

# Biological Data Analysis (CSE 182) Final project: A Proteomics Toolkit

## 1 Logistics

The final 1 or 2 lectures of the class will be devoted to final presentation of the project. You are required to work on one of the projects described below. The project itself should be easy. However, you may need some time to understand what is required, and checkpoints have been created to help you stay on track.

Excellent proteomics resources exist on the web, and you should scan them. A good site to visit for proteomics is <http://prospector.ucsf.edu>.

### Checkpoints

**C1:11/2/04** Submit a one page written report with answers to the following questions:

1. Your project partner's name, if any. Teams of 2 people are recommended.
2. Give the residue mass for the peptide SAMPLER.
3. Draw the MS only isotope distribution corresponding to SAMPLER assuming  $Z = 1$ , and  $C - 13$  as the only isotope. Draw it for the peptide SAMPLESAMPLER, assuming  $Z = 3$ .
4. Output all  $b$ -ions,  $y$ -ions, and  $y - H_2O$  ions for the peptide SHDR. Which neutral losses (Loss of  $H_2O$ , and  $NH_3$ ) are not likely to be seen, and why?

**C2:11/15/04** Submit a short report on the following.

1. Select one of the projects described below. The remaining requirements hold for all projects other than web-integration.
2. Describe using pseudocode and text, your approach to the problem, and a plan to test your code.
3. A diagram of the web-input, and web-output of your project. This will be supplied to the team developing the web-application.
4. Code that can read in the data-set appropriate for your problem, and produce some simple output.

**C3:11/22/04:** Web-integration checkpoint. Demonstrate a web-mockup of all of the proposed tools. This applies only to the team doing the web-integration.

**C4:11/29/04** Submit a report on your final project. Schedule a demonstration with your instructor for that day. Receive test data, and prepare for presentations in the final two classes.

## 2 Project Details

1. : *Isotope Number* and *Charge* prediction:

**Input:** A file containing multiple MS/MS spectra

**Output:** For each peak  $p$  in each spectrum, a number  $i_p \in \{1, 2, 3, 4\}$  signifying position in the isotope cluster, and a number  $Z$  indicating the charge on the peak. For the final submission, you must annotate the peaks of some test spectra.

**Notes:** Note that many of the peaks do not have isotope clusters. They can be given  $Z = 1, i_p = 1$ . The charge can be predicted by taking the reciprocal of the distance between adjacent peaks in a cluster. There may be noise peaks of low intensity in between adjacent isotopic peaks. Training and test spectra will be provided.

2. Signal peak detection

- Input:** A file of MS/MS spectra
- Output:** For each peak, a number  $s$ .  $s = 0$  if the peak is a noise peak, and 1 otherwise. For the final submission, provide results on the training spectra, and return your output on the test spectra.
- Notes:** Noise peaks are typically low-intensity. However, the intensity of signal peaks is also lower at low and high values of  $M/Z$ . therefore, you must select signal and noise peaks within a window. Training spectra will be provided.
3. Parent Mass Correction
- Input:** An MS/MS spectrum, and a parent mass.
- Output:** Corrected parent mass of the peptide. Training spectra will be provided with correct parent mass. You should provide results on the training spectra after assuming either an incorrect mass (shifted away), or with no knowledge of the mass at all. You should also provide your corrected parent mass on test spectra.
- Notes:** Note that a correct annotation of the spectrum should have many suffix and prefix ions corresponding to the same fragment, which can be matched only with the correct parent mass.
4. Visual display of annotated spectrum
- Input:** An MS/MS spectrum with the following annotations on peaks: Signal/Noise, Charge, Isotope-number, Ion-type.
- Output:** A gif file displaying the annotated spectrum. The code should have switches to turn-off annotations (EX: display only the b-ions).
- Notes:** Note that not all peaks may be annotated, so your code should gracefully handle missing annotation. This could be a 3 person project. An option here is to develop a Java applet with zoom and scroll ability.
5. Spectral Annotation, and Theoretical spectrum calculation.
- Input:** For the calculator, the input is a peptide and a list of ion types & their relative intensity. For the annotator, the input includes a spectrum.
- Output:** For the calculator, produce a list of peaks corresponding to a theoretical spectrum of the peptide. For the annotator, annotate all peaks of the spectrum with the following Signal/Noise: Charge: Isotope Number: Fragment ion type.
- Notes:**
6. Isotope calculators
- Input:** The input is either a peptide, or a mass value.
- Output:** The output should contain the molecular composition of the peptide, the first 5 isotopic peaks and their positions assuming a charge of 1. You should test your output against a similar calculator at <http://prospector.ucsf.edu>.
- Notes:** Note that you will need to know the natural isotopes of Carbon, Oxygen, Sulphur, and Nitrogen, and their abundance in nature. If the input is only a Mass value, and not a peptide, you should assume a composition corresponding to the fictional peptide Averagine. See <http://prospector.ucsf.edu>.
7. Web-Integration of tools.
- Input:**
- Output:** A master web-page listing all of the tools. Based on input from other teams, design input forms for all tools, and html versions of the output from the tools. Integrate all of the tools.
- Notes:** This needs some experience with web-programming (HTML/CGI/servelets). Please talk to the instructor for detailed instructions. Team of 3 might be appropriate.
8. *De novo* identification of paired spectra
- Input:** A pair of spectra corresponding to the same (unknown) peptide. One of the spectra corresponds to a PT modified peptide.
- Output:** Use the characteristic shift of the modification to identify the prefix and suffix ions. Use this knowledge to reconstruct the peptide sequence.
- Notes:** Use the spectral deconvolution algorithm and the following paper (Pevzner et al., Genome Res. 2001 Feb;11(2):290-9.) as starting point. Speak to the instructor if you have more questions.