annot8r_physprop - A tool to predict physical properties of peptides

Ralf Schmid and Mark Blaxter

The Institute of Cell, Animal and Population Biology
University of Edinburgh

for the Natural Environment Research Council
Environmental Genomics Thematic Programme Data Centre

Ashworth Laboratories, King's Buildings, Edinburgh, EH9 3JT, UK
p +44 131 650 6760 +44 131 650 6761 f +44 131 650 7489

nematode.bioinf@ed.ac.uk

http://www.nematodes.org/
http://www.earthworms.org/
http://www.tardigrades.org/

-Overview

annot8r_physprop is part of the annot8r suite - a series of scripts to provide annotation for peptide sequences derived from EST datasets and to give the user an idea about quality and reliability of the annotation. annot8r_physprop calculates molecular weight (Mw), isoelectric point (pI), charges at pH5, pH7 and pH9 and gives the associated information needed to estimate the reliability of these calculations. Most of the calculations are done using Bioperl modules.

Calculation of physical properties may appear to be an easy task, nevertheless a few words of warning are in order. The main problem is the assumption that we are dealing with complete sequences. Using partial sequences has the inescapable consequence that the numbers calculated for Mw are too small. In a future version we will try to adress this problem by using sequence overlap with known protein domains to estimate the "completeness" of the sequences. Undefined residues (X) are taken into account in calculating an upper and lower limit for the predicted molecular weight.

The calculation of pI and charges at different pHs depend strongly on the number of charged residues. This again rises the issue of partial sequences. The actual values are predicted assuming independent residues with standard pKa-values as defined in EMBOSS. Different sets of pKa-values can be used when changing the script accordingly. (see http://search.cpan.org/dist/bioperl/Bio/Tools/pICalculator.pm for details). Using standard values for pKa values does not account for the variability of pKa for the same type of amino acid in a different protein environment. Predicting these effects accurately requires a high resolution protein structure and sophisticated electrostatic calculations - these are not feasible with medium or high-throughput approaches so far. The possible oxidation of cysteins to disulfide bonds is not accounted for. Furthermore, the calculations do not consider post-translational modifications such as phosphorylation and glycosylation, or the impact of charged cofactors. However, we have a measure to estimate the accuracy of the calculations: the number of charged residues in the sequence. As a rule of thumb the robustness of the calculations will increase with the number of charged residues.

The only requirement to run annot8r_physprop is a file containing your peptide sequences of interest and a postgresql database if you want to integrate the annotation into a PartiGene database. The script is started by typing "annot8r_physprop.pl" (assuming you have saved it as an executable file in a directory in your path).

There are two options :

1. Calculate physical properties

After selecting option “1” the user is asked for the location of the input sequence file. This file should contain the sequences of interest in FASTA format. For these sequences annot8r_physprop calculates a series of physical properties and creates a TAB-delimited file listing the following for each sequence: the sequence identifier, the length of the sequence, the number of undefined residues, the minimal and maximal Mw considering undefined residues, the number of positively and negatively charged residues (only R, K, D and E are considered), the predicted pI and predicted charges at pH5, pH7 and pH9. This file can either be used to access the information on physical properties directly or as input file for a PartiGene database (see option 2).

2. Databasing

Option 2 takes the data calculated in step one and stores them in a PartiGene database in a table called “physprop”. If a PartiGene database is defined in the .partigene.conf file the user is asked whether he or she wants to use it. There are several alternative options to using the pre-defined PartiGene database. A new database can be created, a table “physprop” can be added to an already existing database or an existing table “physprop” can be upgraded.

A quick tutorial

Before starting with real data we recommend running annot8r_physprop with the test data provided within the annot8r_physprop module. All you need to do is to copy the file test_seq.fsa (a set of test sequences for which physical properties will be calculated) into the directory from where you are starting the script. Input you have to type is printed in *italics*.

1. Copy test_seq.fsa into your project directory.
2. Start annot8r_physprop.pl.
3. Select the “**Calculate physical properties**” option.
4. Enter *test_seq.fsa* as your relative location of the inputfile.
5. Now annot8r_physprop.pl will run all calculations and store the output in the file “test_seq.fsa_phys”.

=> so far you have created a file holding sequence identifiers and the calculated physical properties for the respective sequences. This file can either be used to access the results directly or as an input file for a PartiGene

database (see following steps)

6. Select the "**Databasing**" option.
7. If you have used PartiGene previously, you are asked whether you want to use the PartiGene database as defined in your PartiGene configuration file for the physical properties. For this test run answer "n".
8. Enter *phystest* as database to be created or used.
9. Reply "*y*" to the query to create "phystest".

=> now the results of the calculation are stored in the postgresql database "phystest".

10. Select the "**Quit**" option to quit annot8r_physprop.
11. type *psql phystest* to open your results postgresql database.
12. type *select * from physprop;* (don't forget the semicolon) to view your results.
13. type *\q* to exit postgresql.
14. type *dropdb phystest* to remove the test database.

For further questions, comments, problems or bug reports please contact nematode.bioinf@ed.ac.uk

If you are working in an EG-Awardee lab and have questions or problems please contact the helpdesk helpdesk@envgen.nox.ac.uk