ELSEVIER

Computers in Biology and Medicine 38 (2008) 23–30

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

Improved application of the oscillating method for the isoelectric point determination of protein: Potential connection with protein data banks

Andrés Maldonado, Francisco Vara, Antonio Sillero*

Departmento de Bioquímica, Facultad de Medicina, Instituto de Investigaciones Biomédicas Alberto Sols (UAM/CSIC), Arzobispo Morcillo 4, 28029 Madrid, Spain

Received 16 October 2006; received in revised form 23 May 2007; accepted 11 June 2007

Abstract

The oscillating method (OM) for the theoretical determination of the pI values, one by one, of proteins and other macromolecules has been previously published [Sillero and Maldonado, Comput. Biol. Med 36 (2006) 157–166]. An improved application of the method, here named as improved oscillating method (IOM), allows the pI determination of group of proteins. This characteristic may be useful to explore the pI value and electric charge of family of enzymes. As an example the pI values of 1630 enzymes collected in a Swiss-Prot data bank (www.expasy.org), as belonging to the enzymes ligases (EC 6. 2. 1. *) is presented. The method also permits the determination of the pI value of any group of proteins stored in data banks provided that they can be supplied to the program in a FASTA format. Free access to IOM can be reached at http://www.bq.uam.es/otros/pical3.zip.

Keywords: Electric charge; PH; Proteins; PI theoretical determination; Visual basic; Acid–base residues; PICAL

1. Introduction

Several procedures have been published from our laboratory for theoretical determination of the pI value of proteins and other macromolecules: simplified [1], abridged [1], comprehensive [2] and oscillating methods [3]. In all the cases, two types of acid–base residues are considered, *N* and *P*, depending on whether they are *N*eutral or *P*rotonated, when undissociated, respectively. The name, groups (N_1, N_2, N_3, N_4, P_1, P_2, P_3, P_4) and their pK values are summarized in Table 1. As it is well known, the charge of an each acid–base residue depends on the pH value, as deduced from the Henderson–Hasselbalch formula or from some of its deduced variants [4,5]. The pI value of a protein is attained when the positive and negative charges of its acid–base groups are equals.

In the simplified method only two types of *N* (N_a and N_b) and two types of *P* (P_a and P_b) are considered (Table 1). In this case, the pK values (between brackets) and the components of each subgroup are: $N_a = N_1 = N_2 = N_3$ (4.2); $N_b = N_3 = N_4$ (9.5); $P_a = P_2 = P_3 = P_4$ (9.5); $P_b = P_1$ (6.4). This simplification allows the approximate determination of the pI value of proteins with the help of only appropriate published tables [1]. Other simplified procedures for the determination of pI values of proteins have been also reported [4,6].

In the comprehensive method, the pI value is calculated through the solution of polynomials with a degree equal to the number of groups with different pK values present in the molecule [2]. The coefficients of the polynomials are summatories of complex arrays of products of acid dissociation constants that can be calculated with the help of an algorithm developed in this laboratory [7].

We have recently described a so-called oscillating method (OM) [3] that: (i) calculates the charge of one protein versus pH values; (ii) calculates the pI value with much less mathematical apparatus and with the same accuracy as the comprehensive methods [3]. This method is useful in research and for didactical purposes in regular biochemical courses. The program, written in visual basic, calculates the isoelectric point and the charge of a protein versus pH values in the form of a graph or a table. However, the calculation of the parameters for several

* Corresponding author. Tel.: +34 91 4975413; fax: +34 91 49585401.
E-mail address: antonio.sillero@uam.es (A. Sillero).

0010-4825/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2007.06.004

Table 1
Types and number of groups, chemical structure and pK values of the acid–base residues considered to be present in a protein

Type	Chemical residue	pK value	Number
N_1	Terminal carboxyl	$pK_{N_1} = 3.2$	N_1
N_2	β-carboxyl of aspartate	$pK_{N_2} = 4.0$	N_2
N_3	δ-carboxyl of glutamate	$pK_{N_3} = 4.5$	N_3
N_4	Thiol of cysteine	$pK_{N_4} = 9.0$	N_4
N_5	Phenol of tyrosine	$pK_{N_5} = 10.0$	N_5
N_a	$N_a = N_1 = N_2 = N_3$	$pK_{N_a} = 4.2$	N_a
N_b	$N_b = N_3 = N_4$	$pK_{N_b} = 9.5$	N_b
P_1	Imidazolium of histidine	$pK_{P_1} = 6.4$	P_1
P_2	Terminal ammonium	$pK_{P_2} = 8.2$	P_2
P_3	ε-ammonium of lysine	$pK_{P_3} = 10.4$	P_3
P_4	Guanidinium of arginine	$pK_{P_4} = 12.0$	P_4
P_a	$P_a = P_2 = P_3 = P_4$	$pK_{P_a} = 9.5$	P_a
P_b	$P_a = P_1$	$pK_{P_b} = 6.4$	P_b

proteins had to be made one by one, what in our experience is a limitation.

Here, we present an improved application of the OM that allows the simultaneous calculation of the pI value of several proteins and, above all, permit analysis of the changes in both pI and electric charges of family of proteins. As an example of this application, variation in the pI values of the enzymes collected as ligases by the Enzyme Commission is here presented.

2. Materials

The pK values considered in this work for the N and P acid–base groups of the proteins are specified in Table 1. Note as five different N and five different P groups are annotated. Although not used in this work, the two N groups (N_a and N_b) and the two P groups (P_a and P_b) are also assembled in Table 1.

3. Results and discussion

3.1. General aspects

The application of the IOM and the OM has similar computational cores. For this reason, only codes referring specifically to the IOM are annotated in the Appendix. The interested readers may consult [3] for further computational detail. The program calculates the charge of each acid–base residue (N_1, N_2, N_3, N_4, N_5, P_1, P_2, P_3, P_4) at increasing pH values, starting from zero and with increment of 0.1 pH units. The net charge of the protein (NQ), at the beginning of calculation always a positive value, is consecutively calculated, at each pH value, until a value of NQ (a) < 0 is firstly obtained. In this case, a pH = pI(a) is reached: the net charge of the protein is negative at this point and at pH = pI(a) − 0.1 the charge is positive. Therefore, pI(a) represents the first significant decimal figure of the pI value of the protein. For the calculation of the second decimal figure of pI, the loop start at pH = pI(b) = pI(a) − 0.1, with successive increments of 0.01 pH units. The process can continue oscillating (hence its name) until the precision required (in decimal figures) is reached. As

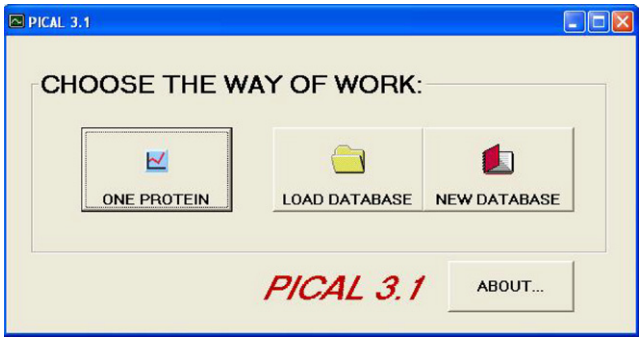

Fig. 1. First screen of the improved oscillating method. It directs to determination of the pI and electric charge of one protein, creation of a new database or to load a database.

the program calculates the charge of the protein at each pH value, these data are afforded in the form of graph or table.

After an intensive use of the OM for more than two years, we have observed several drawbacks in its application: (i) only one protein each time could be approached; (ii) data from this analysis could be printed but not saved; (iii) the program had some internal computational loops that could be, and have been simplified; and (iv) it could not be applied to groups of proteins stored in data banks. Accordingly, the OMs below reported presents important improvements over the one previously published [3].

The first screen of the program (Fig. 1) invites to follows three routes of action: work with one protein, with a stored database or to create a new database.

3.2. Calculation of the parameters of one protein

In this case, the methodology is very similar to that followed in the OM (compare Fig. 2 of this paper with Fig. 3 in [3]). From this second screen (Fig. 2) one get automatically the pI value of the investigated protein and a graph with the charge of the protein versus pH values. The data (Code A in the Appendix) can be: (i) saved into a file (with the same, or different, name of the protein); (ii) printed; and (iii) represented in a table in the

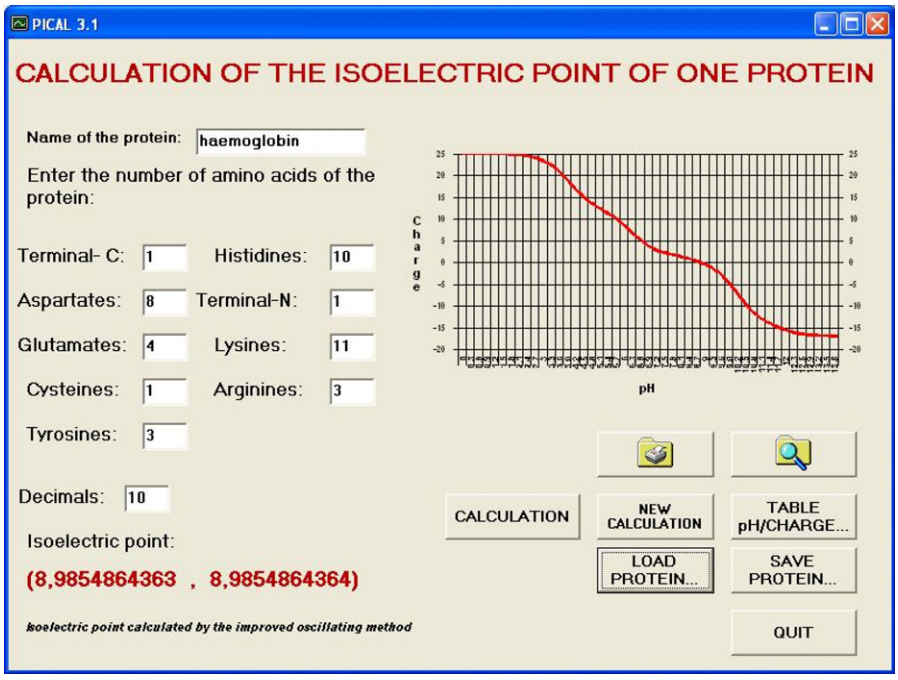

Fig. 2. Determination of the parameters of one protein. The name of the protein and the number of acid–base residues can be introduced. The computer automatically supplies the pI value and the charge of the protein versus pH value. For more explanation see the text.

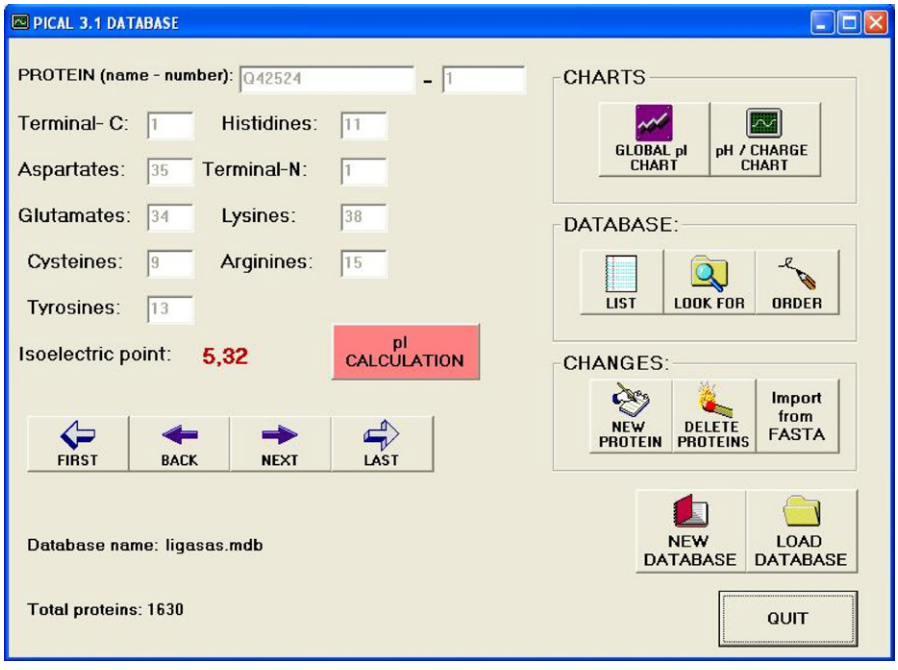

Fig. 3. Main screen to handle proteins included in a file. For more information on this screen see the text.

form of pH versus charge. New calculations can be worked out for new proteins, and saved. A file, corresponding to a protein (haemoglobin..pcl), is included in the program as an example. The isoelectric point can be given with the desired number of decimals. Usually two decimal figures are enough for most purposes (Code B in the Appendix).

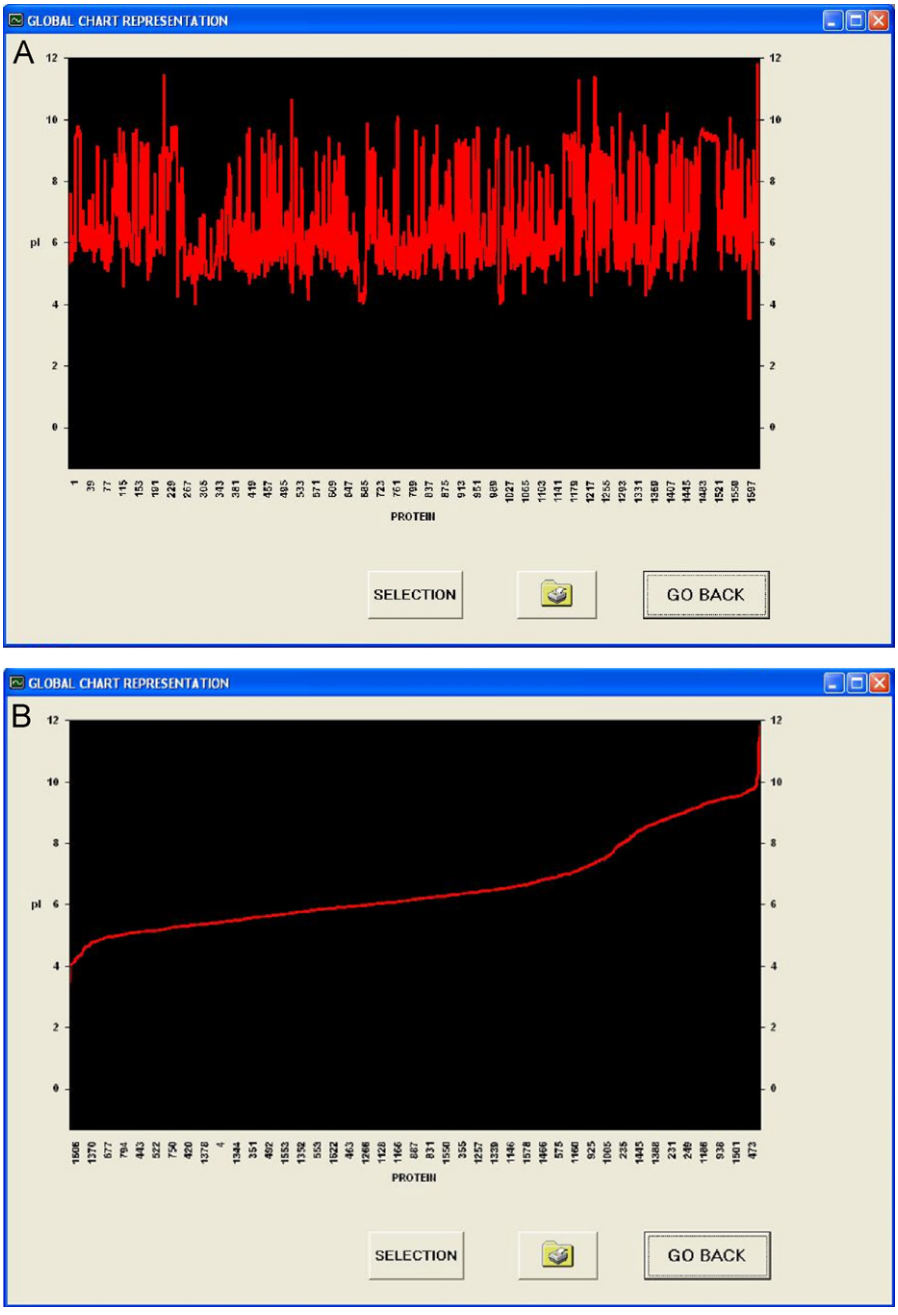

Fig. 4. Variation in the pI values of enzymes ligases represented following a protein order (A) or increasing pI values (B).

3.3. Create a new database

After pressing the corresponding button (Fig. 1, "create new database"), the name of a new file are requested. Once saved, a new screen (Fig. 3) appears with the superimposed message "The data base xxx has been created" containing, of course, zero proteins. New proteins can be manually added, or eliminated with the corresponding buttons included in the frame "Changes". At any moment the following can be requested from the program (Fig. 3): (i) global pI representation of all the proteins in the file (Fig. 4A); (ii) pH/charge representation of the specific protein on the screen; (iii) list of all the proteins present in the file (Code C in the Appendix); (iv) look for a protein in the database; (v) to order the proteins by name, isoelectric point or input order: once this order is established a new global pI chart can be obtained, as requested (Codes D and E in the Appendix) (Fig. 4B).

3.4. Interaction with protein data banks

As an example, the file ligases.mdb is included in the program and can be reached from the program with the button "Import form FASTA" (Fig. 3). The usefulness of the program to handle

this type of files containing a great number of proteins can easily tested as explained above and following the routes specified in Fig. 3.

It is very important that, when importing the file ligases.txt, the program is empty of any previously charged protein file, otherwise the content of ligases will be added to the present file. Because of that it is convenient to create a new database and after that import the file ligase.mbd.

The present method offers the possibility of exploring pI values of any group of proteins collected in data banks, for example the Swiss-Prot data banks. The steps to follow are the following: (a) The Swiss-Prot data bank can be reached with the command www.expasy.org; (b) the search can be started with the desired group of proteins; (c) the data can be saved as a file with an appropriate name; in the context of this work the file must be saved in a FASTA format (xxx.txt); (d) this file can be reached with the "Import from FASTA" button of the program (Fig. 3) and manipulated as above explained. The IOM automatically store the file in a format xxx.mdb appropriate to be handled by the program (F Code in the Appendix).

4. Summary

We have used extensively for already two years the OM for research and didactical purposes in this Medical School. Through its use, the students may get a practical knowledge of the relationship between the composition in acid–base residues of a protein and its pI value and, as a corollary, they also familiarize with the electric charge of a protein in function of the surrounding pH, its migration in an electric field and of its tendency to bind to cationic or anionic matrices.

The main drawback observed with the OM was the impossibility of comparing simultaneously proteins with different acid–base composition. This disadvantage has been solved with the IOM here reported. The method also allows the creation of a bank of model proteins to better appreciate the influence of certain acid–base residues, at determined pH values on the charge of a protein. On the other side, we are actually approaching the utility of this method to analyze superfamilies, families or even domain of proteins collected in data banks, related with their pI values and electric charges.

The knowledge of the pI values may be of interest in many fields of protein research: protein purification; characterization of small peptides or active areas of enzymes, to predict ionic interactions between different proteins or between proteins and other macromolecules, to interpret potential deviation between the theoretical and experimental and the calculated pI of a protein, etc. An important part of modern proteomic deal with the simultaneous analysis of mixtures of proteins contained in biological samples [8–13]. The development of new techniques for the detection of minimal amounts of proteins, together with the use of two-dimensional gel analysis combining separation procedures based on the different molecular weights and isoelectric proteins of proteins have been relevant landmarks in these types of studies. In this regard, the development of theoretical methods to evaluate and predicts the pI values of macromolecules are very important tools to rationalize the results experimentally obtained.

Acknowledgments

This investigation was supported by grants from Dirección General de Investigación Científica y Técnica (BFU2006-04307/BMC).

Appendix

New codes concerning the improved oscillating method (IOM). These codes complement commands on the oscillating method (OM) previously published [3].

A CODE:

```
CommonDialog1.CancelError = True
On Error Resume Next
   'pcl is the extension for pical
   CommonDialog1.Filter = "pical archives" & "(*.pcl)|*.pcl|"
   CommonDialog1.ShowSave
Open CommonDialog1.FileName For Output As #1
Write #1, name, N1, N2, N3, N4, N5, P1, P2, P3, P4
Close #1
End Sub
```

B CODE:

```
////////////////////////OSCILLATING METHOD////////////////////////

r = Val(Text1(9).Text)
pH = 0
n = 1
Do
   Do
   QT = Module1.QT(N1, N2, N3, N4,N5, P1, P2, P3, P4, pH)
   pH = pH + 10^(-n)
```

```
    Loop Until QT <= 0
    pH = pH - 10^(-n)
    If QT = 0 Then
        Label17.Visible = True
        Label13.Caption = pH
        GoTo 30
    End If
    pH = pH - 10^(-n)
n = n + 1
pH = pH + (10^(-n))
Loop Until n > r Or QT = 0'maximum of decimal figures is 16.
pH = Round(pH - (10^(-n)), r)
n = n - 1
pH1 = pH - (10^(-n))
pH2 = pH
Label13.Caption = "(" & pH1 &" , " & pH2 &")"
```

MODULE 1

```
Function QT(ByVal N1, N2, N3, N4, N5, P1, P2, P3, P4, pH As Variant) As Variant
    Q1 = (-N1)/(1 + (10^(3.2 - pH)))
    Q2 = (-N2)/(1 + (10^(4 - pH)))
    Q3 = (-N3)/(1 + (10^(4.5 - pH)))
    Q4 = (-N4)/(1 + (10^(9 - pH)))
    Q5 = (-N5)/(1 + (10^(10 - pH)))
    Q6 = (P1)/(1 + (10^(pH - 6.4)))
    Q7 = (P2)/(1 + (10^(pH - 8.2)))
    Q8 = (P3)/(1 + (10^(pH - 10.4)))
    Q9 = (P4)/(1 + (10^(pH - 12)))
    QT = Q1 + Q2 + Q3 + Q4 + Q5 + Q6 + Q7 + Q8 + Q9
End Function
```

C CODE:

```
Private Sub Command6_Click()
Form6.Show
Form6.Data1.Recordset.Index = "Name"
Form6.Data1.Recordset.MoveFirst
Form6.Caption = Label15.Caption
Unload Me
End Sub
```

```
Private Sub Form_Load()
Data1.DatabaseName = Form4.Data1.DatabaseName
Data1.Refresh
DBGrid1.Refresh
End Sub
```

D CODE:

```
Private Sub Form_Load()
With MSChart1
.ColumnCount = 1
.RowCount = Form4.Label2.Caption
.AutoIncrement = True
End With
Form4.Data1.Recordset.MoveFirst
Do
MSChart1.RowLabel = Form4.Text10.Text
MSChart1.Data = Form4.Label13.Caption
```

```
Form4.Data1.Recordset.MoveNext
Loop Until Form4.Data1.Recordset.EOF = True
MSChart1.Repaint = True
Form4.Data1.Recordset.MovePrevious
End Sub
```

E CODE:

```
Private Sub Command1_Click()
If Option1.Value = True Then
Form4.Data1.Recordset.Index = "Name"
Form4.Data1.Recordset.MoveFirst
End If
If Option2.Value = True Then
Form4.Data1.Recordset.Index = "IP"
Form4.Data1.Recordset.MoveFirst
End If
If Option3.Value = True Then
Form4.Data1.Refresh
Form4.Data1.Recordset.MoveFirst
End If
Unload Me
End Sub
```

F CODE:

```
Open textfile For Input As #1
Value = 0
Do Until EOF(1)
Line Input #1, VarText
If (Left(VarText, 1) = ">") Then
If value > 0 Then
For counter2 = 65 To 90
letter(counter2 - 65) = 0
Next counter2
For character = 1 To Len(content)
temporal = Mid(content, character, 1)
For counter2 = 65 To 90 If temporal = Chr(counter2) Then letter(counter2 - 65) =
letter(counter2- 65) + 1
Next counter2
Next character
Form4.Data1.Recordset.AddNew
Text10.Text = titles
Text1(1).Text = letter(3) 'Asp = D
Text1(2).Text = letter(4) 'Glu = E
Text1(3).Text = letter(2) 'Cys = C
Text1(4).Text = letter(24) 'Tyr = Y
Text1(5).Text = letter(7) 'His = H
Text1(7).Text = letter(10) 'Lys = K
Text1(8).Text = letter(17) 'Arg = R
Form4.Data1.Refresh
title = ""content = ""
End If value = value + 1
titles = title & VarTexto
Else content = content & VarTexto
End If Loop
For counter2 = 65 To 90
letter(counter2 - 65) = 0
Next counter2
```

```
For character = 1 To Len(content)
temporal = Mid(content, character, 1)
For counter2 = 65 To 90 If temporal = Chr(counter2) Then letter(counter2 - 65) =
letter(counter2- 65) + 1
Next counter2
Next character
Form4.Data1.Recordset.AddNew
Text10.Text = titles
Text1(1).Text = letter(3)
Text1(2).Text = letter(4) Text1(3).Text = letter(2)
Text1(4).Text = letter(24)
Text1(5).Text = letter(7)
Text1(7).Text = letter(10)
Text1(8).Text = letter(17)
Form4.Data1.Refresh Close #1
```

References

[1] A. Sillero, J.M. Ribeiro, Isoelectric points of proteins: theoretical determination, Anal. Biochem. 179 (1989) 319–325.

[2] J.M. Ribeiro, A. Sillero, A program to calculate the isoelectric point of macromolecules, Comput. Biol. Med. 21 (1991) 131–141.

[3] A. Sillero, A. Maldonado, Isoelectric point determination of proteins and other macromolecules: oscillating method, Comput. Biol. Med. 36 (2006) 157–166.

[4] D.S. Moore, Amino acid and peptide net charges: a simple calculational procedure, Biochem. Educ. 13 (1985) 10–11.

[5] J.C. Cameselle, J.M. Ribeiro, A. Sillero, Derivation and use of a formula to calculate the net charge of acid–base compounds, its application to aminoacids, proteins and nucleotides, Biochem. Educ. 14 (1986) 131–136.

[6] C.S. Patrickios, E.N. Yamasaki, Polypeptide amino acid composition and isoelectric point. II. Comparison between experiment and theory, Anal. Biochem. 231 (1995) 82–91.

[7] J.M. Ribeiro, A. Sillero, An algorithm for the computer calculation of the coefficients of a polynomial that allows determination of isoelectric points of proteins and other macromolecules, Comput. Biol. Med. 20 (1990) 235–242.

[8] M.W. Reeves, M.C. Drummond, M. Tager, Partial purification and characterization of the multiple molecular forms of staphylococcal clotting activity (coagulase), J. Bacteriol. 148 (1981) 861–868.

[9] A. Santagostino, G. Giagnoni, P. Fumagalli, D. Pavesi, E. Torretta, Isoelectric point determination of human and camel beta-endorphin, alpha-endorphin and enkephalins, Biochem. Biophys. Res. Commun. 104 (1982) 577–582.

[10] P.M. Larsen, L. Storgaard, S.J. Fey, Proteins present in bovine papillomavirus particles, J. Virol. 61 (1987) 3596–3601.

[11] G. Henriksson, A.K. Englund, G. Johansson, P. Lundahl, Calculation of the isoelectric points of native proteins with spreading of pKa values, Electrophoresis 16 (1995) 1377–1380.

[12] R. Gras, M. Muller, E. Gasteiger, S. Gay, P.A. Binz, W. Bienvenut, C. Hoogland, J.C. Sanchez, A. Bairoch, D.F. Hochstrasser, R.D. Appel, Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection, Electrophoresis 20 (1999) 3535–3550.

[13] K. Hiller, M. Schobert, C. Hundertmark, D. Jahn, R. Munch, JVirGel: calculation of virtual two-dimensional protein gels, Nucleic Acids Res. 31 (2003) 3862–3865.

Andrés Maldonado born in 1984 is a fifth year medical student whose hobbies are computer programming and chess. At present he is an Erasmus student in Berlin Humboldt Universität (Charité).

Francisco Vara is Ph.D. degree in biochemistry (Universidad Complutense de Madrid) in 1982. He was a postdoctoral Fellow in the Imperial Cancer Research Fund, London (UK) (1984–1986). At present he is Assistant Professor of Biochemistry and Molecular Biology in the Medical School of Universidad Autónoma de Madrid and is in charge of the Web page of the Department of Biochemistry.

Antonio Sillero received both the Ph.D. degree in Chemistry (Universidad Complutense de Madrid) and the M.D. Ph.D. degree (Universidad de Granada) in 1966. He was a postdoctoral Fellow of both the U.S. National Institutes of Health and the Jane Coffin Childs Memorial Fund at the Department of Biochemistry of the New York University Medical Center (1967–1970). His research interest have included: control of the irreversible enzymes of glycolysis and gluconeogenesis in liver, metabolism of fructose, RNA bacteriophages, development of *Artemia*, proteases, purine nucleotide metabolism, degradation, synthesis and function of dinucleoside polyphosphates, theoretical determination of electric charges of macromolecules, enzyme kinetics. Lately, he studies the response of yeast to oxidative stress and the mechanism of action of bisphosphonates. At present, he is Professor of Biochemistry and Molecular Biology in the Medical School of Universidad Autónoma of Madrid.