

The modal distribution of protein isoelectric points reflects amino acid properties rather than sequence evolution

Georg F. Weiller^{1,2}, Gilles Caraux³ and Nicole Sylvester²

¹ARC Centre of Excellence for Integrative Legume Research, Australian National University, Canberra, Australia

²Research School of Biological Sciences, Australian National University, Canberra, Australia

³Département d'Informatique Fondamentale et Appliquée, LIRMM, Montpellier, France

Two-dimensional gel electrophoresis, a routine application in proteomics, separates proteins according to their molecular mass (M_r) and isoelectric point (pI). As the genomic sequences for more and more organisms are determined, the M_r and pI of all their proteins can be estimated computationally. The examination of several of these theoretical proteome plots has revealed a multimodal pI distribution, however, no conclusive explanation for this unusual distribution has so far been presented. We examined the pI distribution of 115 fully sequenced genomes and observed that the modal distribution does not reflect phylogeny or sequence evolution, but rather the chemical properties of amino acids. We provide a statistical explanation of why the observed distributions of pI values are multimodal.

Keywords: Extremophiles / Genome / Isoelectric point / Multimodal distribution

Received	19/12/02
Revised	5/9/03
Accepted	8/9/03

1 Introduction

Proteins exhibit very sharp pI s and can therefore be very well separated by IEF [1]. Images of 2-DE gels depicting the proteomes of various organisms have been widely published, and interactive websites have been established where the individual protein spots of these images provide a link to further information on the respective protein. For organisms for which the genomic sequence is determined it is also possible to create an image of their theoretical proteome by plotting the theoretical pI of all predicted proteins against their theoretical M_r . Bjellqvist *et al.* [2, 3] have shown that the pI of a denatured linear protein, as observed in 2-DE, can be calculated with high accuracy. While most 2-DE gels utilise narrow pI gradients only revealing expressed proteins within a limited pI range (e.g., pH 3–7, or pH 7–9), theoretical proteomes typically display the distribution of M_r s and pI s of the entire proteome over the entire pI range (pH 0–14).

Our bioinformatics laboratory forms part of an ongoing *Sinorhizobium* proteome project for which we have produced several theoretical proteome plots. Curiously, a roughly bimodal distribution of the pI s of the proteins was observed in all plots, with the scarcity of proteins within the pH 7.4–7.6 range being the most significant feature. Other investigators have made similar observations. Already in the early 1980s Gianazza and Righetti [4] showed that protein M_r s and pI s are not normally distributed. Urquart *et al.* [5] observed a bimodal pI distribution in *Mycobacterium bovis*. VanBoegelen *et al.* [6] analysed several bacteria and also observed a bimodal distribution of pI s with peaks centred around pH 5.5 and pH 9. The authors assumed that this distribution reflected the intracellular pH of the cell. As proteins are generally poorly soluble near their pI [7] it would appear that most proteins have evolved to avoid a pI close to the pH of the cytoplasm, which is assumed to be near neutrality. Schwartz *et al.* [8] extended this analysis to several eukaryotes for which they reported a trimodal distribution. They confirmed the previously observed bimodal distribution for bacteria and archaea and presented an additional peak near pH 7 as being characteristic for eukaryotes representing nuclear proteins. The authors also speculate that the pI of eukaryotic proteins reflects the subcellular localization, and suggest that theoretical pI calculations may

Correspondence: Dr Georg F. Weiller, Bioinformatics Laboratory, Genomic Interactions Group, Research School of Biological Sciences, Australian National University, Canberra, ACT 0200, Australia

E-mail: gweiller@rsbs.anu.edu.au

Fax: +61-(0)-2-6125-9709

provide a way of tentatively assigning subcellular localizations to proteins that are identified in sequenced genomes but not further characterised. Independently of the studies above, we have analysed the pI distributions of 115 completely sequenced genomes or chromosomes spanning all three kingdoms of life as well as mitochondria, chloroplasts and viruses. Our results differ in several aspects from previous reports and we arrive at a different conclusion. Here, we briefly describe the breadth of observed variations and offer an explanation for the distribution of protein pI values found in different organisms.

2 Methods

The predicted protein sequences of 115 complete genomes (available in January 2001) were obtained from the National Centre for Biotechnology Information [9] and the Protein Information Resource [10]. The data span 115 different genomes and included 8 Archea, 32 Eubacteria, 5 Eucaryotes (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Plasmodium falciparum*), 48 mitochondria, 9 chloroplast, 9 viruses of prokaryotes and 4 viruses of eukaryotes. Programs for predicting the M_r and pI from predicted proteins and for binning the data given in the histograms were written in C. The pI values were calculated using a stepwise approach with the pH value increasing by 0.1 until the positive and negative charges in the protein were equal. Note that the calculated pI depends considerably on the set pK values assumed for the ionisable groups. When several different sets of published pK values were used the predicted pI of some proteins differed by up to 1 pH unit. We have observed that the shapes of the pI distributions vary considerably if other pK values are used, but the multimodal shape per-

sists. All the data presented here use the pK values of amino acids described by Bjellqvist *et al.* [2], which were defined by examining polypeptide migration between pH 4.5 to 7.3 in an IPG gel environment with 9.2 M and 9.8 M urea at 15°C or 25°C, and are therefore close to the values obtained in 2-DE.

The programs for generating random sequences according to a specified multinomial model were written in C using *rand1* and *gasdev* as random number generators [11]. The first one was used to generate random sequences according to a specified multinomial distribution model. The second one is a normally distributed random deviation generator. We used it to approximate multinomial distribution by a normal distribution. The pI distribution histograms were plotted with MS Excel (Microsoft Corporation) and Sigmaplot (Statistical solutions, Saugus, MA, USA).

3 Results and discussion

3.1 Distribution of estimated protein pI in genomic sequences

Figure 1 shows the theoretical pI distributions found in 135 621 proteins derived from 115 fully sequenced genomes or chromosomes. The first three histograms summarise the proteins of eukaryota, eubacteria and archea, respectively. Most proteins form part of one of two major clusters with peaks at pH 5.5 and 9.5, respectively, representing the previously observed bimodal distribution of pI values. Very few proteins have a pI between 7 and 8 except around pH 7.8 where a small and narrow third peak is found. On the acidic side of the spectrum the distribution raises sharply. No protein in the analysed dataset has a pI below 3.0 and only 394 (less than 0.3%) pro-

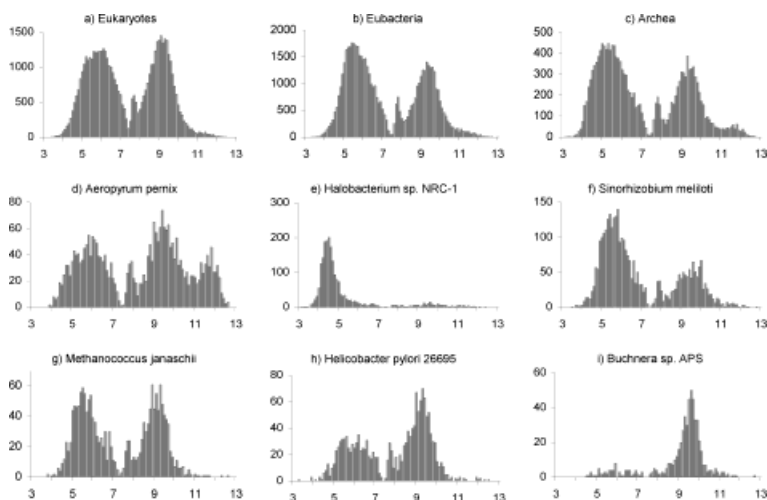


Figure 1. Histograms of computationally estimated pI distributions in steps of 0.1 pI . The pI values ranging from 3–13 are given on the abscissa and the number of proteins on the ordinata.

teins have a pI below 4. In contrast, the alkaline side has a long tail. Here, 9706 proteins (more than 7%) have a pI of 10 or above. This is due to a small group of proteins which have a alkaline pI that forms a fourth peak around pH 12. This fourth peak is more visible in archaea and is particularly pronounced in the archaeon *Aeropyrum pernix* (Fig 1d).

Besides this, the distribution in the different kingdoms is remarkably similar. The only apparent differences being that in eukaryota the main alkaline peak is slightly higher than the acidic peak, while the reverse is true for both prokaryotic kingdoms. These differences are however, not specific for the kingdoms. For instance, the distributions of *S. cerevisiae* and *D. melanogaster* are very similar to the eubacterial histogram (panel b) with the main acidic peak being the most prominent (data not shown). The last 6 panels (d–i) of Fig. 1 show some individual prokaryotic species. Note that the pH of the four peaks is the same for all organisms, however, the relative abundance of proteins belonging to the four peaks differs largely in individual species. These differences do not reflect phylogeny.

In most organisms neither of the two major peaks contain more than about 75% of the proteins. These distributions are exemplified in panels f–h. In a small number of organisms the discrepancy is however, more pronounced, with *Halobacterium* (panel e) and *Buchnera* (panel i) forming the extremes in the data that we examined. At least for these species, and possibly also for *Aeropyrum* (panel d), the relative amount of proteins in the individual peaks appears to reflect the environment for which the species are adapted as their extreme pI distribution correspond to extreme environments and all of these organisms can be classified as extremophiles. The adaptation of the extreme halophilic archaeon *Halobacterium sp.* NRC-1 to its environment has been analysed by Kennedy *et al.* [12]. On the other extreme of the distribution spectrum is the eubacterium *Buchnera sp.* Although a close relative to *Escherichia coli* [13], which has a pI distribution similar to the one given in Fig. 1b, *Buchnera* is an obligate endocellular symbiont that can only survive in the bacteriocytes of aphids. The extremely high average pI (pH = 9.6) of this species has previously been observed by Shigenobu *et al.* [14] and has been attributed to the high lysine usage of the species. *Aeropyrum* is an extreme thermophile archaeon growing at temperatures of up to 100°C [15]. It is however, not clear whether the fourth and alkaline peak (pH 12) observed in this species is associated with heat resistance, as such a peak is absent from other thermophiles like the archaeon *Methanococcus janaschii* (Fig. 1g).

While the relative abundance of proteins forming the individual peaks differs in different organisms, the pH of the peaks and especially of the troughs remains constant,

e.g., there is always a trough at pH 7.4 and at pH 8.0. We were interested in whether this characteristic shape of pI distribution reflects genetic evolution. Have all proteins evolved to be charged at physiological pH as has been suggested? There are two impediments to answering this question; first we do not know the pI of native proteins, as theoretical calculations can only estimate the pI of denatured polypeptides. Further, the actual physiological pH cannot be measured *in vivo*, and exact estimates therefore, do not exist. However, it may be reasonable to assume that the physiological pH is often close to neutrality and that the distribution of native protein pI/s follows the distribution of denatured polypeptides. Have sequences evolved to avoid pI values coinciding with the observed troughs?

To answer these questions we have analysed the pI distribution of randomly generated protein sequences and found that the distributions closely follow the distributions of actual sequences. Clearly, nucleotide composition as well as amino acid composition varies widely between organisms, and this will result in differences in the average pI of the proteins. In simulations that model the spread and variation of sequence length as well as amino acid compositions of real organisms, random sequences have produced the same pI distribution as actual sequences. We conclude therefore that sequences have not evolved to conform to a specific pI, but rather the observed pI distribution is a statistical consequence that follows from the chemical properties of the amino acids that make up the proteins. Below we give a statistical explanation as to why the modal distribution of pI values is to be expected without assuming any selective pressures.

3.2 Distribution of estimated protein pI in the absence of evolutionary constraints

By definition, the pI of a protein is the pH at which the positive charge is equal to the negative charge. At this pH, the sum of charges, over all amino acids, is null. In proteins there are four ionisable groups that can assume positive charges. These are the three amino acids lysine (K), arginine (R) and histidine (H) as well the N-terminus (N-term). Negative charges can be assumed by the four amino acids tyrosine (Y), cysteine (C), aspartate (D) and glutamate (E) and the C-terminus (C-term) of the protein. Let f^+ be the sum of all positive charges and f^- be the sum of all negative charges of a given protein. Both values depend on the pH and the pK values of the ionisable groups in the protein. We define $I^+ = \{K, R, H, N\text{-term}\}$ as the set of amino acids charged positively and $I^- = \{Y, C, D, E, C\text{-term}\}$ as the set of amino acids charged negatively.

Hence, the positive charge f^+ of a given protein and its negative charge f^- can be defined by:

$$f^+ = \sum_{i \in I^+} n_i f_i^+ \quad \text{and} \quad f^- = \sum_{i \in I^-} n_i f_i^- \quad \text{Eq. 1}$$

where f_i^+ is the elementary charge assumed by an amino acid of type i , and n_i is the number of amino acids of type i in the current protein. We have $n_{N\text{-term}} = n_{C\text{-term}} = 1$. By definition, the pI value of a protein is the solution of the Eq.

$$f^-(pH) + f^-(pH) = 0 \quad \text{Eq. 2}$$

For a given amino acid, f_i^+ or f_i^- is estimated from Henderson-Hasselbach's Eq. by:

$$f_i^+(pH) = \frac{1}{1 + 10^{pH - pK_i}}$$

$$f_i^-(pH) = f_i^+(pH) - 1 \\ = \frac{10^{pH - pK_i}}{1 + 10^{pH - pK_i}}$$

where $0 \leq f_i^+(pH) \leq 1$ and $0 \leq f_i^-(pH) \leq 1$.

All f_i^+ functions have the same shape and can be deduced from another one by translation. The same is true for the f_i^- functions. They are a sigmoid function of pH with an inflection point at $pH = pK_i$. This point is a symmetric point where

$$f_i^+(pK_i) = 1/2 \quad \text{or} \quad f_i^-(pK_i) = 1/2.$$

In each function f_i^+ decreases with increasing pH , ranging from 1 to 0. Reciprocally, f_i^- is an increasing function of pH , ranging from 0 to 1. The f_i^+ curve has a relatively steep zone extending about 1.0 pH unit on either side of the pK_i . This zone is the well-known buffering region as large changes in the charge result in only slight variations of pH . The entire graph resembles a step centred at the pK_i .

From these properties, and from Eq.(1), the positive charge f^+ is a descending stair function (Fig. 2) varying from

$$n_{I^+} = \sum_{i \in I^+} n_i$$

to 0.

Each step is associated with an amino acid and is centred on its pK value. The height of each step is n_i and depends only on the frequency of the associated amino acid in the protein. Note that the position (pH) of each step is independent of the protein!. Only its height (charge) changes depending on the amino acid composition.

For amino acids with close pK values the steps are smoothed forming a single step as the buffering region of both amino acids intersects (e.g., K and R in Fig. 2).

Symmetrically, and for the same reasons, f^- is a rising stair function varying from 0 to

$$n_{I^-} = \sum_{i \in I^-} n_i.$$

The pI value is the abscissa of the intersection point between the ascending and descending stair graphs (Fig. 3).

3.3 The influence of the shape of f^+ and f^- on pI position

Equation 2 does not have a simple solution and it would be very difficult to study the relationship between pI and the amino acid composition analytically. Consequently, we followed the influence of each amino acid graphically and used simulation to exhibit the distribution of pI around a set of random sequences. First, it is easy to establish from Fig. 3 that each amino acid charge does not have the same effect on pI values. If the number of histidines (n_H) increases in a sequence, the height of the associated

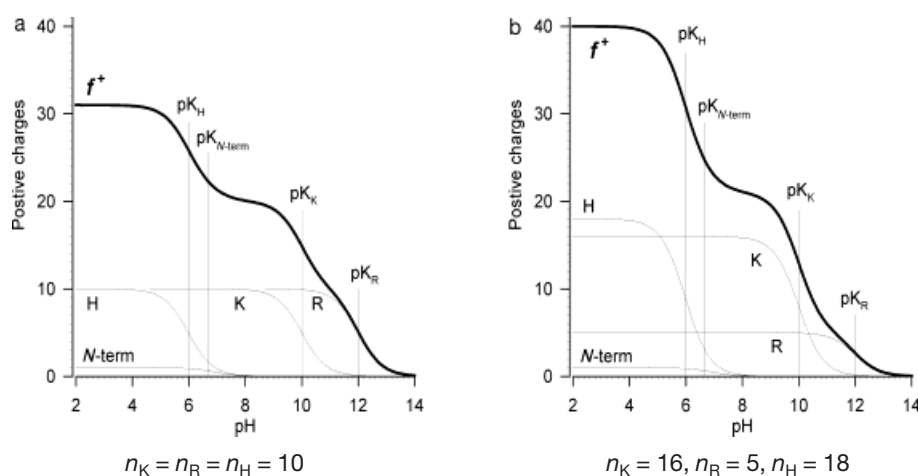


Figure 2. Positive charge curve (f^+) as a function of pH . The additive contribution of each amino acid is given in thin lines and their numbers (n_K , n_R , n_H) are indicated for part a) and b) of the Figure.

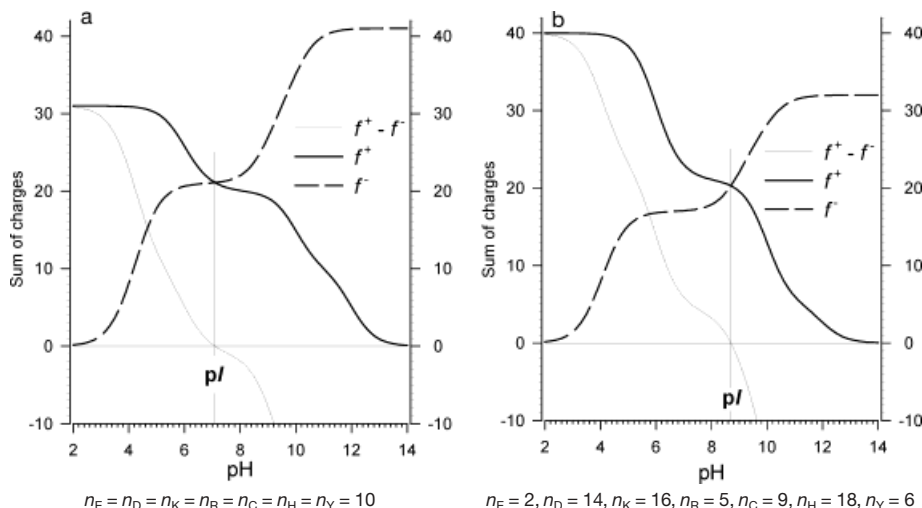


Figure 3. Positive (f^+) and negative (f^-) charge curves. The pI value is the abscissa of the intersection point between f^+ and f^- , e.g. where the sum of charges ($f^+ - f^-$) is 0. The numbers of each amino acid ($n_E, n_D, n_K, n_R, n_C, n_H, n_Y, n_H$) are indicated for part a) and b) of the Figure.

(left most) step in f^+ increases accordingly. This does not significantly modify the curve to the right hand side of this step. In Fig. 3 for instance, the number of histidines (n_H) changes from 10 (a) to 18 (b). Note that this modification does not alter the pI value by much. In general, an amino acid with a positive charge and a low (acidic) pK does not normally have a large effect on the pI. For the same reasons, an amino acid with a negative charge and a high (alkaline) pK does not have a large effect on the pI.

In contrast, a change in the occurrence of an amino acid with positive charge and high pK, or an amino acid with negative charge and a low pK, have a strong effect on the pI of a protein. In Fig. 3, the value of n_K increases from 10 to 16, but at the same time, n_R decreases from 10 to 5. Both amino acids are associated with the same step of the f^+ curve. The total height of the step associated with these amino acids grows slightly and the pI moves slightly to the right. The stair shapes of the f^+ and f^- functions explain the uneven effect that a change in the number of charged ionic groups has on the resulting pI of the protein. If the pI appears in a zone where the f^+ curve is relatively flat, a small change in f^- has a larger effect on the pI than if f^+ is in a buffer zone (near a pK value). The reverse is true for f^- . For example, in Fig. 3a the pI is 7.1 and the f^- curve is relatively flat in this zone:

$$\frac{\partial f^-(pI)}{\partial pH} = 0.92.$$

If f^+ increased by one charge unit, the pI would increase to 7.8 ($\Delta_{pI} = +0.7$).

By contrast, in Fig. 3b the pI is 8.7 and f^- curve is growing in this zone:

$$\frac{\partial f^-(pI)}{\partial pH} = 5.82.$$

If f^+ would increase by one charge unit, the pI would only raise to 8.8 ($\Delta_{pI} = +0.1$).

Thus, the density distribution of pI values must be low on the areas where f^+ or f^- are flat. For example, pI distribution should be particularly small between 7 and 8, where f^+ and f^- are both relatively flat.

3.4 Simulated pI distributions

To test this prediction let us consider that the amino acid composition of a sequence is the result of a random process. Each n_i is now a random variable with a specific distribution, and the pI therefore represents a random value. We are interested by its distribution but as we do not know the exact distribution of n_i for a set of similar proteins we cannot know the distribution of the pI without further assumptions. For a fixed sequence length

$$l = n_{l^+} + n_{l^-} = \sum_{i \in \{l^+, l^-\}} n_i,$$

we suppose that the n_i are coming from a multinomial distribution with l trials and p_i probabilities. Thus, for a fixed amino acid, the mean and SD of the number of its occurrences are:

$$v_i = E(n_i) = l p_i \quad \text{and} \quad \sigma_i = \sqrt{V(n_i)} = \sqrt{l p_i (1 - p_i)}.$$

We have to choose values for parameters l and p_i . To get realistic values for l we modelled a set of sequence lengths on an existing proteome (we chose *Sinorhizobium meliloti*), and randomly chose l from this set in our simulations. For p_i we explored two cases. For the first one, we studied the theoretical case where $v_i = \text{Cst}$ (Fig. 3a), i.e.

where the probability of occurrence is the same for each ionisable amino acid. For the second case, we chose an heterogeneous set of values for p_i given in Fig. 3b.

Furthermore, from Eq. (1) we can see that f^+ and f^- are linear expressions of the f_i^+ and f_i^- functions. In these expressions each f_i^+ and f_i^- are weighted by n_i . As these weights are integers the pI distribution reflects a discontinuous phenomenon when amino acid distributions change. This discontinuity complicates our study by introducing local modifications in pI distribution. Then, in a first step, we chose to eliminate discontinuity effects by considering that n_i has a continuous value. We simulate n_i by the normal approximation of a multinomial distribution. We suppose that

$$n_i \equiv N\left(v_i = l p_i, \sigma = \sqrt{l p_i (1 - p_i)}\right).$$

We obtain an excellent approximation of multinomial distribution if we round n_i to the nearest integer. Without rounding n_i , we find a smooth distribution for pI with two peaks

(Fig. 4) for each case studied. We also confirm our initial prediction by observing that the pI distribution is small for a pH between 7 and 8 (relatively flat zone for f^+ and f^-).

If we round n_i , obtained with normal approximation, or if we use a multinomial generator, we reintroduce n_i discontinuity. Then f^+ and f^- fluctuate across a finite set and the pI cannot take all pH values. Some of them are impossible, some are more frequent values. A region where we can find the effects of these discontinuity points is for a pH between 7 and 8. In this zone the pI is more sensitive. As we discuss below, any modification of f^+ or f^- has an amplified effect on the pI . This is the reason why, in Fig. 5, a multimodal distribution arises. Between pH 7 and 8 we observe a zone with a low pI value distribution as well as a zone with a local peak. This peak corresponds to the third peak found in the pI distribution of most organisms and is most apparent in large protein sets. This peak does not have any biological significance but arises from the discrete definition of f^+ and f^- across n_i .

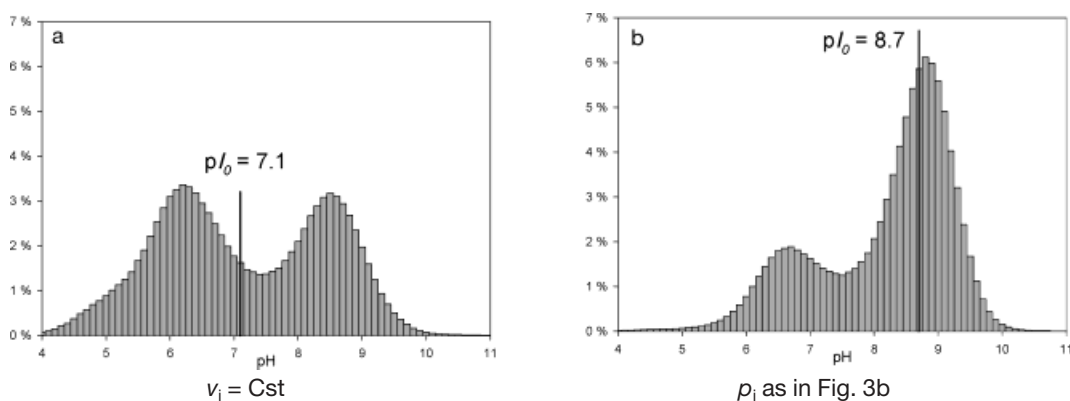


Figure 4. Simulated pI distribution using a continuous normal distribution perturbation. A line indicates pI_0 , the pI of a sequence with $n_i = v_i$, see section 3.4.

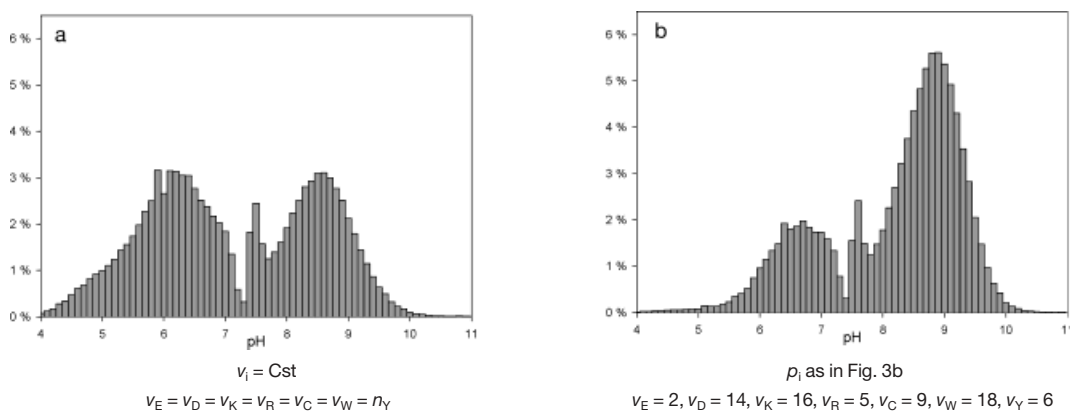


Figure 5. Distribution of pI with a multinomial distribution, see section 3.4.

4 Concluding remarks

Positive and negative charges of a protein are respectively, descending and ascending functions of the pH. The underlying sigmoid shape of the pK curves of the individual chargeable groups causes both functions to assume an uneven stair-like shape. The intersection of the curves represents the protein's pI and is more likely to occur at a pH where the curves are steep. Consequently, some pI values are attained by a large number of different compositions of charged amino acids while combinations of amino acids resulting in other pI values can hardly be obtained. An uneven distribution of pI values must therefore be expected when a sufficiently large number of proteins are examined. Our simulations show that the distribution of randomly generated pI distributions concur with the multimodal distributions of pI values observed in all organisms. Most proteins form part of one of the two major clusters with peaks at pH 5.5 and 9.5 respectively, although minor peaks can be detected at pH 7.8 and pH 12. Note that the exact positions of these peaks depend on the pK values assumed for the pI calculations. This characteristic location of the distribution peaks is observed in the organisms of all kingdoms and is also found in chloroplasts, mitochondria and viruses. However, depending on the amino acid usage, the number of proteins that form part of the different clusters differ in different organisms. In extreme cases the acidic or alkaline cluster may be missing almost entirely. These unusual distributions have only been found in extremophiles, and almost certainly reflect the extreme environment in which these organisms live.

5 References

- [1] Svensson, H., *Acta Chem. Scand.* 1962, 16, 456–466.
- [2] Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N. *et al.*, *Electrophoresis* 1993, 14, 1023–1031.
- [3] Bjellqvist, B., Basse, B., Olsen, E., Celis, J. E., *Electrophoresis* 1994, 15, 529–539.
- [4] Gianazza, E., Righetti, P. G., *J. Chromatogr.* 1980, 193, 1–8.
- [5] Urquhart, B. L., Atsalos, T. E., Roach, D., Basseal, D. J. *et al.*, *Electrophoresis* 1997, 18, 1384–1392.
- [6] VanBogelen, R. A., Schiller, E. E., Thomas, J. D., Neidhardt, F. C., *Electrophoresis* 1999, 20, 2149–2159.
- [7] Arakawa, T., Timasheff, S. E., *Methods Enzymol.* 1985, 114, 49–77.
- [8] Schwartz, R., Ting, C. S., King, J., *Genome Res.* 2001, 11, 703–709.
- [9] <ftp://ncbi.nlm.nih.gov/genbank/genomes/>
- [10] <http://www-nbrf.georgetown.edu/pirwww/search/genome.html>
- [11] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., *Numerical Recipes in C: The art of Scientific Computing*, Cambridge University Press, Cambridge, England 1992, pp. 280–289.
- [12] Kennedy, S. P., Ng, W. V., Salzberg, S. L., Hood, L., Das-Sarma, S., *Genome Res.* 2001, 11, 1641–1650.
- [13] Baumann, P., Moran, N. A., Baumann, L., in: Dworkin, M. (Ed.), *The Prokaryotes*, Springer, New York, USA 2000.
- [14] Shigenobu, S., Wantanabe, H., Hattori, M., Sakaki, Y., Ishikawa, H., *Nature* 2000, 407, 81–86.
- [15] Sako, Y., Nomura, N., Uchida, A., Ishida, Y. *et al.*, *Int. J. Syst. Bacteriol.* 1996, 46, 1070–1077.