

Protein Prediction

Molecular Weight, Isoelectric Point,
Localization

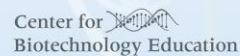


Center for  Biotechnology Education

Hi, I'm Bob Lessick at Johns Hopkins. In this lecture we will examine how to input a protein sequence and output three types of predictions. The isoelectric point and molecular weight are very simple to calculate. We will also examine prediction to where a protein goes within a cell after it is translated or synthesized.

Objectives

- Define molecular weight, isoelectric point
- Predict molecular weight, isoelectric point
- List cellular locations for which proteins may be targeted
- Predict subcellular location of a protein



We want to first look at exactly what is meant by isoelectric point and molecular weight for a protein.

And then go ahead and perform those calculations.

We'll shift to targeting or localization. Where in a cell can proteins go? Let's look at some possibilities.

Finally, let's try to predict where proteins go based on sequence. Many proteins have short stretches of sequences that serve as tags that molecules involved in transport might recognize.

Molecular Weight

- Measured in daltons (Da)
 - Same as unified atomic mass unit
 - Equivalent to grams/mol
 - But it is mass of one molecule, not a mole
- Technically measure of mass, not weight
 - Like grams, kilograms (mass units)
 - Not like ounces, pounds (weight units)
- Still common to call it molecular weight



Center for  Biotechnology Education

Let's start with molecular weight. Molecular weight is measured in daltons. The word "daltons" is not capitalized, but the abbreviation "Da" is capitalized. For those of you familiar with chemistry, a dalton is synonymous with a unified atomic mass unit, which is close to the weight of a single proton or single neutron. The unit is named for John Dalton, an English chemist/physicist known for atomic theory.

The molecular weight in daltons is equivalent to grams per mole, except instead of measuring the mass of a mole of molecules, the dalton refers to the mass of a single molecule.

I should not that molecular weight is technically an incorrect term, since the dalton is a unit of mass, not a unit of weight. The same is true for grams and kilograms, both units of mass.

Weight is measured in the English system with ounces and pounds, primarily in the United States. What's the difference? Weight is affected by gravity and mass is not.

Although "molecular mass" is a more correct term, "molecular weight" is generally accepted and more commonly used.

Isoelectric Point (pI)

- Defined as pH at which entire protein has no net charge
 - If $pI < 7$, protein is acidic
 - If $pI > 7$, protein is basic
- Acidic proteins are negatively charged at neutral pH
- Basic proteins are positively charged at neutral pH



Center for  Biotechnology Education

Isoelectric point is usually abbreviated as pI , a lower case p and a capital I . That looks a lot like pH and there is a reason for that. The isoelectric point is the pH at which the protein has no net charge—that is, all of the positively charged atoms and equally balanced by the same quantity of negatively charged atoms.

So, if the isoelectric point is less than 7—remember, 7 is neutral pH , then that protein is acidic. If the pI is greater than 7, then that is a basic protein.

To put it another way, acidic proteins have a net negative charge at neutral pH .

Meanwhile, basic proteins have a net positive charge at pH 7.

Physical Properties Prediction

- Not difficult to predict molecular weight, isoelectric point
- Compute pI/Mw tool
 - http://web.expasy.org/compute_pi/
 - Enter protein, output pI, MW
- Both are important measures, also useful in protein separation/purification



Center for  Biotechnology Education

These calculations are among the easiest in sequence-based bioinformatics. For molecular weight, each amino acid has its own molecular weight. Add them up and subtract the weight of water for each peptide bond in the chain. That's because an H₂O molecule is released upon peptide bond formation. Isoelectric point is a little more tricky, but not that difficult.

There is a tool called the Compute pI/MW tool—pI for isoelectric point and MW for molecular weight.

You input the sequence and the output contains both the isoelectric point and molecular weight,

I should add that these are important measures—there's a reason to do this. In addition to knowing those important parameters, both measures are useful for developing strategies to separate and purify proteins in the lab.

pI/Mw Output

ExPASy
Bioinformatics Resource Portal

Compute pI/Mw

HBB_HUMAN (P68871)

Hemoglobin subunit beta (Beta-globin) (Hemoglobin beta chain) [Contains: LVV-hemorphin-7]
Homo sapiens (Human).

The parameters have been computed for the following feature:

FT CHAIN 2 147 Hemoglobin subunit beta.

Considered sequence fragment:

```
1 11 21 31 41 51
| | | | |
1 VHLTPEEKS AVTALNGKVN VDEVGGEALG RLLVVPWTQ RFFESFGDLS TPDVAMGNFK 60
61 VKANGKKVLG AFSDDLALND NLKGTATLS ELMCKKLMD PENFELGGV LVCVLARHG 120
121 KEPTFFVQAA YQGVVAGVAN ALAKKYH
```

> Fasta

Molecular weight (average): 15867.22

Theoretical pI: 6.81



Center for Biotechnology Education

Here is the output for human beta globin.

The calculated molecular weight is 15,867.22 daltons. Most would more commonly round that off to 15.9 kilodaltons or even 16 kilodaltons, meaning 16,000 daltons.

The isoelectric point is 6.81, slightly acidic, but very close to neutral. At neutral pH, this protein would have a very slight net negative charge.

Protein Localization

- Where does a protein go after translation?
- Secretory pathway
 - Includes ER, Golgi, plasma membrane
- Cytoplasm
- Other organelles
 - Nucleus, mitochondrion, peroxisome, lysosome
 - Chloroplast (plants only)



Center for  Biotechnology Education

Now let's transition to protein localization. That's a more difficult prediction, since the signals that are found in proteins are short. Let's examine a few places proteins could go in a eukaryotic cell after translation.

The secretory pathway is a mechanism for transporting molecules out side of the cell. Sometimes cells need to release proteins for a variety of possible reasons. Try to think of a few reasons that might be useful. The wikipedia entry on "secretory pathway" explains it very well. Proteins that utilize the secretory pathway are deposited into the cell's endoplasmic reticulum as they are being translated. From there, they travel to the Golgi apparatus and then via secretory vesicles to the plasma membrane. Some proteins stay in the plasma membrane and some are released to the outside.

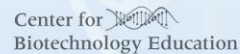
However, other proteins are destined to stay in the cell. Some function primarily in the cell's cytoplasm. The cytoplasm is an aqueous region that is basically everything but the organelles of a cell.

Here are some of those organelles. The nucleus hosts the genomic DNA, the mitochondria are the energy centers, the peroxisome is where fats are broken down and the lysosome is a digestive center. These are all places that require specific enzymes and proteins in order to function properly, so getting those proteins to those locations is very important.

And in plant or photosynthetic cells only, a possible destination is the chloroplast.

Localization Prediction

- General predictors
 - Input protein sequence, output location
 - Some list several possibilities
- Specialized predictors
 - MITOPRED predicts mitochondrial proteins
 - Output is likelihood protein localizes to the mitochondrion



There are many prediction programs available on the web. Many are general predictors. Put in an amino acid sequence and the program predicts its destination. Some list possibilities from most likely to less likely.

Some programs are specific for a single organelle. For instance, a program called MITOPRED, well, you might not be surprised that that predicts whether or not a protein goes to the mitochondrion. It's a yes/no question, and the output of such programs is usually pretty simple and expressed as a percentage, for example the prediction might say that a protein is 78% likely to be mitochondrial based on its sequence.

Some localization servers

Server	URL
ProLoc-GO	http://iclab.life.nctu.edu.tw/prolocgo/index.php
SignalP	http://www.cbs.dtu.dk/services/SignalP/
TargetP	http://www.cbs.dtu.dk/services/TargetP/
Predotar	http://urgi.versailles.inra.fr/predotar/predotar.html
PSORT	http://www.psort.org/
CELLO	http://cello.life.nctu.edu.tw/
MultiLoc2	http://abi.inf.uni-tuebingen.de/Services/MultiLoc2
Euk-mPloc	http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/

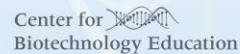


Center for  Biotechnology Education

Here are just some of those web servers and there are quite a few out there.

Exercise: Three servers

- Run the yeast Pex8p protein on three servers
 - Accession: NP_011591
 - Retrieve from NCBI in FASTA format
- Three servers
 - Predotar: <https://urgi.versailles.inra.fr/predotar/predotar.html>
 - CELLO: <http://cello.life.nctu.edu.tw/>
 - LocTree: <https://www.rostlab.org/services/loctree3/>



Let's do an exercise. First, find the yeast pex8p protein at NCBI. The accession number is NP underscore 011591 and let's get that protein sequence in FASTA format.

Then go to these servers and enter the protein sequence. Make sure it is set for eukaryotic prediction if that is an option. And we'll see what we get from Predotar, Cello, and LocTree.

CELLO



MBC
About CELLO

Molecular Bioinformatics Center
National Chiao Tung University

CELLO v.2.5: subCELLular LOCALization predictor

1. Choose Eukaryotes

ORGANISMS	SEQUENCES
☒ Gram negative	☐ DNA
☐ Gram positive	☒ Protein
☒ Eukaryotes	

2. Delete Example Protein!

Paste the query sequences in FASTA format below

>1086005 (Genbank) Outer membrane/Extracellular (Autotransporter) (major ring-forming surface protein precursors)
MTKISDVEQNTLKRKSSLRKFTFPLIATLAFSLASSFVDAAGAGQAPVNAEGITVTYVQAKTATVSGSD
GNATFTTHSANTVNGTADPAVTAHIEVNIANTVNTVQKPAQAGQDAEAKFVNLTFPGGLASSTAKTFTL
HLGAGADALGRIHLGAGNATLNTHTWSLADGPPVIVHCAITFDFTSSGAGTGNIVTGNMTSOTGNIIT
DQFMCI PRHSELHSGATVQAPATVLSHETTSISINVTFRKSGWCCIAGNATGSLGHPVTFKQVNTISG
VIASTGGVNTLNFONATVDATGGNTLIIQSSGITFRTHGVNSFTLTHATITFAAAGGDFAKATVQGNKSAVQ
GVNTLHFFFAKLESTPAKAPAFANITATGAGNIVFPQGLVNAULTSLDGGINTLVMTNIVNFPILLTQV
VTTFPRAASNTLIFQNTSTSGHAGNTLQVAVVNPFAHGSQATFSTVPAFHLMLKQAGSLRAGADA
HARANDQASQIQVILKFNHGSANATGILATNGSTATVLRHTTLAULFPQAGQVNTVGGNHSANTVLEAPVDA
SATITVGGTLOGHNTSHVNVNSQTSFVHLIFANADNRGTFLLNGATGSLVSDAFGGQFRDLGAGKVLQTVQSG
TQSSLCGGNTLQAGNGLUSSPFAFICALIATVSNAEFATQSLPVPZLVPSGHSISFGSGNSFVNTILEG
VAVGSITALTNGANTTGGNTSGIVSLKESPSVLEHSEHGHSTHNGALGSAHLLQSSGAGTGGVALLNLI
IASGNGHNGNGAAVFGSSVSFGKQNDYTALQNTVILATGGSSNVFRTWFLNLLVQANSNHTTASDQQA
SLOGGNALFRVTVADANQNGAGGGRNATLNGHSFHSGSLVGLYSRVTVTQCFCDISHPFQMSIYPR
YBSSGTERADVAVLVNDEGQASVNTTTSPTDTPKMLTAVHAGVETENHAGNNTFAQLGSLFQSG
TQSTSSGHTGSSQQAHDQVITYTFLSQAVNTSEANGLATATALASHYTYLANIDSLHNMKELSRFSSGPRGM

Or upload from file: to file chosen



JOHNS HOPKINS
UNIVERSITY

Center for  **Biotechnology Education**

This is cello—I believe it is pronounced cello because it is a cell prediction, but it’s a Taiwanese-based program and web server so I’m not positive, maybe it is “chell—oh.” There are two very important things to do before running the program.

First, switch to Eukaryotes—the default is set for Gram-negative bacteria. Gram-negative bacteria have no organelles, so that prediction is very different—mainly secretory or retained in the cell.

Then, note that there is a protein sequence is already there. That is an example sequence that is there by default so that users can see how it works. Be sure to delete that sequence before you enter your sequence!

CELLO Results

CELLO RESULTS		
SeqID: gi 6321514 ref NP_011591.1 Pex8p [Saccharomyces cerevisiae S288c]		
Analysis Report:		
SVM	LOCALIZATION	RELIABILITY
Amino Acid Comp.	Peroxisomal	0.520
N-peptide Comp.	Peroxisomal	0.922
Partitioned seq. Comp.	PlasmaMembrane	0.511
Physico-chemical Comp.	Peroxisomal	0.810
Neighboring seq. Comp.	PlasmaMembrane	0.664
CELLO Prediction:		
	Peroxisomal	2.934 *
	PlasmaMembrane	1.590
	Nuclear	0.098
	Mitochondrial	0.093
	Extracellular	0.076
	Cytoplasmic	0.075
	Golgi	0.038
	ER	0.033
	Lysosomal	0.026
	Chloroplast	0.021
	Cytoskeletal	0.009
	Vacuole	0.008



JOHNS HOPKINS
UNIVERSITY

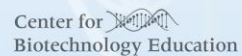
Center for  Biotechnology Education

Cello examines five different parameters to make its prediction. The amino acid composition, the N-terminus and the physico-chemical properties suggest that this protein goes to the peroxisome. Based on examination of the partitioned sequence and neighboring sequences, it appears to be a plasma membrane protein.

Cello is an example of a jury prediction strategy. The consensus, based on the results of the five different predictions suggest strongly that this protein goes to the peroxisome. Do a web search for Pex proteins—that will give you a good idea where this protein actually goes.

Summary

- The compute pI/Mw tool will accurately predict:
 - Isoelectric point (pI)
 - Molecular weight (in daltons)
- Many localization servers available
 - Detect signals in sequence for localization
 - Some algorithms more suited to certain organelles



So to wrap it up, we know that we can enter a protein sequence and get a reliable calculation of the isoelectric point and molecular weight.

And that many localization predictors are out there. Some are general and some are specific for certain organelles. Give some of them a try. Good luck!