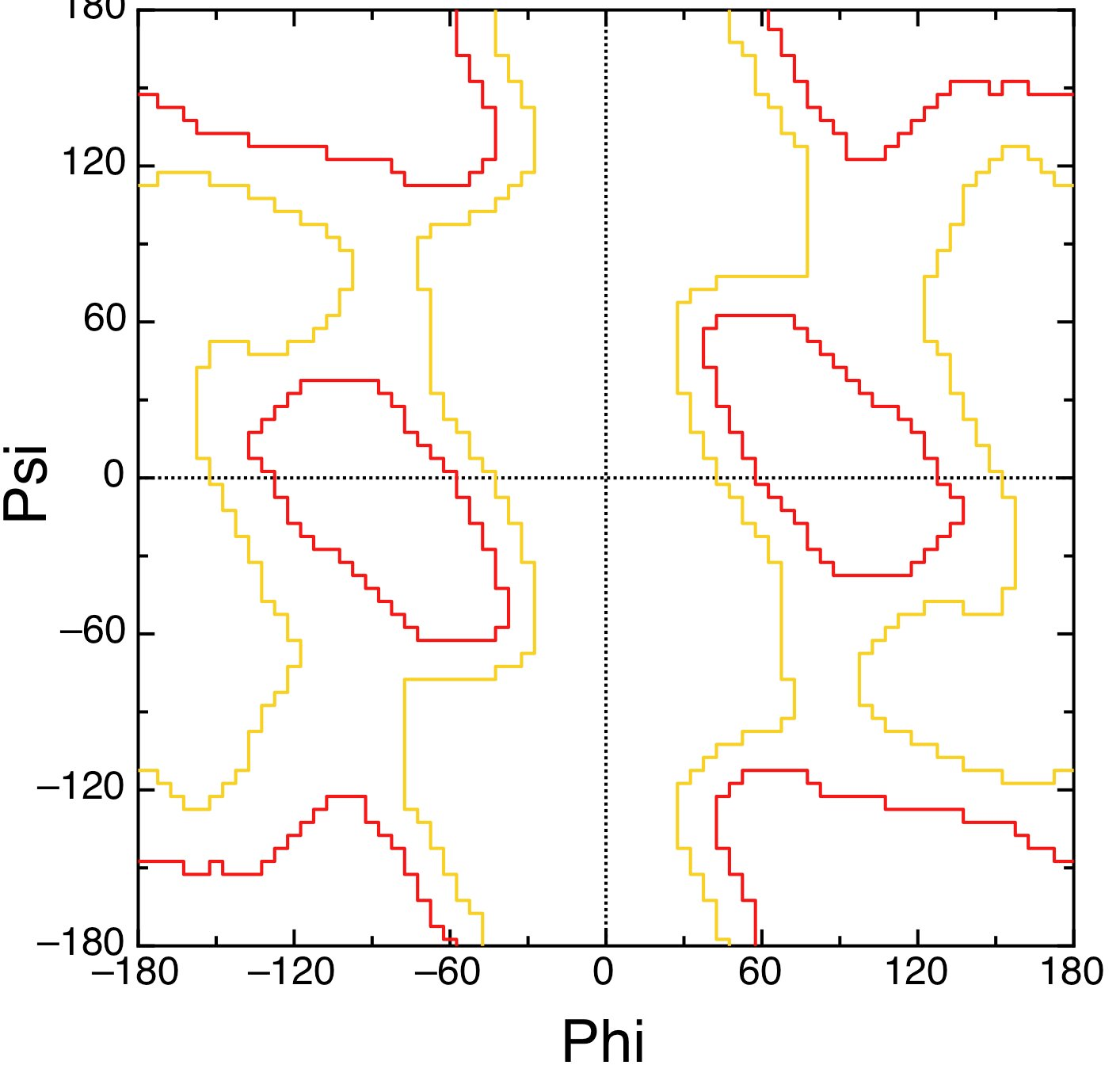
2. Uniprot is a commonly used database and it consists of several sub-databases. (a) Describe what type of information can be found in Uniprot (5p) ? (b) What are the differences between Swiss-Prot and TrEMBL (3p) ? What is Uniref50 (2p) ?

3. What is the “power law–like distributions of paralogous family size in a genome”? Explain how this describes the different protein family sizes (5p). Which fundamental evolutionary principle might have caused this distribution (5p)?

4. Describe and draw the TMHMM method that is used for membrane protein topology predictions (5p). How is the length of the helices modeled (2p)? What is the “positive inside rule” and how is it taken into account in the model (3p)?

5. (A) How many degrees of freedom (e.g., angles) do you need to describe the backbone conformation of a protein chain of length N residues? (5p)

(B) What amino acids does this Ramachandran plots represent (5p)?



6. Given an identity matrix (score of 1 for matches and 0 for mismatches) and a gap penalty of -1 per gap what is the optimal local alignment and score of the following sequence pairs (5p per pair):

IKLALL / IKAL

WYIKLLLL / WIKLLLK

7. Describe how PSI-BLAST works (5p). Describe also how it can be both faster and better than a standard Smith-Waterman alignment (i.e. a dynamic programming alignment) (5p).

8. (a) Describe how a supervised machine learning method works (5p)? Describe some factors that are important to obtain the best result. (b) In particular what is overtraining (3p) and (c) how can it be avoided (2p)?

*Long question (20p).*

9. Recently there has been huge progress in the ability of predicting contacts between interacting residues in a protein. (a) Describe how these contacts are identified (5p) and what was the major breakthrough that enabled these methods to improve significantly in the last few years (5p). (b) Now imagine that you want to use these methods to predict interactions between proteins. How would you develop a method that (i) predicts contact between homo-dimers and (i) between two non-homologous bacterial proteins that are known to interact? Describe potential problems and pit-falls and how you would test the method.