Quiz 2 (100 pts)

Cpt S 223 – Fall 2013

Due 11-13-2013

You may work with others on this assignment including working together during code implementation. But you may not share the actual code.

Create a program called "**peptides**" that will create a k-d tree from a peptide database (comma separated value formatted file) based on mass and normalized elution time (NET). The file will contain a list of peptide sequence strings and NET. Your program should compute the mass of a peptide based on the amino acid characters. You must create a hash table, including the hash function, to do so. You **may not use a map** or data structure from the standard template library. Use the monoisotopic mass from the table at the end of this document.

The program should also read a file containing list of 2-D points of mass and NET, called the *observed* list. Each item in this file will also have a number called the ID (for index).

Then for each item in the observed list, perform a nearest neighbor search returning the closest peptide sequence and elution time. Your program should print (in **CSV** format) to standard output the list of all found peptides. Your distance function should be a Euclidean distance based on mass and NET. (See your notes from class.)

Example data files are given on the course Angel site (lms.wsu.edu).

Your program will be run as following:

```
> peptides peptideDatabase.csv observedList.csv
```

Failed inputs should say "usage: peptides databaseFile observedListFile"

Example Output (only showing one hit)

```
Observed ID, Peptide, NET, Mass, Observed NET, Observed Mass
0, AGGVGGK, 0.1494728, 523.42, .1495, 523.426
```

Example Peptide Database

```
Peptide, NET
AGGVGGK,0.149476528
AGMFGK,0.148264542
APTAAAK,0.147068828
SSPGGVK,0.149400458
AHYGGF,0.203524396
VFGGGTK,0.199178353
MVPAVR,0.166774005
ADGSPVK,0.084761672
```

Example Observed List

ID, NET, Mass
0, 0.149, 523.42
0, 0.447, 825.42
0, 0.346, 573.42

Amino Acid Mass Table

1-letter code	3-letter code	Chemical formula	Monoisotopic
A	Ala	C_3H_5ON	71.03711
R	Arg	$C_6H_{12}ON_4$	156.10111
N	Asn	$C_4H_6O_2N_2$	114.04293
D	Asp	$C_4H_5O_3N$	115.02694
C	Cys	C_3H_5ONS	103.00919
E	Glu	$C_5H_7O_3N$	129.04259
Q	Gln	$C_5H_8O_2N_2$	128.05858
G	Gly	C_2H_3ON	57.02146
H	His	$C_6H_7ON_3$	137.05891
I	Ile	$C_6H_{11}ON$	113.08406
L	Leu	$C_6H_{11}ON$	113.08406
K	Lys	$C_6H_{12}ON_2$	128.09496
M	Met	C_5H_9ONS	131.04049
F	Phe	C_9H_9ON	147.06841
P	Pro	C_5H_7ON	97.05276
S	Ser	$C_3H_5O_2N$	87.03203
T	Thr	$C_4H_7O_2N$	101.04768
W	Trp	$C_{11}H_{10}ON_2$	186.07931
Y	Tyr	$C_9H_9O_2N$	163.06333
V	Val	C_5H_9ON	99.06841