Journal of Computer Science & Systems Biology - Open Access
Research Article JCSB/Vol.2 January-February 2009

Web Based Theoretical Protein pI, MW and 2DE Map

Itaraju J. B. Brum, Daniel Martins-de-Souza, Marcus B. Smolka, José C. Novello, Eduardo Galembeck*

Depto. de Bioquímica, Instituto de Biologia, UNICAMP, Campinas, Sao Paulo, Brazil

*Correspondent author: Prof. Dr. Eduardo Galembeck, Dept. of Biochemistry, Biology Institute, State University of Campinas - UNICAMP, Campinas, SP, 13083-970 - Brazil, Tel/Fax: +55 19 3521-6138; E-mail: eg@unicamp.br

Availability: http://pro-161-70.ib.unicamp.br/~itaraju/tools/pimw
(For local installation, please contact: itaraju@gmail.com)

Received January 22, 2009; Accepted February 24, 2009; Published February 27, 2009

Citation: Brum IJB, Martins-de-Souza D, Smolka MB, Novello JC, Galembeck E (2009) Web Based Theoretical Protein pI, MW and 2DE Map. J Comput Sci Syst Biol 2: 093-096. doi:10.4172/jcsb.1000020

Abstract

The genomic projects have provided a far wide amount of information that still requires be analyzing and interpreting. That would be impossible to be done without the development of well adapted computational tools that might help the analysis of these data we have collected so far. Due to the need for analyzing proteomes we developed a tool, implemented through the CGI that can simulate the two-dimensional electrophoresis from a whole genome.

Keywords: pI; Isoelectric point; MW; Molecular weight; Two-dimensional gel electrophoresis; Protein tool; Bioinformatics

Introduction

The proteomics field aims to identify proteins and quantify their presence in a cell or tissue (Wilkins et al., 1998). Two-dimensional electrophoresis (2DE) is one of the most used techniques in proteomics. It enables one to separate complex mixtures of proteins according to their molecular weight (MW) and isoelectric point (pI), resulting in an experimental map where spots represent proteins (Williams and Hochstrasser, 1997).

It is possible to implement the algorithms to calculate protein pI and MW (Bjellqvist et al., 1993) and it can be used in the theoretical computation of these properties from the available amino-acid sequences. Prior knowledge of pI and MW values of proteins are valuable information that can direct proteomic experiments. These values are also used as a characteristic of proteins under study and/or used for protein identification.

There are tools that have already been performing such kind of calculation on the Internet (Hiller et al., 2003; Gasteiger et al., 2005). One of them, the ExPASy Server tool for computing pI and MW properties from amino-acid sequences, 'Compute pI/MW Tool' (http://www.expasy.org/tools/pi_tool.html; Gasteiger et al., 2005) is classically employed for this purpose. It enables submission of one amino-acid sequence or a list of Swiss-Prot/TrEMBL ID entries from which pI/MW theoretical values will be computed. Nevertheless, it is not possible to submit a user-defined FASTA file, making it difficult for the end-user to compute pI/WM for a large dataset of user sequences or sequences from which it is not yet known their Swiss-Prot/TrEMBL ID values.

Due to the Open Reading Frames (ORF) availability from whole genomes, if such algorithms are used for calculation

of pI and MW from a list of proteins it can also be used to represent graphically the electrophoresis in two dimensions, as a theoretical 2DE map, showing spots for each proteins according to their pI and MW. It is done in JVirGel tool (http://www.jvirgel.de; Hiller et al., 2003). JVirGel has many features for visualizing a virtual, or theoretical, 2DE gel that can also be constructed from user submitted FASTA file. It lacks an easy interface for simple retrieving computed pI/MW values and other properties, like amino-acid composition table. Both sited tools do not enable the customization of constant parameters used for computing pI values, which would be an interesting feature for biologists with specific applications.

Here we present a pI/MW prediction tool available through a web site that enables the construction of theoretical 2DE maps using user submitted data in a FASTA file. It is focused on simplicity for retrieving computed values and customization of parameters for pI computation.

Program Overview

The first step for using pI/MW prediction tool is filling a web form where sequence data and configuration parameters are specified. The FASTA format (http://www.ncbi.nlm.nih.gov/blast/fasta.shtml, Feb-2009) is expected for submitted sequences, so that one can submit an entire file with potentially thousands of amino-acid sequences. Submission of few sequences, even a single one, is also allowed. The submitted data and specified parameters are processed and a resulting web page shows computed properties and amino-acid composition table, as in the example shown in Figure 1.

The initial web form is also used to select the categories of the desired information to be shown in the resulting report, such as ORF identification (present in FASTA file), the amino-acid sequence of individual ORFs and amino-acid composition table. Another feature is the results exhibition in a raw text, in opposite to the ordinary web pages — the output is suitable for exporting the results to other programs like spreadsheet and databases managers.

The theoretical 2DE Map, can be plotted or not, depending on the user preferences. Otherwise, if the map is displayed, one can change the map's default scale. Figure 2

ORF:

XF0002 (XF-03E01-GL09)

Sequence:

MRFRLQRETFLKPLAHVVNVVERRQTRSIL ANLLIKVNEDQLSLTGTDLEVEMISKTIIE
DAESGEITIPARKIYEIVRALPDSSQLSVY QSDDKITLQAGRSRFTLATLPANDFPSIDK
IEVTERIIIPEVLLKELIERTAFAMAQQDV RYYLNGLLFDLRDTKLRCVATDGHRLALCE
TELEQAKDLKRQIILPRKGVMELQRLLEGS DRQIELEIARNHIRMKSFDVTFTSKLIDGS
FPDYEGVIPIGADREVKVAREVLRDALQRA AILSNEKYRGVRIEVSPGQLKINAHNPEQE
EAQEEIEAQTIVDGLAIGFNVNYLLDALSS LRGDFVNIQLRDSNSSALIRESNSEKSLQV VMPLRL

MW:	pI:
41549.68	5.36

Amino-acid composition

Ala (A)	26	7.1%	Met (M)	6	1.6%
Cys (C)	2	0.5%	Asn (N)	14	3.8%
Asp (D)	23	6.3%	Pro (P)	12	3.3%
Glu (E)	34	9.3%	Gln (Q)	19	5.2%
Phe (F)	11	3.0%	Arg (R)	32	8.7%
Gly (G)	15	4.1%	Ser (S)	24	6.6%
His (H)	4	1.1%	Thr (T)	17	4.6%
Ile (I)	32	8.7%	Val (V)	25	6.8%
Lys (K)	17	4.6%	Trp (W)	0	0.0%
Leu (L)	46	12.6%	Tyr (Y)	7	1.9%

Total: 366

Figure 1: The resulting page shows information about submitted sequences like pI, MW, amino-acid sequence, amino-acid composition.

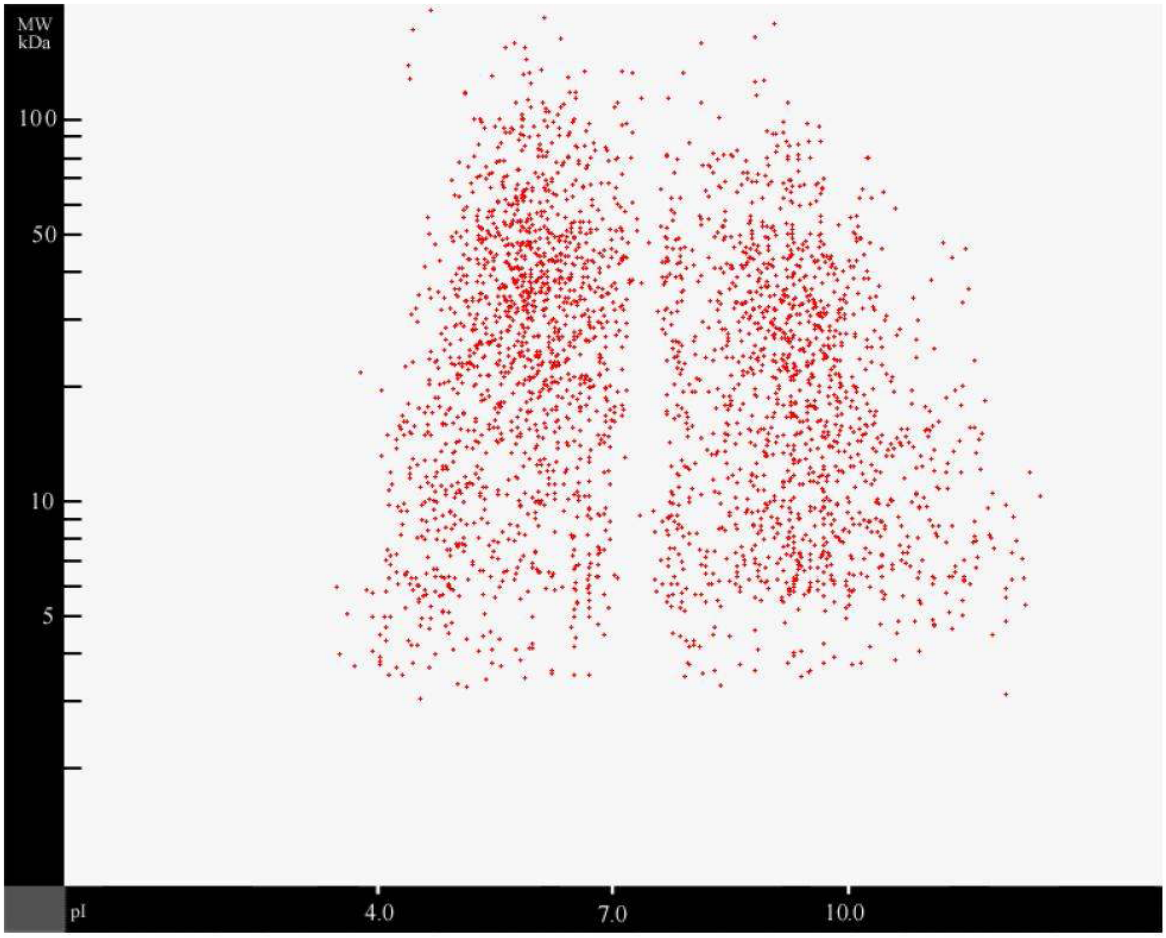

Figure 2: Theoretical 2DE map created by the pI/MW prediction tool. Sequences from the 2,830 ORFs from Xylella fastidiosa's proteome were used. Each red dot represents a spot of a protein which sequence was submitted.

shows a theoretical 2DE map for the 2,830 ORFs from *Xylella fastidiosa*'s proteome as an example. Once the pI calculation is influenced by nine pK numerical values, the initial form also enables one to type pK values other than the default ones. The pK values are used by the pI calculation software component leading to a customizable calculation.

The computed values of pI and MW are not significantly different from the ones computed with other tools (data not shown). However, the experimental conditions have to be observed when computing these theoretical values. The presence of posttranslational modification in proteins structure or the not exposure of all charged residues to the medium, due to e.g. conformational 3D protein folding, are factors that would impact in great differences when comparing experimental and computed values for the considered properties.

Design

The pI/MW prediction and theoretical 2DE Map tools were implemented in Perl language and use CGI (Common Gateway Interface) in order to be available on the Internet. The theoretical 2DE map returned is a GIF image plotted as graphic with pI and MW identified as the coordinated axes. For GIF file generation, FLY (http://martin.gleeson.com/fly) program was employed.

For computing molecular weight (*MW*) of an individual amino-acid sequence we employ Equation 1, which is the sum of the molecular weight of each amino-acid residue *i* in the considered sequence. The amount of each of 20 common amino-acids is expressed as n_i. As amino-acid residues molecular weight (MW_i) is the weight of single monomers that make up the pepidict sequence, formed after the pepedict chemical reaction in which one water molecule is

lost, we need to add the molecular weight of a water molecule (MW_{H2O}), in Equation 1, in order to preserve the actual sum of weights of termini amino-acids.

$$MW = MW_{H2O} + \sum_{i=1}^{20} n_i \cdot MW_i$$

Property pI is computed by finding the value of pH that is the root of Equation 2, which is the theoretical net charge of an amino-acid sequence in a given pH (Bjellqvist et al., 1993). Equation 2 assumes that all charged residues in a peptide of protein are exposed to the medium, so that they can be found in a protonated or deprotonated state depending on the medium pH. The amount of each basic or acid amino-acid present in the sequence, Nb_i and Na_i respectively, and their dissociation constants pKb_i and pKa_i, respectively, are considered for computing the pI value. As the amine or carboxyl groups in terminal ends of a protein sequence can also be charged, constants pKb_i, for N-termini, and pKa_i, for C-termini, are also considered in Equation 2, and their values vary according to which amino-acid is present in the C- and N-termini. Default dissociation constants are taken from (Bjellqvist et al., 1993).

$$cl = \sum_i \frac{Nb_i \cdot 10^{-pH}}{10^{-pH} + 10^{-pKb_i}} + \sum_i \frac{Na_i \cdot 10^{-pH}}{10^{-pH} + 10^{-pKa_i}} - Na_i$$

Discussion & Conclusion

The presented tool was designed to perform a high throughput analysis on pI and MW from a FASTA file. It has also a flexible and customizable interface, which permits the construction of theoretical 2DE Electrophoresis Maps and compare than with reference maps. The report page is also customizable depending on the pK values set and the desired information on it. Its format can be easily imported to spreadsheets and databases.

Users can benefit from a a simple interface that directly combines, in the results pages, information that would require one of: laborious submission of individual sequences and manual work for organizing results or computer programming skills for creating customized programs.

Acknowledgment

Financial support: FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo).

References

1. Bjellqvist B, Hughes GJ, Pasquali C, Paquet N, Ravier F, et al. (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their ammino acid sequences. Electrophoresis 14: 1023-1031. » CrossRef » Pubmed » Google Scholar
2. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, et al. (2005) Protein Identification and Analysis Tools on the ExPASy Server. In: John M. Walker (ed): The Proteomics Protocols Handbook. Humana Press pp 571-607. » CrossRef » Google Scholar
3. Hiller K, Schobert M, Hundertmark C, Jahn D, Münch R (2003) JVirGel: calculation of virtual two dimensional protein gels. Nucleic Acids Res 31: 3862-3865.» CrossRef » Pubmed » Google Scholar
4. Wilkins MR, Gasteiger E, Tonella L, Ou K, Tyler M, et al. (1998) Protein Identification with N and C terminal Sequence Tags in proteome Projects. J Mol Biol 278: 599-608.» CrossRef » Pubmed » Google Scholar
5. Williams KL, Hochstrasser DF (1997) Introduction to Proteome. In: Wilkins MR, Williams KL, Appel RD, Hochstrasser (eds) Proteome Research: New Frontiers in Functional Genomics. Springer-Verlag Berlin Heidelberg. Germany pp1-11.