Prediction of protein and peptide isoelectric point based on sequence and structure features

Łukasz P. Kozłowski

Warsaw, 13.12.2018
Outline of the project

1.1 IPC version 2.0
- sequence (1st Student)
- structure (2nd Student)

1.2 Proteome-pI version 2.0

Timeline: ~24 months

Output:
- 2-3 publications expected (hopefully only 2)
- 2 Master’s theses
**Introduction**

**Isoelectric point**

*No Net Charge on Protein*

$p_I = pH$

Low pH  

1. Increasing pH  
2. pH $>$ $p_I$  
Protein Has Net Positive Charge

3. High pH  
4. pH $<$ $p_I$  
Protein Has Net Negative Charge

---

www.creative-proteomics.com
Introduction
Introduction

acidic media
low pH

neutral form

basic media
high pH

PDB: 4DKL
Isoelectric point importance - isolating the proteins

1. Separation of proteins by pl value

2. Soaking the gel in SDS solution and fitting it on an SDS PA gel

3. Separating the proteins by molecular mass with SDS PAGE

Approximately 1000 E. coli proteins on a single 2D gel

Two-dimensional electrophoresis (2D-PAGE)
Isoelectric point importance - isolating the proteins

Ovitrap mass spec machine

Isoelectric Point-Based Fractionation with Liquid Chromatography and Mass Spectrometry (2D-LC-MS)
Isoelectric Point-Based Fractionation
with
Liquid Chromatography and Mass Spectrometry
(2D-LC-MS)
Isoelectric point importance – X-ray crystallography

Crystal → Diffraction pattern → Electron density map → Protein model

Source: www.creative-biostructure.com
**Isoelectric point importance – X-ray crystallography**

Protein crystals do not grow in the buffer with pH around isoelectric point.

**Protein isoelectric point** as a predictor for increased **crystallization** screening efficiency

KA Kantardjieff, B Rupp - Bioinformatics, 2004 - academic.oup.com

Source: www.creative-biostructure.com
Current state of the art in Isoelectric Point calculation

Henderson–Hasselbalch equation

\[
HA + H_2O \overset{\text{equilibrium}}{\Leftrightarrow} A^- + H_3O^+
\]

\[
HA \overset{\text{equilibrium}}{\Leftrightarrow} A^- + H^+; K_a = \frac{[A^-][H^+]}{[HA]}
\]

\[
pK_a = -\log_{10}(K_a)
\]

**Acid dissociation constant (Ka)**

the strength of an acid in solution (the equilibrium constant for a chemical reaction known as dissociation in the context of acid–base reactions)
Henderson–Hasselbalch equation

\[ \text{negative charged} \]
\[ \sum_{i=1}^{n} \frac{-1}{1 + 10^{pK_i - p\text{H}}} \]

\[ \text{positively charged} \]
\[ \sum_{i=1}^{n} \frac{1}{1 + 10^{p\text{H} - pK_i}} \]

\[ \text{NQ} = QN_1 + QN_2 + QN_3 + QN_4 + QN_5 + QP_1 + QP_2 + QP_3 + QP_4 \]
Henderson–Hasselbalch equation

\[
\sum_{i=1}^{n} \frac{-1}{1 + 10^{pK_a - pH}} = \sum_{i=1}^{n} \frac{1}{1 + 10^{pH - pK_a}}
\]

\[
NQ = QN_1 + QN_2 + QN_3 + QN_4 + QN_5 + QP_1 + QP_2 + QP_3 + QP_4
\]

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>NH₂</th>
<th>COOH</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>H</th>
<th>K</th>
<th>R</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMBOSS</td>
<td>8.6</td>
<td>3.6</td>
<td>8.5</td>
<td>3.9</td>
<td>4.1</td>
<td>6.5</td>
<td>10.8</td>
<td>12.5</td>
<td>10.1</td>
</tr>
<tr>
<td>DTASelect</td>
<td>8.0</td>
<td>3.1</td>
<td>8.5</td>
<td>4.4</td>
<td>4.4</td>
<td>6.5</td>
<td>10.0</td>
<td>12.0</td>
<td>10.0</td>
</tr>
<tr>
<td>Solomon</td>
<td>9.6</td>
<td>2.4</td>
<td>8.3</td>
<td>3.9</td>
<td>4.3</td>
<td>6.0</td>
<td>10.5</td>
<td>12.5</td>
<td>10.1</td>
</tr>
<tr>
<td>Sillero</td>
<td>8.2</td>
<td>3.2</td>
<td>9.0</td>
<td>4.0</td>
<td>4.5</td>
<td>6.4</td>
<td>10.4</td>
<td>12.0</td>
<td>10.0</td>
</tr>
<tr>
<td>Rodwell</td>
<td>8.0</td>
<td>3.1</td>
<td>8.33</td>
<td>3.68</td>
<td>4.25</td>
<td>6.0</td>
<td>11.5</td>
<td>11.5</td>
<td>10.07</td>
</tr>
<tr>
<td>Patrickios</td>
<td>11.2</td>
<td>4.2</td>
<td>-</td>
<td>4.2</td>
<td>4.2</td>
<td>-</td>
<td>11.2</td>
<td>11.2</td>
<td>-</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>8.2</td>
<td>3.65</td>
<td>8.18</td>
<td>3.9</td>
<td>4.07</td>
<td>6.04</td>
<td>10.54</td>
<td>12.48</td>
<td>10.46</td>
</tr>
<tr>
<td>Lehninger</td>
<td>9.69</td>
<td>2.34</td>
<td>8.33</td>
<td>3.86</td>
<td>4.25</td>
<td>6.0</td>
<td>10.5</td>
<td>12.4</td>
<td>10.0</td>
</tr>
<tr>
<td>Grimsley¹</td>
<td>7.7</td>
<td>3.3</td>
<td>6.8</td>
<td>3.5</td>
<td>4.2</td>
<td>6.6</td>
<td>10.5</td>
<td>12.04</td>
<td>10.3</td>
</tr>
<tr>
<td>Toseland</td>
<td>8.71</td>
<td>3.19</td>
<td>6.87</td>
<td>3.6</td>
<td>4.29</td>
<td>6.33</td>
<td>10.45</td>
<td>12.0</td>
<td>9.61</td>
</tr>
<tr>
<td>Thurlkill</td>
<td>8.0</td>
<td>3.67</td>
<td>8.55</td>
<td>3.67</td>
<td>4.25</td>
<td>6.54</td>
<td>10.4</td>
<td>12.0</td>
<td>9.84</td>
</tr>
<tr>
<td>Nozaki_Tanford</td>
<td>7.5</td>
<td>3.8</td>
<td>9.5</td>
<td>4.0</td>
<td>4.4</td>
<td>6.3</td>
<td>10.4</td>
<td>12.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Dawson²</td>
<td>8.2</td>
<td>3.2</td>
<td>8.3</td>
<td>3.9</td>
<td>4.3</td>
<td>6</td>
<td>10.5</td>
<td>12.0</td>
<td>10.0</td>
</tr>
<tr>
<td>Bjellqvist³</td>
<td>7.5</td>
<td>3.55</td>
<td>9.0</td>
<td>4.05</td>
<td>4.45</td>
<td>5.98</td>
<td>10.0</td>
<td>12.0</td>
<td>10.0</td>
</tr>
</tbody>
</table>
Henderson–Hasselbalch equation

\[ NQ = QN_1 + QN_2 + QN_3 + QN_4 + QN_5 + QP_1 + QP_2 + QP_3 + QP_4 \]
Optimization

Brute force attack

Checking all possible combinations is not very tractable as even for 9 variables (charged amino acid pKa) in range of pH of 3 (±1.5 pH of average for given amino acid pKa) with 0.01 precision gives 1.9683 × 10^{22} possibilities. Far too many to compute.

\[
\sum_{i=1}^{n} \frac{-1}{1 + 10^{pK_n - pH}} \quad \sum_{i=1}^{n} \frac{1}{1 + 10^{pH - pK_p}}
\]

negative charged \quad \text{positively charged}

\small
NQ=QN1+QN2+QN3+QN4+QN5+QP1+QP2+QP3+QP4

Basinhopping optimization using truncated Newton algorithm

This produces suboptimal results in more reasonable time with less than few dozens of iterations with pKa optimized with high precision.

In the nutshell, the basinhopping algorithm is iterative search procedure with each cycle composed of the following features:

• random perturbation of the coordinates
• local minimization
• accept or reject the new coordinates based on the minimized function value of the Metropolis criterion of standard Monte Carlo algorithm

As an initial seed previously published pKa values were used. To limit search space truncated Newton algorithm was used with 2 pH units bounds for pKa variables (e.g. if starting point for Cys pKa was 8.5 the solution was allowed in the interval [6.5, 10.5]).

For more details how those algorithms works read Wales & Doye, 1997 (doi: 10.1021/jp970984n)
Current state of the art in Isoelectric Point calculation

<table>
<thead>
<tr>
<th>Method</th>
<th>Protein dataset</th>
<th>Peptide dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSD</td>
<td>%</td>
</tr>
<tr>
<td>IPC_protein</td>
<td>0.874</td>
<td>0</td>
</tr>
<tr>
<td>Toseland</td>
<td>0.934</td>
<td>14.9</td>
</tr>
<tr>
<td>Bjellqvist</td>
<td>0.944</td>
<td>17.7</td>
</tr>
<tr>
<td>Dawson</td>
<td>0.945</td>
<td>17.8</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>0.955</td>
<td>20.5</td>
</tr>
<tr>
<td>Rodwell</td>
<td>0.963</td>
<td>22.8</td>
</tr>
<tr>
<td>Grimsley</td>
<td>0.968</td>
<td>24.2</td>
</tr>
<tr>
<td>Solomon</td>
<td>0.970</td>
<td>24.8</td>
</tr>
<tr>
<td>Lehninger</td>
<td>0.970</td>
<td>25.0</td>
</tr>
<tr>
<td>ProMoST</td>
<td>0.978</td>
<td>27.3</td>
</tr>
<tr>
<td>Nozaki</td>
<td>1.024</td>
<td>41.3</td>
</tr>
<tr>
<td>Thurlkill</td>
<td>1.030</td>
<td>43.4</td>
</tr>
<tr>
<td>DTASelect</td>
<td>1.032</td>
<td>44.1</td>
</tr>
<tr>
<td>EMBOSS</td>
<td>1.056</td>
<td>52.3</td>
</tr>
<tr>
<td>Sillero</td>
<td>1.059</td>
<td>53.2</td>
</tr>
<tr>
<td>Patrickios</td>
<td>2.392</td>
<td>3201.8</td>
</tr>
<tr>
<td>Avg_pl*</td>
<td>0.965</td>
<td>23.3</td>
</tr>
</tbody>
</table>

Expected results

1) Better tools for prediction of isoelectric point:
   
   a) sequence based predictor *
   
   b) structure based predictor *

2) Update of *Proteome-pl* database *

* parts which can be published as separate publications
Use deep learning for Isoelectric Point prediction
Deep learning architecture

**generating features**

- AAindex
  - e.g. polarity, bulkiness, hydrophobicity, etc.
- secondary structure
  - PSIPRED (sequence)
  - DSSP (structure)
- solvent accessibility
  - ACCpro (sequence)
  - DSSP (structure)
- number of charged aminoacids with their neighborhood

**deep learning**

Dataset
- protein/peptide sequences

one-hot-encoding

GPU
- TensorFlow

- Convolution
  - MaxPooling
  - LSTM
  - dense layer
  - output

exemplary network architecture

predicted
- $pl$
The plan

Isoelectric point

Peptides
- with PTM (1)
  - Datasets:
    - Gauci et al. 2008
    - Lengqvist et al. 2011
    - Halligan et al. 2004
    - Bekker-Jensen et al. 2017
    - Panizza et al. 2017
- without PTM (2)
  - Datasets:
    - Kozlowski, 2016
    - Atanassov & Urlaub

Proteins
- sequence
  - with PTM (3)
  - without PTM (4)
  - only Eukaryota (5)
  - only Prokaryota (6)
- structure (7)
  - structure provided by the user (PDB file)
  - homology based model (HHblits & MODELLER)

Dataset: Kozlowski, 2016
Datasets:

a) The literature (summarized in Kozlowski, 2016)

a) The collaborators:

- prof. Henning Urlaub (University Medical Center & Max Planck Institute for Biophysical Chemistry, Gottingen, Germany)

- dr Ilian Atanassov (Protein Core Facility at Max Planck Institute for Biology of Ageing in Colone, Germany).
Proteome-pI database

Database of pre-computed isoelectric points for proteomes from different model organisms (5029 species).

Goals of the database include making statistical comparisons of the various prediction methods (18 algorithms implemented) as well as facilitating biological investigation of protein isoelectric point space. Isoelectric point, the pH at which a particular molecule carries no net electrical charge, is important parameter for many analytical biochemistry and proteomics techniques, especially for 2D gel electrophoresis (2D-PAGE), capillary isoelectric focusing (cIEF), liquid chromatography–mass spectrometry (LC-MS) and X-ray protein crystallography.

2D plots of predicted molecular weight and isoelectric point can be useful for initial identification of proteins in the sample and limiting the complexity of the further analyses.

21,721,250 sequences from 5,029 proteomes with isoelectric point predicted using 18 algorithms

Check some of the most frequently used proteomes

- **Homo sapiens** (92173 proteins)
- **Mus musculus** (58774 proteins)
- **Arabidopsis thaliana** (33445 proteins)
- **Drosophila melanogaster** (23296 proteins)
- **Danio rerio** (43305 proteins)
- **Xenopus tropicalis** (35350 proteins)
- **Caenorhabditis elegans** (27795 proteins)
- **Escherichia coli** (4314 proteins)
- **Bacillus subtilis** (4205 proteins)
- **Mycobacterium tuberculosis** (4650 proteins)
- **Salmonella enterica** (5128 proteins)
- **Vibrio cholerae** (3784 proteins)
- **Helicobacter pylori** (1563 proteins)
- **Phage lambda** (63 proteins)
- **T4 phage** (278 proteins)
- **Herpes simplex virus 1** (77 proteins)

If you are interested in the analysis of isoelectric point for proteins coming from all organisms use one of the files

- **PDB** (339k proteins)
- **Swiss-Prot** (550k proteins)
- **UniProtKB/TrEMBL** (63M proteins)
- **nr** (55M proteins)

Proteome-pI database

Proteome-pI database

Current state: good db for 2D-PAGE

After the update: good db also for MS

Two fellowships for master students in NCN OPUS grant
*(Deep learning of high-throughput biological data)*
Warsaw, Poland

The objective of the project entitled *Prediction of protein and peptide isoelectric point based on sequence and structure features* (2018/29/B/NZ2/01403) is a development of the next generation of tools and databases for the isoelectric point prediction using the deep learning.

We are seeking a highly motivated and talented students for working in the field of computational proteomics.

**Requirements:**
- A proactive approach for achieving tasks and objectives of the project, good interpersonal skills, collaborative attitude
- Ability to program in languages such as Python (preferred), R, C++ or other
- Knowledge of Unix (Ubuntu)
- Writing master thesis under the supervision of dr Łukasz Kozłowski (the project PI)
- Bachelor's degree in life science, computer science or related area; being within master studies already
- Very good written and oral communication skills in English
- Knowledge of biology, bioinformatics or computational biology

**We offer:**
- Two scholarship positions for 18 months (1 500 PLN per month)
- Presenting the results during the national and international conferences
- Laptop (plus access to the computer cluster)
- Development of a career in an interdisciplinary and international team studying biologically relevant questions

**More details:** During the project the deep learning will be conducted on massive amounts of the data from the mass spectrometry experiments and PDB and UniProt databases. The machine learning will be performed using TensorFlow & Keras.


**Related resources**
- IPC, (Isoelectric Point Calculator), [http://isoelectric.org](http://isoelectric.org)
- Proteome-pI (Proteome Isoelectric Point Database), [http://isoelectricpointdb.org](http://isoelectricpointdb.org)

**Deadline:** December 31st 2018

**How to apply:** by email

Proponowane tematy prac magisterskich:

1. Przewidywanie punktu izoelektrycznego białek i peptydów z sekwencji za pomocą głębokich sieci neuronowych

2. Przewidywanie punktu izoelektrycznego białek ze struktury trzeciorzędowej za pomocą głębokich sieci neuronowych
Thank you for your time

Any other questions & comments

email: lukasz.kozlowski.lpk@gmail.com  www: bioinformatic.netmark.pl