Math 2300 **DISCRETE MATH** Spring 2017
Dr. Duval **Project 2**
due Tuesday, April 4

Mass Spectrometry

In this project, we'll use counting to look at mass spectrometry in glycobiology. Mass spectrometry is a technique to detect the molecular weights (masses) of various pieces in a large molecule. The resulting "spectrum" is a display of the relative frequencies of these molecular weights. For the purposes of this project, we'll ignore the relative frequencies, and just pay attention to which molecular weights show up at all. (We'll also ignore the substantial science that goes into producing these spectra from actual samples of biological and chemical material.)

If we could only get the molecular weight of the whole molecule, it wouldn't help as much in finding the makeup of the molecule. So the mass spectrometry also involves fragmentation, where the molecule is fragmented into smaller molecules. There are general rules describing how and where certain types of molecule will fragment, but for our purposes, it is enough to know that the molecule does not always fragment all the way into its constituent atoms. Furthermore, the different molecules that make up a sample will fragment differently. As a result, what the spectrum shows is the molecular weight of various fragments (not necessarily all of them) you could get from the molecule. The goal is to reconstruct the original molecule from the molecular weight of the fragments.

We'll restrict our attention further to a class of molecules originally described to me as [true story] glycon "trains". The glycon refers to a string of glycon monomers. There are 10 possible such monomers. They are referred to as trains because they are arranged in a straight line, like a train, with each car of the train holding one glycon monomer. Each monomer may appear more than once in a train. The glycon train is also followed by another train of amino acids, but we'll ignore that, and just focus on the glycon part. In theory, the trains could be of unlimited length, but one database shows most of the lengths being at most 8 cars long, so assume all glycon trains have length at most 8 cars. Note that the molecular weight of a glycon train is the sum of the molecular weights of the glycon monomers that compose it (including repeats).

For every question in this project, find the actual numerical value being asked for, as well as the computation that leads to it. For instance, if the question is "how many different permutations of 3 letters are there?", you would answer "$3! = 6$". Also explain why your calculation is correct, so that a classmate would understand your reasoning. (In our example, that would mean explaining why 3! is the correct calculation, not why 3! equals 6.)

1. How many different glycon trains, whose lengths are exactly 8 cars, are possible?

2. Without the fragmentation, we would only be able to find the total molecular weight of each train. What is the maximum number of possible different **molecular weights** of glycon trains whose length is exactly 8 cars? [There is a subtle point here: If two monomers have the same molecular weight, or if one monomer has the same molecular weight as two monomers put together, you might get an unexpected repeat. Ignore this possibility for now, and assume that the molecular weights of these monomers are

known with enough precision that such repeats are unlikely.] Recall that the molecular weight of a glycon train is the sum of the monomer cars that compose it (including repeats).

3. In practice, we won't know in advance how long the glycon train is. What is the maximum number of possible different **molecular weights** of glycon trains whose length is at most 8 cars?

4. In fact, two of the glycons are similar enough that their molecular weights are almost identical. Answer questions 2 and 3 taking this into account.

5. Compare the numerical values of the answers to questions 1 and 2. Use this to explain why fragmentation is helpful in reconstructing molecules.

   Reconstructing a molecule from is spectrum is not easy (see the final question, where you have to do it yourself by hand), and so [true story] some biologists were interested in developing an algorithm to automate the process. Good algorithms are hard, so instead they were recommended to create a database of all the possible trains, and, for each one, compute all the possible fragment weights. Then instead of having to work an algorithm each time, they could just look it up in this database. This would take extra time initially (to compile the database), but make things go faster afterwards (searching in a database is a well-known procedure, and there are many ways to do it efficiently). But then it's important to know how big that database will be.

6. You've already computed (in a previous question) how many entries in the database there are (the number of possible glycon trains). Now let's figure out how much data each molecule needs to store. We'll start with an example. If a train is ABC, then the possible fragments are: A, B, C, AB, BC, ABC, for a total of 6 possible fragments, and hence 6 pieces of data for a train of length 3. Repeat this for trains of length 4, 5, 6, 7, and 8. State a general formula for the number of possible fragments of a train of length exactly n; your answer should **not** include a summation. Justify your answer.

7. (Bonus question.) Try your hand at reconstructing a train from its spectrum, by hand. (Alternative: Write a computer program to do it; but then you have to describe in some detail how you wrote your program.) Here are the molecular weights (rounded to the nearest integer for simplicity of the assignment) of the 10 monomers:

monomer	weight
HEXOSE	162
HexNAc	203
DEOXYHEXOSE	146
NeuAc	291
NeuGc	307

monomer	weight
PENTOSE	132
SO3H	80
PO3H	80
KDN	250
HexA	176

   Here are the molecular weights of some (not all!) of the fragments of a train with **at most** 6 cars:

$$176, 291, 426, 541, 717, 791, 967$$

   Determine the train. Explain in detail what you did to figure out your answer.