Molecular Weight/Isoelectric Point

1. Purpose

This document describes the process by which protein parameters are calculated, involving isoelectric point (pI) and molecular weight (MW). Values are calculated by a BioPerl related script, in a precompute process using GenBank files, and are presented to the user in VIPR.

2. Method Description

The method calculates physical parameters of proteins given only their sequence. All proteomics studies that separate proteins on the basis of their charge and size generate quantitative values. By calculating these values in a precompute, users can search for molecules within a given range and size, to help determine what they have been studying and purifying.

3. Input Data Preparation

Input to the method is a protein fasta (.faa) amino acid text file, such as the protein sequence:

```
>PB1_F2_frag
MEQEQDTPWTQSTGHINIQKRGNGQQTQRLEHLNSTRLTGHCLRTMSQVDMHKQTVSWQ
```

4. Output Data, Processing and Display

Input is fasta files; Usage is:

$ perl calc_MW_IsoelectricPt.pl <inputfile> <outputfile>

e.g. $ perl get_pI_MW.pl FSC198.faa output_FSC198.txt

Output is a table in delimited format:

```
gi          iep   mw_lower     mw_upper     diff
110669658   8.71  57676.4      57676.4      t
110669659   5.11  41674.3      41674.3      t
110669660   6.68  30305.3      30951.1      f
```

The lower and upper MW difference occurs ('f') if a sequencing abnormality reveals an amino acid residue whose identity cannot be determined precisely ('X'). For example, a string of four X's can result in the calculation of MW that differ in lower and upper range by 645 g/mol (i.e. all glycine vs. all tryptophan).

Results such as for the NA fragment above are presented back to the user as shown:

Isoelectric pt/Molecular Weight (SOP)

Isoelectric pt	Molecular Weight	Evidence Code
6.9	255972.8	RCA

The precision of the method is pI is +/- 0.01 charge unit. MW precision is to the closest 0.01 g/mol.

5. Reference

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C.,… Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome research*, *12*(10), 1611–1618.

DOI: http://www.genome.org/cgi/*doi*/*10.1101*/*gr.361602*

##

```
#!/usr/bin/perl

use warnings;
use strict;

use Bio::SeqIO;
use Bio::Tools::SeqStats;
use Bio::Tools::pICalculator;

#
# Chris Larsen, Anuj Bhatia, and Tom Briggs
# UT Southwestern Medical Center and Vecna Technologies
#
# January 21, 2008
#
# Output format:
# gi <tab> iep <tab> mw_lower <tab> mw_upper <tab> do_lower_and_upper_differ
#

if (scalar(@ARGV) != 2) {
  die("Usage: $0 INPUT_FILE OUTPUT_FILE\n");
}
my ($INFILE, $OUTFILE) = ($ARGV[0], $ARGV[1]);

open (OUTFILE, "> $OUTFILE");

print OUTFILE "gi\tiep\tmw_lower\tmw_upper\tdiff\n";

# Fasta File Object
my $entries = Bio::SeqIO->new('-file'   => $INFILE,
                              '-format' => 'Fasta');

# pI Calculator Object
my $calc = Bio::Tools::pICalculator->new(-places => 2,
                                         -pKset => "EMBOSS");

# iterate through each sequence
while (my $seq = $entries->next_seq()){
        #get gi
        my $seqid = $seq->id;
        my @gi = split(/\|/, $seqid);

  my $strippedDefline = 1;
  if (scalar(@gi) > 1) {
    $strippedDefline = 0;
  }

        #get mw (two var array)
        my $seq_stats = Bio::Tools::SeqStats->new($seq);
        my $weight = $seq_stats->get_mol_wt();

        #get iep
        $calc->seq($seq);
        my $iep = $calc->iep;

  my $differ = ($$weight[0] != $$weight[1]) ? 't' : 'f';
  my $giNumber = $strippedDefline ? $gi[0] : $gi[1];
        print OUTFILE $giNumber . "\t" . $iep . "\t" . $$weight[0] . "\t" . $$weight[1] . "\t" . $differ . "\n";
}
```