Weak Acids and Bases and Isoelectric Points of Proteins*

Many beginning students of biochemistry often have tremendous difficulty learning and applying concepts of acid-base chemistry to biomolecules such as amino acids, peptides, and proteins. This situation is unfortunate because an understanding of these concepts helps the students secure a solid foundation from which a deeper appreciation of many aspects of biochemistry can be attained. For example, a thorough understanding of the acid-base chemistry of biomolecules allows one to predict the order of elution of amino acids, peptides, or proteins from an ion-exchange chromatography column, to understand and appreciate general acid-base catalysis of enzymes, to appreciate the resolution that techniques such as 2-D electrophoresis provides, and to appreciate ways in which regulatory proteins such as hemoglobin change in response to pH changes in their environment.

The teaching of these concepts can, of course, be approached in a number of ways, and the authors of two previous papers in this *Journal* (*1*, *2*) present two very good but different approaches to teaching these concepts. In the first paper (*2*), the author discusses the use of a Maple program to calculate the charge of a peptide at different pH values. This approach, which requires the use of a commercially available program, is somewhat obviated by the widespread availability of freely accessible programs on the Internet that will make the same calculations. The use of these free programs, however, has the potential to obscure the concepts being taught because the students do not know what the algorithm is doing, and the process thus becomes somewhat of a “black box”. The authors of the other paper (*1*) describe the use of a graphical approach to calculate the isoelectric point (pI) and charge of small peptides. This latter approach works very well, but it is limited to small peptides and thus would quickly become unwieldy if larger polypeptides or proteins were used.

In the work reported herein, an alternative approach is presented that focuses on the use of a spreadsheet program such as Microsoft Excel to help teach students some of the fundamental concepts of acid-base chemistry, especially as applied to important biomolecules such as amino acids, peptides, and proteins. The use of a spreadsheet allows students to enter their own equations, which requires an understanding of the underlying theory, and consequently the results of the calculations are likely to be more apparent than they otherwise would be if a computer program is used. In addition, the use of a spreadsheet potentially enhances student learning because the students are given the freedom to explore what happens when, say, they use different pK_a values in the equations because different sources list slightly different values, or because they have learned (or will learn) that the pK_a values of the side chains of some amino acids (when incorporated into a peptide or protein) are known to be significantly different from the pK_a values of the respective side chains when the amino acid is free in solution. As the students use the spreadsheet to perform the calculations, they can create figures to help them visualize what might otherwise be an abstract concept. In essence, the students are "in control" of their learning, which generally is considered to lead to more effective understanding (*3*). Finally, since spreadsheet programs such as Excel are readily accessible at educational institutions, their use does not require the purchase of additional software.

In this article, I first discuss the acid-base properties of relatively simple compounds that have one ionizable functional group, and I show how a spreadsheet can be used to help enhance one's understanding of the acid-base chemistry of these relatively simple weak acids. Next, I discuss the slightly more complicated amino acids and show how a spreadsheet can be used to calculate the pI of an amino acid by calculating the net charge of an amino acid as a function of pH. I then discuss how an understanding of the acid-base properties of amino acids can be

applied to oligopeptides and proteins, and I show how a spreadsheet can be used to make similar calculations with these molecules. Finally, the results obtained with a spreadsheet are compared with those obtained via programs that can be found on the internet. These comparisons reveal that both the spreadsheet approach and the Internet-based program approach provide similar or essentially identical estimates of the pIs of simple proteins (i.e., those composed of one subunit); however, when more complex proteins (i.e., those made up of two or more, possibly different subunits) are treated, the limitations of the Internet-based programs become apparent as they are not designed to estimate the pIs of complex proteins. Instead, the spreadsheet approach is shown to be flexible enough to allow reasonable estimates (as judged by closeness to experimentally determined values) of the pIs of such complex proteins. It is hoped that the spreadsheet approach described herein will prove useful to other teachers and students of biochemistry.

Weak Acids

We begin our discussion of the acid-base properties of weak acids by considering a compound that has a carboxyl functional group. For now, we will not specify the rest of the molecule; instead, we will simply refer to the rest of the molecule by using a capital R, as is often done in organic chemistry. Since a carboxyl group can be protonated, our example molecule could be represented as RCOOH. But a carboxyl group can lose a proton, and therefore our example compound also could be represented as $RCOO^-$. So when our compound is protonated it is neutral; when it is deprotonated it carries a negative charge. We might wonder about the best way to represent our compound when it is dissolved in an aqueous solution at relatively low pH, say pH = 2. Would most of the molecule be in the protonated, neutral form? Or, would most of the molecule be in the unprotonated, negatively-charged form? What about the situation at relatively high pH, say pH = 13, or at intermediate pH values? To answer these

questions we need to consider some basic acid-base chemistry, and we need to perform some algebra. Since our compound is a weak acid, we represent its dissociation by the following reaction:

$$\mathrm{RCOOH} \rightleftharpoons \mathrm{H^+ + RCOO^-} \tag{1}$$

The left-hand side eq 1 should contain a water molecule, and the right-hand side of eq 1 should contain a hydronium ion in place of a proton, but these species have been omitted for the sake of clarity. The equilibrium constant (or acid dissociation constant) for the reaction expressed by eq 1 is:

$$K_a = \frac{[\mathrm{H^+}][\mathrm{RCOO^-}]}{[\mathrm{RCOOH}]} \tag{2}$$

where K_a is the acid dissociation constant. Technically, eq 2 should contain a term that allows the equilibrium constant to be dimensionless. Eq 2 also should contain the activities of the species on the right-hand side of the equation, or the concentrations on the right-hand side of the equation should be multiplied by activity coefficients. So, an unstated assumption of eq 2 is that the activity coefficients are all unity, which generally is reasonable for dilute solutions of weak acids. If we divide both sides of eq 2 by ($[\mathrm{H^+}]K_a$) we obtain the following:

$$\frac{1}{[\mathrm{H^+}]} = \frac{1}{K_a}\left(\frac{[\mathrm{RCOO^-}]}{[\mathrm{RCOOH}]}\right) \tag{3}$$

Taking the log of both sides of eq 3 yields:

$$\log\frac{1}{[\mathrm{H^+}]} = \log\left(\frac{1}{K_a}\left(\frac{[\mathrm{RCOO^-}]}{[\mathrm{RCOOH}]}\right)\right) \tag{4}$$

Since $\log(xy) = \log x + \log y$ and $\log(x/y) = \log x - \log y$, we can rewrite eq 4:

$$\log 1 - \log[\mathrm{H^+}] = \log 1 - \log K_a + \log\left(\frac{[\mathrm{RCOO^-}]}{[\mathrm{RCOOH}]}\right) \tag{5}$$

Finally, we simplify eq 5 by removing the removing the "log1" terms (since log1 = 0) and substituting in the working definitions that pH = $-\log[H^+]$ and $pK_a = -\log K_a$:

$$\mathrm{pH} = pK_a + \log\frac{[\mathrm{RCOO}^-]}{[\mathrm{RCOOH}]} \tag{6}$$

Eq 6 is a form of the Henderson-Hasselbalch equation, which has been modified to apply to our example. The inherent assumptions of eq 2 mentioned above are carried over to eq 6. Thus, useful working definitions such as pH = $-\log[H^+]$ are understood to be approximate and not strictly true. Since we are interested in the best way to represent our compound at different pH values, we see that if we isolate the ratio $[RCOO^-]/[RCOOH]$ we can make progress in this regard:

$$\frac{[\mathrm{RCOO}^-]}{[\mathrm{RCOOH}]} = 10^{(\mathrm{pH}-pK_a)} \tag{7}$$

If the pK_a of our compound is known then we can calculate the ratio of the unprotonated, negatively charged form to the protonated, neutral form at any pH value. For example, if the pK_a of our compound is 5.0, then at pH = 2.0 the above ratio is $10^{(2.0-5.0)} = 10^{-3.0} = 0.001$ or 1/1000. So at pH = 2, approximately one molecule of our compound is in the negatively charged carboxylate form ($RCOO^-$) and 1000 are in the neutral, carboxyl form (RCOOH). (The fraction of molecules in the carboxylate form and the fraction of molecules in the carboxyl form are 1/1001 and 1000/1001, respectively, which is an important point that will be elaborated upon shortly.) We could repeat the above calculation for the case where the pH = 13, or for cases where the pH has some intermediate value. We see that it becomes somewhat repetitive at this point, and this is precisely the type of problem for which a spreadsheet is particularly well-suited. The steps outlined below are specific for Microsoft Excel, but other spreadsheet programs should work in a similar manner.

An Excel spreadsheet that one might construct is shown in Figure 1. This spreadsheet is constructed by entering a zero in cell A2, and then selecting cell A3 and typing the "=A2+1". When the "Enter" key is pressed, the value "1" appears in this cell. (We could use a smaller increment, and in some of the later examples we will find it necessary to do so.) To rapidly fill in this column with increasing values to 14 (the highest pH that we are considering), we place the cursor at the lower right corner of cell A3 and note that the appearance of the cursor changes from an open plus sign to a solid plus sign. When we see the solid plus sign we drag the cursor down to "fill in" the remaining values of column A. The values in column A increase by increments of one because the function we typed in cell A3 (i.e., =A2+1) made cell A2 a relative reference. Thus, when we position the cursor at the lower left corner of cell A3 and drag the cursor down to cell A4 we can see in the formula bar "=A3+1".

(The formula bar is not shown in Figure 1, but it is adjacent to the symbol "f_x" in an Excel sheet.) The subsequent cells fill in accordingly until we reach the desired value (in this case 14). Next, we select cell B2 and type the function "=10^(A2–5.0)", which effectively says that the value of the cell will equal 10 raised to the power that is the difference of the pH value in the adjacent cell (A2) minus the pK_a value (5.0),

	A	B	C	D
1	pH	ratio	fraction	(-1)*fraction
2	0	0.00001	9.9999E-06	-9.9999E-06
3	1	0.0001	9.999E-05	-9.999E-05
4	2	0.001	0.000999001	-0.000999001
5	3	0.01	0.00990099	-0.00990099
6	4	0.1	0.090909091	-0.090909091
7	5	1	0.5	-0.5
8	6	10	0.909090909	-0.909090909
9	7	100	0.99009901	-0.99009901
10	8	1000	0.999000999	-0.999000999
11	9	10000	0.99990001	-0.99990001
12	10	100000	0.99999	-0.99999
13	11	1000000	0.999999	-0.999999
14	12	10000000	0.9999999	-0.9999999
15	13	100000000	0.99999999	-0.99999999
16	14	1E+09	0.999999999	-0.999999999
17				
19				
19	ratio = [RCOO$^-$]/[RCOOH]			
20	fraction = [RCOO$^-$]/([RCOO$^-$] + [RCOOH])			

Figure 1. Spreadsheet used for calculating various parameters (e.g., ratio of carboxylate to carboxyl form) as a function of pH.

as shown in eq 4. When we press the "Enter" key we see this value, which corresponds to the ratio of the carboxylate form to the carboxyl form at pH = 0, which is 0.00001. Next, we fill-in column B as described above for column A, and we see that the value of the ratio increases dramatically to 1×10^9 at pH = 14.

To visualize this increase in the ratio as a function of pH we can use Excel to construct a graph (which Excel calls a "chart") as shown in Figure 2A. We use Excel to construct such a graph by highlighting the two columns of data (columns A and B in Figure 1) and then selecting "Insert", "Chart", "XY (Scatter)", and then "Next". When we choose these options we step through a series of windows where, among other choices, we can specify or verify that our data are arranged in columns as opposed to rows, we can label the chart and/or label the axes, and finally we can decide where we want Excel to place our newly constructed chart (the default is an object in the current worksheet). An inspection of Figure 2A shows that the value of the ratio of the carboxylate form to the carboxyl form increases exponentially, which is not surprising since the pH

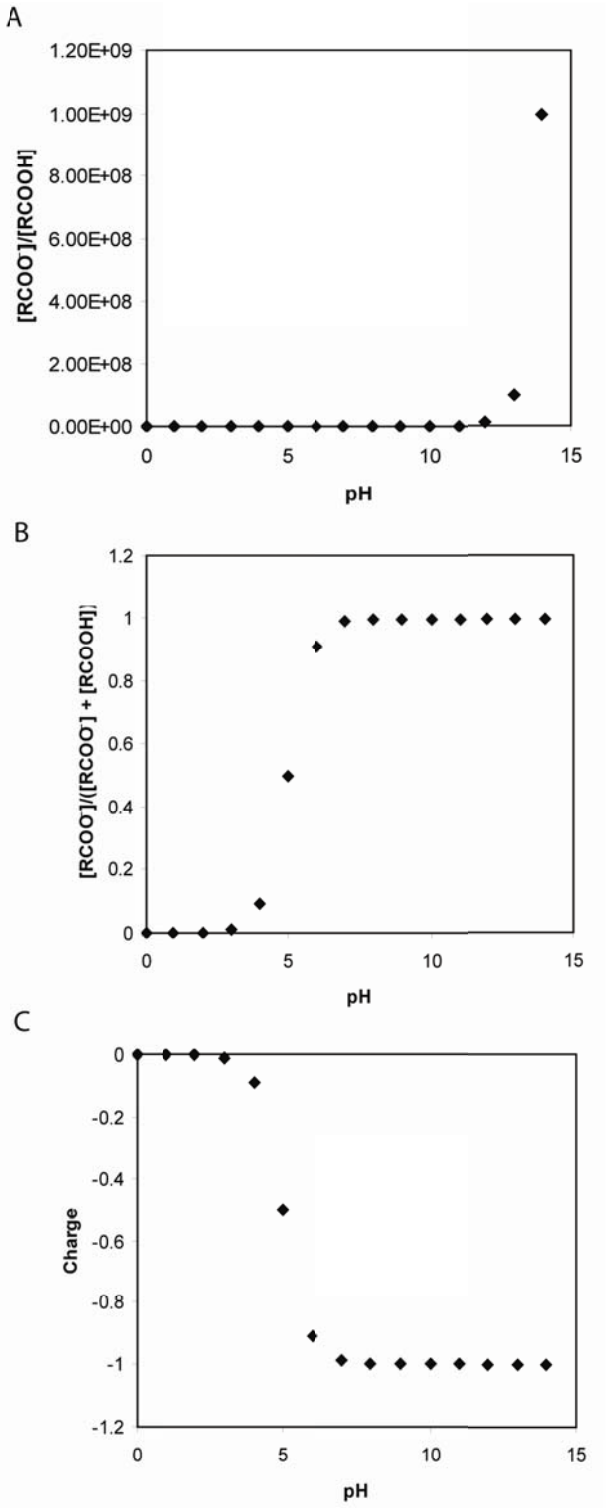


Figure 2. (A) ratio of carboxylate to carboxyl form as a function of pH; (B) fraction of carboxylate form as a function of pH; (C) charge as a function of pH.

scale is a log scale. To fit all of the data of columns A and B of Figure 1 onto the graph shown in Figure 2A, the value of the *y*-axis had to increase in increments of 2×10^8. Thus all of the increases in the above ratio (up to pH = 12) appear flat on the graph shown in Figure 2A. Consequently, our graph is not as informative as it might be, and although we could change the scale of the ordinate axis of Figure 2A to a log scale, it is preferable in this case to think of another way to represent the compound. One way that we can represent the compound is to determine what fraction of the compound is in the carboxylate form at a particular pH value. Given the dramatic increase of the ratio of the carboxylate form to the carboxyl form with increasing pH, we expect that the fraction of carboxylate form to total possible forms (the sum of the carboxylate forms and carboxyl forms) should approach one as pH increases. This concept that the ratio of one form to the other form is not the same as the fraction of one form to the total of possible forms is, in my view, the key to unlocking much of the difficulty with acid-base concepts applied to biomolecules. For example, in cell B7 of Figure 1 we see that the ratio of carboxylate to carboxyl forms is one at pH = 5. This result makes sense when we recall eq 7, which indicates that the value of $[RCOO^-]/[RCOOH]$ is $10^{(pH - pKa)}$. In our present example, this difference is zero and thus the ratio equals one: $10^{(5.0 - 5.0)} = 10^0 = 1$. The fraction of carboxylate to total possible forms, however, is one part carboxylate to two total parts: 1/(1 + 1) = 1/2 = 0.5, which is the value seen in the adjacent cell (C7) under the column heading “fraction” (column C) in Figure 1. We create this column in our spreadsheet by selecting cell C2 and typing “= 10^(A2–5.0)/(10^(A2–5.0) + 1)”, which converts our ratio to the desired fraction. Now, when we press the “Enter” key the value we see corresponds to the fraction of carboxylate forms to total possible forms at pH = 0 . We again fill in the remainder of column C as described above, and we see that the value of the fraction approaches unity at higher pH values as we anticipated.

We construct a plot to visualize these data by highlighting column A (still our independent variable) and then, while holding down the “Control” key we highlight column C and then select “Insert”, “Chart”, etc… as described above. The resulting graph is shown in Figure 2B, and it is an improvement over the graph shown in Figure 2A. We can improve things further, however, if we plot the charge of our compound as a function of pH. To make this change, we select cell D2 and type “= (–1)*10^(A2–2.35)/(10^(A2–2.35) + 1)”, which is the same function statement used in cell C2 but the difference is that now we multiply this function statement by negative one. We now see that as the fraction of carboxylate form approaches unity, the charge approaches negative one, which is consistent with the picture that we might construct in our minds. As more and more of the carboxyl groups become unprotonated, more and more carry a full negative charge. A plot showing the charge on the molecule as a function of pH is shown in Figure 2C. This graph is constructed by highlighting columns A and D and selecting “Insert”, “Chart”, etc… as described above.

A standard question that a beginning student might encounter regarding the compound under consideration is: What is the average, net charge of the compound at pH = 5? The student is now in a position to answer such a question because he or she can look at cell C7 of Figure 1 and realize that half the molecules are in the carboxylate form and thus have a charge of minus one, and half the molecules are in the carboxyl form and thus have a charge of zero. So the average, net charge is 0.5 [$(0.5 \times -1) + (0.5 \times 0) = 0.5$]. The above question is very typical of questions that often are asked about more complicated molecules such as amino acids. We will consider such questions after we discuss amino acids, and it is hoped that the beginning student will see that these questions are very answerable, especially with the help of a spreadsheet when one is first learning the subject.

Before we finish with the above example, it is instructive to note that at pH = 7 the ratio of carboxylate to carboxyl forms is 100 (cell B9, Figure 1) and the fraction of carboxylate to total possible forms at this same pH is 100/101 ≈ 0.99009901 (cell C9, Figure 1). Similarly at pH = 3 the ratio of carboxylate to carboxyl forms is 1/100 or 0.01 (cell B5, Figure 1), but the fraction of carboxylate to total possible forms is 1/101 ≈ 0.00990099 (cell C5, Figure 1). These observations tell us that once the pH that we are considering is either two units above or below the pK_a of a particular group, then we often can make the simplification that almost all of the compound is either unprotonated (for cases when the pH is two units above the pK_a) or protonated (for cases when the pH is two units below the pK_a). Mathematically, it means that we can make the approximation that the fraction of the compound in the carboxylate form is ~1 when the pH is two units higher than the pK_a, and the fraction of the compound in the carboxyl form is ~1 when the pH is two units lower than the pK_a. The approximations improve when the separation between pH and pK_a is greater than two units (column C, Figure 1). To validate these claims, we consider the following problem that is representative of the type that might be asked concerning a weak acid: A weak acid (HA) has one ionizable group, and the pK_a of this group is 5.0. If one has 100 mL of a 0.1 M solution of this compound at an initial pH = 7.0, what is the final pH after 20 mL of 0.1 M HCl are added?

To answer this question, we consider the starting conditions and realize that since the initial pH is two units higher than the pK_a of the compound, we can make the assumption that essentially all of the ionizable groups of the compound are in the unprotonated (conjugate base or A^-) form. Since we have 0.01 mol of the compound (0.1 mol/L × 0.1 L = 0.01 mol), we assume that we have 0.01 mol of the A^- or unprotonated, conjugate base form. We add 0.002 mol of H^+ (from the completely dissociating HCl). Consequently, we will protonate 0.002 mol of

the 0.01 mol of the A^- form. This reaction forms 0.002 mol of the HA form. We still have 0.008 mol of the A^- form left, and we insert this ratio into the Henderson-Hasselbalch equation and solve for pH as follows:

$$\mathrm{pH} = \mathrm{p}K_\mathrm{a} + \log\frac{[\mathrm{A}^-]}{[\mathrm{HA}]} \Rightarrow \mathrm{pH} = 5.0 + \log\frac{0.008}{0.002} \approx 5.602 \approx 5.6$$

If we do not make the above approximation, we calculate that at the initial pH of 7.0 we have ~$9.900990099 \times 10^{-5}$ mol of HA already present (1/101 × 0.01 mol). When we add the 0.002 mol H^+, we form an additional 0.002 mol of HA. So our adjusted mol of HA is ~$2.099009901 \times 10^{-3}$ (0.002 mol + $9.900990099 \times 10^{-5}$ mol). At the initial pH of 7.0 we have ~$9.900990099 \times 10^{-3}$ mol A^- present (100/101 × 0.01 mol), but this amount decreases by 0.002 mol to ~$7.900990099 \times 10^{-3}$ mol when the 0.002 mol H^+ are added. When we use these values to insert a new ratio into the Henderson-Hasselbalch equation and solve for pH we get the following:

$$\mathrm{pH} = \mathrm{p}K_\mathrm{a} + \log\frac{[\mathrm{A}^-]}{[\mathrm{HA}]} \Rightarrow \mathrm{pH} = 5.0 + \log\frac{7.900990099\times10^{-3}}{2.099009901\times10^{-3}} \approx 5.576 \approx 5.6$$

Thus, we see that the answer that we get when we use the above approximation is within 0.5 % of the answer that we get when we do not use the above approximation. The former approach is much easier and less time-consuming than the latter approach and, to the level of accuracy that often is required (and justified given the inherent approximations in the Henderson-Hasselbalch equation), it provides the same answer as latter approach.

We now move on to our second example molecule. This new molecule is an amine, which we represent as RNH_2. Since amines are weakly basic, our compound can at times be protonated and therefore represented as RNH_3^+. So, when our compound is unprotonated it is neutral, but when it is protonated it carries a positive charge. This situation is somewhat opposite

of what we encountered with the carboxyl compound, but we can still proceed in a similar manner. Again, we consider the best way to represent our compound when it is dissolved in aqueous solution at some particular pH. We follow the same line of reasoning as before and obtain the following:

$$RNH_3^+ \rightleftharpoons H^+ + RNH_2 \tag{8}$$

$$K_a = \frac{[H^+][RNH_2]}{[RNH_3^+]} \tag{9}$$

$$\frac{1}{[H^+]} = \frac{1}{K_a}\left(\frac{[RNH_2]}{[RNH_3^+]}\right) \tag{10}$$

$$\log\frac{1}{[H^+]} = \log\left(\frac{1}{K_a}\left(\frac{[RNH_2]}{[RNH_3^+]}\right)\right) \tag{11}$$

$$\log 1 - \log[H^+] = \log 1 - \log K_a + \log\left(\frac{[RNH_2]}{[RNH_3^+]}\right) \tag{12}$$

$$pH = pK_a + \log\frac{[RNH_2]}{[RNH_3^+]} \tag{13}$$

$$\frac{[RNH_2]}{[RNH_3^+]} = 10^{(pH - pK_a)} \tag{14}$$

Eq 14 provides the ratio of the basic, amino form to the acidic, ammonium form, but based on the previous example we anticipate that eventually we will want the fraction that is in the ammonium form since this fraction, when multiplied times positive one, will tell us the charge at a particular pH. So, it is convenient to rearrange eq 14 as follows:

$$\frac{[RNH_3^+]}{[RNH_2]} = 10^{(pK_a - pH)} \tag{15}$$

If we assume that the pK_a of the ammonium form of our compound is 10.0, then we can construct a spreadsheet in which we calculate the ratio of ammonium form to amino form, the fraction of ammonium form to total possible forms, and one times this fraction at pH values ranging from zero to 14. This spreadsheet is shown in Figure 3 and is constructed in the same manner as described above for the spreadsheet shown in Figure 1, but the function statements in cells B2 and C2 are "=10^(10.0–A2)" and "=10^(10.0–A2)/(10^(10.0–A2) + 1)", respectively. (The function statement in cell D2 of Figure 3 is identical to that of cell C2 and only is included for completeness.)

	A	B	C	D
1	pH	ratio	fraction	(+1)*fraction
2	0	1E+10	1	1
3	1	1E+09	0.999999999	0.999999999
4	2	100000000	0.99999999	0.99999999
5	3	10000000	0.9999999	0.9999999
6	4	1000000	0.999999	0.999999
7	5	100000	0.99999	0.99999
8	6	10000	0.99990001	0.99990001
9	7	1000	0.999000999	0.999000999
10	8	100	0.99009901	0.99009901
11	9	10	0.909090909	0.909090909
12	10	1	0.5	0.5
13	11	0.1	0.090909091	0.090909091
14	12	0.01	0.00990099	0.00990099
15	13	0.001	0.000999001	0.000999001
16	14	0.0001	9.999E-05	9.999E-05
17				
19				
19	ratio = $[RNH_3^+]/[RNH_2]$			
20	fraction = $[RNH_3^+]/([RNH_3^+] + [RNH_2])$			

Figure 3. Spreadsheet used for calculating various parameters (e.g. ratio of ammonium form to amino form) as a function of pH.

To visualize the above parameters as functions of pH we construct three graphs, as shown in Figure 4. Once again we see that a graph of the ratio of one form to the other form (this time the ammonium form to the amino form) as a function of pH is not very instructive (Figure 4A), but the graph showing the fraction of the compound in the ammonium form as a function of pH conforms to our expectations that this fraction is essentially unity until the pH approaches the pK_a of our compound (Figure 4B). In other words, the protonated form predominates at pH values well below the pK_a of our compound. The graph showing the charge of the compound as a function of

pH is shown in Figure 4C, and it is apparent that this graph is identical to that shown in Figure 4B. The reason why these two graphs appear the same is because the compound carries a positive charge when it is protonated, but the compound is neutral when it is unprotonated. Consequently, the fraction of the compound in the ammonium form is multiplied by positive one, which does not change the value of the fraction; hence the identical appearance of the graphs shown in Figures 4B and 4C.

Before we move on to consider amino acids, it is worthwhile to note that at pH 8.0 the ratio of the ammonium form to the amino form is 100 (cell B10, Figure 3), and the fraction of the compound in the ammonium form at this pH is 100/101 ≈ 0.99009901 (cell C10, Figure 3). When the pH is 12, the ratio of the ammonium form to the amino form is 0.01 (cell C14, Figure 3), and the fraction of the compound in the ammonium form at this pH is 1/101 ≈ 0.00990099 (cell D14, Figure 3). Again, these observations tell us that at pH values either two

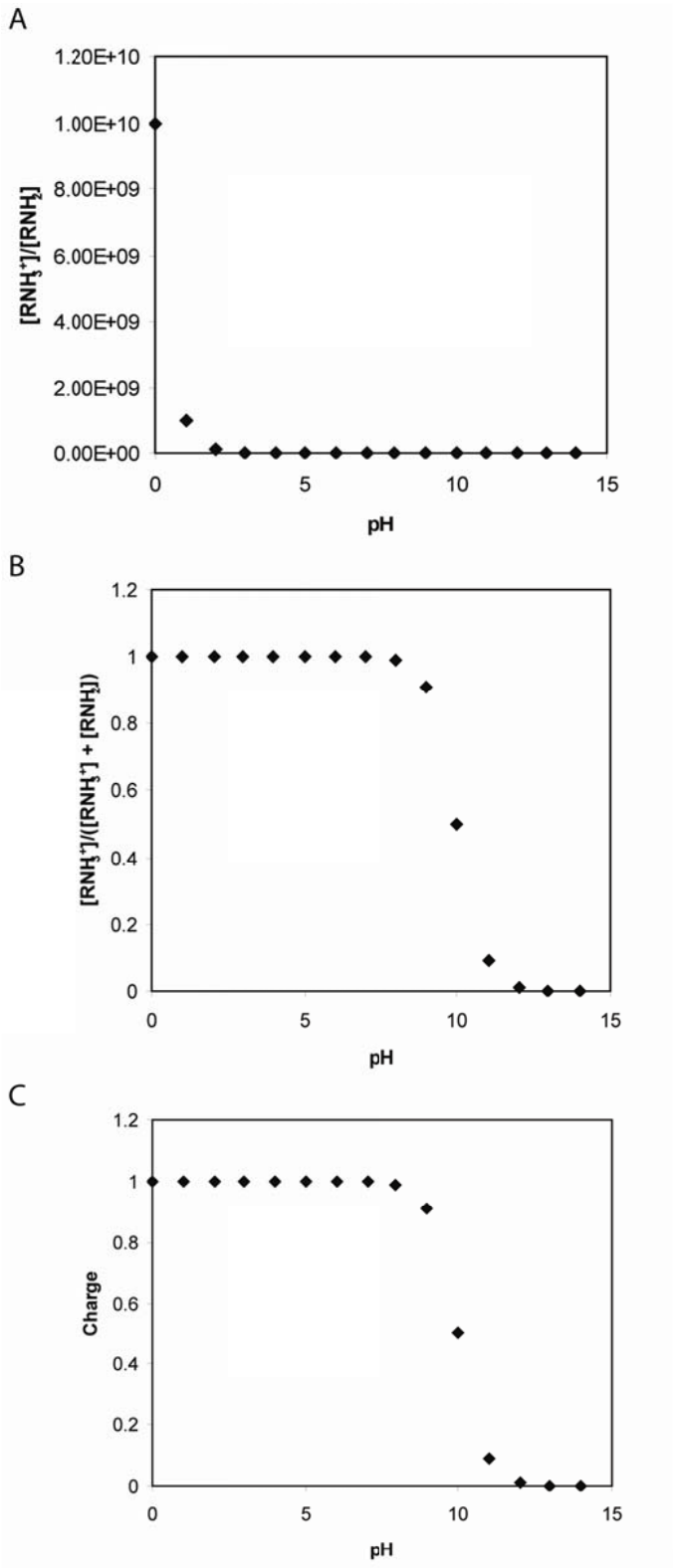
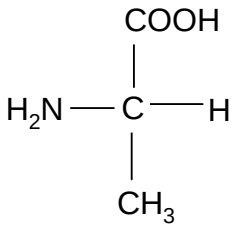


Figure 4. (A) ratio of ammonium to amino form as a function of pH; (B) fraction of ammonium form as a function of pH; (C) charge as a function of pH.

units below or two units above the pK_a of the compound, we can make the approximation that almost all of the compound is protonated or deprotonated, respectively. Mathematically, we can make the approximation that the fraction of the compound in the ammonium form is ~1 when the pH is two units lower than the pK_a, and the fraction of the compound in the amino form is ~1 when the pH is two units higher than the pK_a. And, as mentioned above, the approximations progressively improve as the separation between the pK_a and the pH that we are considering becomes greater than two units (column C, Figure 3).

Amino Acids

We are now ready to consider amino acids, and the amino acid that we will start with is one of the simplest: L-alanine. When drawn in a modified Fisher projection, L-alanine can be represented as follows (albeit in a very unlikely protonation state):



We see that L-alanine has both a carboxyl and an amino group attached to a central, chiral carbon. This carbon is called the α-carbon; thus, the carboxyl group is called the α-carboxyl group and the amino group is called the α-amino group. We also see that a methyl group is attached to the α-carbon. This group is referred to as the R-group or side chain as indicated in Table 1. The L-designation is part of the D,L-system of nomenclature in which glyceraldehyde is used as the reference compound to decide on the D,L-configuration of various organic molecules. All of the 20 standard amino acids found in proteins (with the exception of glycine, which does not have a chiral carbon) have the L-configuration. For our purposes we will not be concerned

with this distinction, and from this point on I simply will refer to various amino acids without the L-designation.

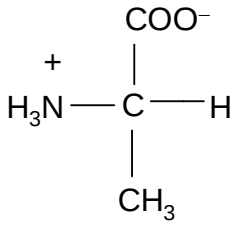
Table 1. One- and three-letter codes, mnemonics for memorizing the one-letter codes, side-chains, and pK_a values of the 20 standard amino acids.

Obvious ones	Mnemonic	Side-chain	pK_1	pK_2	pK_R
A = Ala = Alanine		$-CH_3$	2.35	9.87	
C = Cys = Cysteine		$-CH_2SH$	1.92	10.70	8.37
G = Gly = Glycine		-H	2.35	9.78	
H = His = Histidine		$-CH_2$-imidazole	1.80	9.33	6.04
I = Ile = Isoleucine		-*sec*-butyl	2.32	9.76	
L = Leu = Leucine		-isobutyl	2.33	9.74	
M = Met = Methionine		$-CH_2CH_2SCH_3$	2.13	9.28	
P = Pro = Proline		$-CH_2CH_2CH_2$-αNH	1.95	10.64	
S = Ser = Serine		$-CH_2OH$	2.19	9.21	
T = Thr = Threonine		$-CH(OH)CH_3$	2.09	9.10	
V = Val = Valine		-isopropyl	2.29	9.74	
Phonetic ones					
F = Phe = Phenylalanine	(**ffffff**enylalanine)*	-benzyl	2.20	9.31	
N = Asn = Asparagine	(asparagi**nnnnnn**e)*	$-CH_2CONH_2$	2.14	8.72	
R = Arg = Arginine	(a**rrrrr**ginine)*	$-(CH_2)_3$-guanidinium	1.82	8.99	12.48
Y = Tyr = Tyrosine	(t**yyyyy**rosine)*	$-CH_2$-phenol	2.20	9.21	10.46
Non-obvious ones					
D = Asp = Aspartate	**D**asp	$-CH_2COO^-$	1.99	9.90	3.90
E = Glu = Glutamate	Glu**E**	$-CH_2CH_2COO^-$	2.10	9.47	4.07
K = Lys = Lysine	hj**Kl**mnop	$-(CH_2)_4NH_3^+$	2.16	9.06	10.54
Q = Gln = Glutamine	**Q**tamine	$-CH_2CH_2CONH_2$	2.17	9.13	
W = Trp = Tryptophan	bulkiest letter/residue*	$-CH_2$-indole	2.46	9.41	

*suggested by Professor Ann Palmenberg of the University of Wisconsin–Madison

Once again, we consider how best to represent our compound at various pH values. To help us decide, we consult a suitable reference (*4*) and see (as consolidated in Table 1) that the pK_a values of the α-carboxyl and α-amino (more properly called α-ammonium) groups of alanine are 2.35 and 9.87, respectively. By convention, the pK_a of the α-carboxyl group is referred to as pK_1, and the pK_a of the α-ammonium group is referred to as pK_2. These pK_a values are somewhat lower than we might expect. For example, propionic acid (or propanoic acid) also has three carbons like alanine; unlike alanine, however, propionic acid does not have an α-amino group, and the pK_a of propionic acid is 4.87 (*4*). Isopropyl amine has an amino functional group attached to the second of three carbons, similar to alanine, but the isopropylammonium ion has a

pK_a of 10.63 (*5*). The depressed pK_a values of the groups in alanine are explained on the basis of electron withdrawing effects of both the α-carboxyl and the α-amino groups (*6, 7*). Nonetheless, given these pK_a values and given what we learned above, we might expect that the α-carboxyl groups would essentially all be in the carboxylate form at pH levels higher than two units above pK_1. Similarly, we might expect that the α-amino groups would essentially all be in the ammonium form at pH levels lower than two units below pK_2. Thus, at several intermediate pH levels the predominant form of the molecule is better represented as follows:



This form is called the zwitterion form (where zwitter is the German word for "hybrid" or "hermaphrodite"). We can go further and use a spreadsheet to get a better understanding of the acid-base behavior of this compound as a function of pH. When we construct our spreadsheet we use smaller increments in the increase of pH values for reasons that will become apparent. A portion of the spreadsheet that one might construct is shown in Figure 5.

	A	B	C	D
1	pH	(-1)*fraction COO^-	(1)*fraction NH_3^+	Net Charge
2	0	-0.004446972	1	0.995553
3	0.01	-0.004550084	1	0.9954499
.	.	.	.	.
.	.	.	.	.
237	2.35	-0.5	0.99999997	0.5
.	.	.	.	.
.	.	.	.	.
613	6.11	-0.99982625	0.99982625	0
.	.	.	.	.
.	.	.	.	.
989	9.87	-0.99999997	0.5	-0.5
.	.	.	.	.
.	.	.	.	.
1402	14	-1	7.41255E-05	-0.9999259

Figure 5. Portion of a spreadsheet used for calculating various parameters (e.g., net charge) as a function of pH for the amino acid alanine.

The function statement in cell B2 of Figure 5 is "=(–1)*10^(A2–2.35)/(10^(A2–2.35) + 1); the function statement in cell C2 of Figure 5 is "=(1)*10^(9.87–A2)/(10^(9.87–A2) + 1); and the function statement in cell D2 of Figure 3 is "=B2+C2". We also could have entered "=(–1)*10^(A2–2.35)/(10^(A2–2.35) + 1) + (1)*10^(9.87–A2)/(10^(9.87–A2) + 1)", although it would have been more work to do so. Nonetheless, what this last function statement is saying is that the net charge that is observed is equal to minus one times the fraction of the carboxylate form plus positive one times the fraction of the ammonium form, which should agree with the picture we create in our minds. For example, at pH = 0 almost all of the α-carboxyl groups should be protonated and therefore neutral, but to an even greater extent essentially all of the α-amino groups also should be protonated and positively charged. Thus the net charge is the sum of the fractions of the two groups (adjusted to account for the presence of a negative or positive charge). At pH = 0 this sum is very close to one (cell D2, Figure 5), which we anticipated. Since the incremental increase in pH is by 0.01 units, it takes 1401 rows to reach a pH of 14. We might expect that at this pH essentially all of the α-carboxyl groups are unprotonated and thus negatively charged and almost all of the α-amino groups also are unprotonated and thus neutral. Thus we expect that the net charge should be very close to minus one, and this is indeed what we find (cell D1402, Figure 5).

When the pH equals pK_1 (or the pK_a of the α-carboxyl group; cell A237, Figure 5), we see that the net charge of alanine is 0.5 (cell D237, Figure 5). We can rationalize this result two ways. First, we see that the fraction of the compound having carboxylate groups is 0.5; thus, this fraction multiplied by minus one equals –0.5 (cell B237, Figure 5). The fraction of the compound having ammonium groups is almost one (cell C237, Figure 5); thus this fraction multiplied by positive one is almost one. The net charge is the sum of cells B237 and C237,

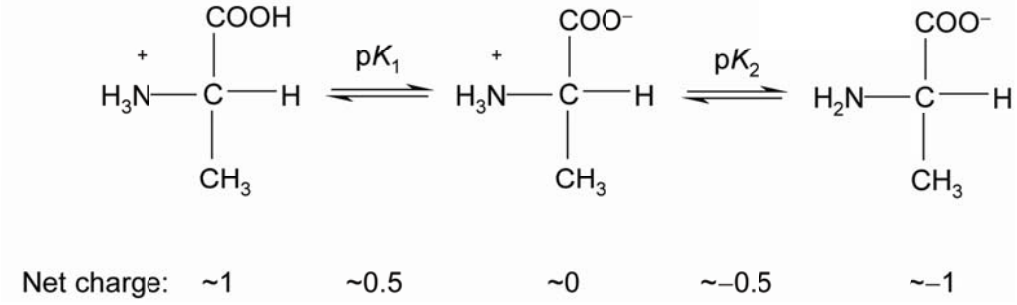
which is, within rounding, 0.5 (cell D237, Figure 5). Another way to rationalize this result is to realize that at pH = 2.35, half the molecules have a net charge of one (those with protonated α-carboxyl and protonated α-amino groups) and half the molecules have a net charge of zero (those with unprotonated α-carboxyl and protonated α-amino groups). Thus $(0.5 \times 1) + (0.5 \times 0) = 0.5$.

When the pH equals pK_2 (or the pK_a of the α-amino group; cell A989, Figure 5), we see that the net charge of alanine is –0.5 (cell D989, Figure 5). We can rationalize this result two ways. First, we see that the fraction of the compound having carboxylate groups is almost one; thus, this fraction multiplied by minus one almost equals minus one (cell B989, Figure 5). The fraction of the compound having ammonium groups is 0.5 (cell C989, Figure 5); thus this fraction multiplied by positive one is still 0.5. The net charge is the sum of cells B989 and C989, which is, within rounding, –0.5 (cell D989, Figure 5). Another way to rationalize this result is to realize that at pH = 9.87, half the molecules have a net charge of zero (those with unprotonated α-carboxyl and protonated α-amino groups) and half the molecules have a net charge of minus one (those with unprotonated α-carboxyl and unprotonated α-amino groups). Thus $(0.5 \times 0) + (0.5 \times -1) = -0.5$.

We sum up our observations to this point as follows: at low pH values (e.g., pH = 0) the net charge is close to positive one; as the pH increases the net charge decreases such that when the pH equals pK_1 the net charge decreases to 0.5; when the pH equals pK_2 the net charge is –0.5; and when the pH equals 14 the net charge is minus one. At some point, the net charge went from positive to negative, which means we passed through a point where the net charge was zero. This point corresponds to a pH of 6.11 (row 613, Figure 5). This point is midway between the two pK_a values; in fact, it is the average of two pK_a values, and it should agree with the picture we might form in our minds. When almost all of the α-carboxyl groups just finish

deprotonating, but before any of the α-amino groups (which are almost all in the ammonium form) start deprotonating, essentially all of the molecules are in the zwitterion form and the net charge will equal zero. The pH at which the net charge of a compound is zero is defined as the isoelectric point or pI (iso is the Greek root for "equal"). The concept of pI will play a major role in our subsequent discussions. Many of the above observations are presented in a succinct form in Scheme 1.

Scheme 1.



We can visualize the above changes in charge as a function of pH (or vice versa) by constructing the graphs shown in Figure 6. In Figure 6A we show three traces, as opposed to the discreet data points shown in Figures 2 and 4. The reason why we use lines to form the traces is that we have so many data points (1401) that they stack up on one another and form a thick "line". We remove the individual data points and replace them with a line by moving our cursor to any data point and depressing the right-button of our mouse. We then select "Format Data Series" where we see a new window in which various options are displayed. In this window we specify the size of the data-point symbols ("none"), which Excel calls "markers", and we select the option to connect the data points with a line (of whatever pattern we choose).

The lower trace in Figure 6A is similar to the graph shown in Figure 2C and thus represents a compound that has one ionizable functional group – a carboxyl group, albeit with a lower pK_a than is typically found in such compounds. In fact, we obtain the lower trace of Figure

6A by plotting column B of Figure 5 (headed as "(–1)*fraction COO^-") as a function of pH. The upper trace seen in Figure 6A is similar to the graph shown in Figure 4C and thus represents a compound that has one ionizable group – an ammonium group, albeit with a pK_a somewhat lower than is typically found in such compounds. We obtain this trace by plotting column C of Figure 5 (headed as "(1)*fraction NH_3^+") as a function of pH. The middle trace in Figure 6A represents a compound that has both of the above functional groups with their respective pK_as. Thus, this middle trace represents the amino acid alanine, and we obtain this trace by plotting column D of Figure 5 (headed as "Net Charge") as a function of pH. We see that the net charge is close to one at low pH, it is close to minus one at high pH, and it is close to zero at several intermediate pH values, where we anticipated the zwitterion form would

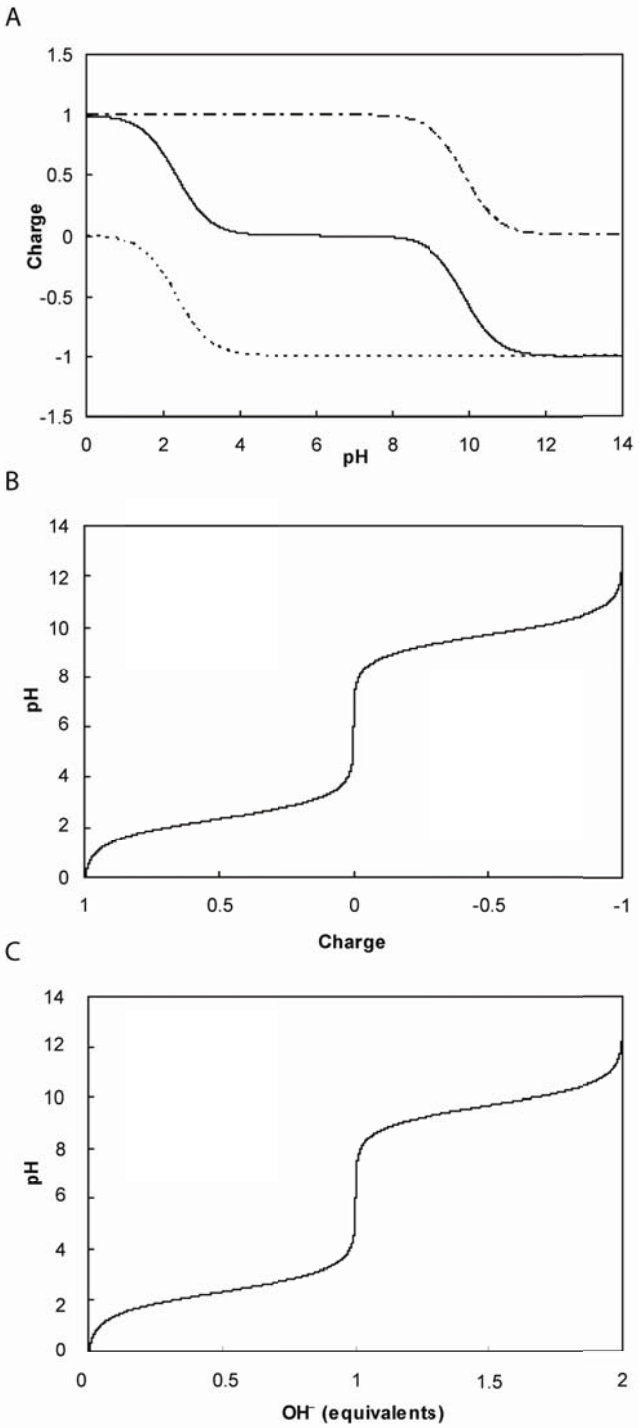


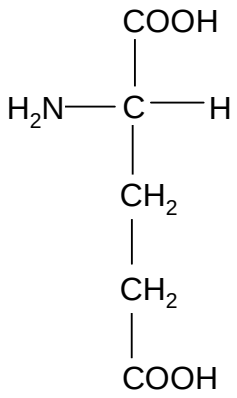
Figure 6. (A) Charge as a function of pH for a compound that has a carboxyl group with a pK_a of 2.35 (lower trace), an ammonium group with a pK_a of 9.87 (upper trace), and a compound (alanine) that has both a carboxyl and ammonium functional group with the above, respective, pK_as (middle trace); (B) middle trace of (A) with the axes inverted; (C) pH as a function of OH^- equivalents for the amino acid alanine.

predominate. If instead we plot pH as a function of Net Charge (in other words, if we invert the axes of the middle trace of Figure 6A) we obtain the graph shown in Figure 6B. The pattern of the graph shown in Figure 6B is identical to that seen in Figure 6C, which represents an idealized titration curve (pH as a function of OH^- equivalents) of the amino acid alanine, and such titration curves are often depicted in textbooks of biochemistry. Figure 6A and 6B both provide a nice visual picture of what must be the approximate pI, even if one did not know the values of pK_1 and pK_2. Figure 6C provides a nice picture of the buffering regions about pK_1 and pK_2, which again could be estimated from the graph. Figure 6C also makes clear that when we have added just enough base (in fact, one equivalent of base) to titrate completely the α-carboxyl group but not the α-amino group, then the corresponding pH will be ~6 (which we know to be 6.11 from cell A613 of Figure 5).

Before we consider a more complicated amino acid, it is worth noting that since alanine does not contain a side chain that has an ionizable group, its acid-base chemistry is similar to and representative of 12 of the other 20 standard amino acids because none of these has an ionizable side chain either. Thus, what we learn about alanine holds for these other amino acids, but it is important to note that the pK_as of the α-carboxyl and α-ammonium groups of these other amino acids will be somewhat different than those of alanine, so any spreadsheets that we may construct will have to account for these different values. While I do not feel that beginning students of biochemistry (or practicing biochemists for that matter) need to remember the exact values of the pK_as of these amino acids, I do feel that it is important to have a general idea of these values. For example, all the α-carboxyl groups of all of the 20 standard amino acids have pK_a values in the neighborhood of ~2 (average of all twenty is 2.1), and all the α-ammonium groups of all of the 20 standard amino acids have pK_as in the neighborhood of ~9.5 (average of

all twenty is 9.5) (Table 1). There are some notable exceptions, but these values are good “ball park” figures.

Our next example molecule is the amino acid glutamic acid (or glutamate), which is represented in the modified Fisher projection as follows (albeit, again, in a very unlikely protonation state):



Glutamate is more complicated than alanine because it has a side chain that itself contains an ionizable functional group – namely a carboxyl group. The pK_as of glutamate are as follows: pK_1 = 2.10; pK_2 = 9.47; and pK_R (the pK_a of the side-chain carboxyl group, which also is called the γ-carboxyl) = 4.07 (Table 1; (*4*)). Given these values, we expect that at pH = 0 almost all of the α-carboxyl groups will be protonated and neutral, almost all of the α-amino groups will be protonated and positively charged, and almost all of the γ-carboxyl group also will be protonated and neutral for an overall net charge of positive one. At pH = 14, we predict that almost all of the α-carboxyl groups will be unprotonated and negatively charged, almost all of the α-amino groups will be unprotonated and neutral, and almost all of the γ-carboxyl group also will be unprotonated and negatively charged for an overall net charge of minus two. But what about the protonation states of glutamate at intermediate pH levels? What is the net charge when the pH equals pK_1 or pK_2 or pK_R? What is the pI of glutamate?

We can obtain approximate answers to the above questions relatively quickly by reasoning our way through a "mental" titration given what we learned above. When the pH = pK_1, half of the α-carboxyl groups are unprotonated, but practically none of the α-ammonium groups are unprotonated and very few (~1%) of the γ-carboxyl groups are unprotonated. So at this pH, approximately half of the molecules have a net charge of one (i.e., those with a protonated and neutral α-carboxyl, protonated and positive α-ammonium, and protonated and neutral γ-carboxyl), and half of the molecules have a net charge of zero (i.e., those with an unprotonated and negative α-carboxylate, protonated and positive α-ammonium, and protonated and neutral γ-carboxyl). Thus, we predict that the average net charge at this pH is about 0.5.

When the pH = pK_R, most (~99%) of the α-carboxyl groups are unprotonated, practically none of the α-ammonium groups are unprotonated, and half of the γ-carboxyl groups are unprotonated. We find then that approximately half of the molecules have a net charge of zero (i.e., those with an unprotonated and negative α-carboxylate, protonated and positive α-ammonium, and protonated and neutral γ-carboxyl), and half of the molecules have a net charge of −1 (i.e., those with an unprotonated and negative α-carboxylate, protonated and positive α-ammonium, and unprotonated and negative γ-carboxylate). Thus, we predict that the average net charge at this pH is about −0.5.

When the pH = pK_2, almost all of the α-carboxyl groups are unprotonated, half of the α-ammonium groups are unprotonated, and almost all of the γ-carboxyl groups are unprotonated. So we find that approximately half of the molecules have a net charge of −1 (i.e., those with an unprotonated and negative α-carboxylate, protonated and positive α-ammonium, and unprotonated and negative γ-carboxylate), and half of the molecules have a net charge of −2 (i.e.,

those with an unprotonated and negative α-carboxylate, unprotonated and neutral α-amino, and unprotonated and negative γ-carboxylate). Thus, we predict that the average net charge at this pH is about −1.5. We estimate the pI by noting that the average net charge at pK_1 is about 0.5, and the average net charge at pK_R is about −0.5. Thus, the sign change occurs midway between these two pK_as, and that is where the pI is located. The average of pK_1 and pK_R is 3.085.

We can check our reasoning by constructing a spreadsheet. An example of a spreadsheet that one might construct is shown in Figure 7. As was the case with alanine, we increase the pH values by 0.01 units per cell. The function statement in cell B2 of Figure 7 is "=(–1)*10^(A2–2.1)/(10^(A2–2.1) + 1)" to correspond to charge of the α-carboxyl group as a function of pH; the function statement in cell C2 of Figure 7 is "=(1)*(10^(9.47–A2)/(10^(9.47–A2) + 1)" to correspond to charge of the α-amino group as a function of pH; the function statement in cell D2 of Figure 7 is "=(–1)*(10^(A2–4.07)/(10^(A2–4.07) + 1)" to correspond to charge of the γ-carboxyl group as a

	A	B	C	D	E
1	pH	(-1)*fraction αCOO^-	(-1)*fraction αNH_3^+	(-1)*fraction γCOO^-	Net Charge
2	0	-0.007880684	1	-8.51066E-05	0.9920342
3	0.01	-0.008062769	1	-8.70888E-05	0.9918501
.	.	.	.	.	.
.	.	.	.	.	.
212	2.1	-0.5	0.999999957	-0.010601595	0.4893984
.	.	.	.	.	.
.	.	.	.	.	.
310	3.08	-0.905212604	0.999999593	-0.092830064	0.0019569
311	3.09	-0.907169936	0.999999583	-0.094787396	-0.001958
.	.	.	.	.	.
.	.	.	.	.	.
409	4.07	-0.989398405	0.999996019	-0.5	-0.489402
.	.	.	.	.	.
.	.	.	.	.	.
949	9.47	-0.999999957	0.5	-0.999996019	-1.499996
.	.	.	.	.	.
.	.	.	.	.	.
1402	14	-1	2.95112E-05	-1	-1.99997

Figure 7. Portion of a spreadsheet used for calculating various parameters (e.g. net charge) as a function of pH for the amino acid glutamate.

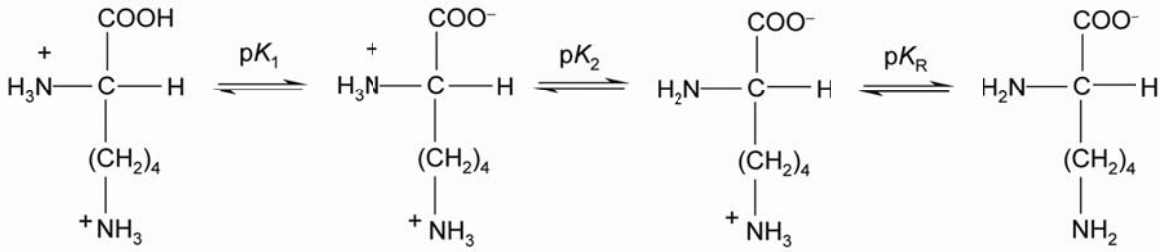
function of pH; and the function statement in cell E2 of Figure 7 is "=B2+C2+D2", which, as seen before with alanine, provides the net charge of the molecule as a function of pH. We also can construct our spreadsheet such that we only have two columns, one for pH and the other for the net charge. If we choose to do this our function statement in the first cell of the "net charge" column needs to be the sum of the function statements (typed out in full) that we used in cells B2, C2, and D2 of Figure 7.

We see in cell E2 of Figure 7 that at pH = 0 the net charge of glutamate is about 0.99, which, as we anticipated, is close to one. We see in cell E1402 of Figure 7 that at pH = 14 the net charge of glutamate ≈–2.0, which we also anticipated. We see that at a pH = pK_1, the net charge is about 0.489 (cell E212, Figure 7), which is close to our above estimate of 0.5; at pH = pK_R, the net charge is about –0.489 (cell E409, Figure 7), which is close to our above estimate of –0.5; and at pH = pK_2, the net charge is about –1.499996 (cell E949, Figure 7), which again is very close to our above estimate of –1.5. We also see in Figure 7 that the sign changes from positive in cell E310 to negative in cell E311. The pH values in the corresponding cells (A310 and A311) are 3.08 and 3.09, respectively. The average of these two cells (3.085) is the average of pK_1 and pK_R, which corresponds to the pI as we anticipated. Thus, we can still find the pI by averaging two pK_as, but it is important that we identify the two relevant pK_as.

If instead of glutamate we consider the amino acid aspartic acid (or aspartate) we find that its pI also is the average of its pK_1 and pK_R. The reason why this is so is that the side chain of aspartate also contains an acidic carboxyl group (called a β-carboxyl), and this group has a pK_a of 3.90 (Table 1; (*4*)). There are five other amino acids that have side chains that contain ionizable functional groups in the pH range that we typically consider (i.e., 0 – 14). These amino acids include: cysteine, which has a thiol functional group; tyrosine, which has a phenol

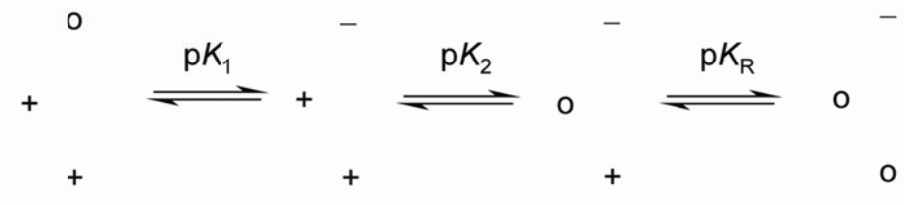
functional group; histidine, which has an imidazole functional group; lysine, which has an ε-ammonium group; and arginine, which has δ-guanidinium group (Table 1; (*4*)). For the amino acids that have basic side chains (e.g., histidine, lysine, arginine) we find the pI by taking the average of pK_2 and pK_R. To see why the pI is the average of the pK_2 and pK_R of these amino acids, it helps to perform another mental titration as shown in Scheme 2 for the amino acid lysine.

Scheme 2.



An inspection of Scheme 2 shows that the two pK_as that bracket the neutral species are pK_2 and pK_R; thus, the average or midpoint between these two pK_as represents the pI. Had we shown the titration of histidine, pK_R would have come before pK_2 because the side-chain of histidine titrates before the α-ammonium group because its pK_a is lower than that of the α-ammonium group. To save time, one can perform a "symbolic" titration in which only plus (+), neutral (0), and minus (–) symbols are used to indicate the different protonation/charge states rather than drawing out the full structures each time; such a titration (of an amino acid that has a basic side-chain with pK_R > pK_2) is shown in Scheme 3.

Scheme 3.

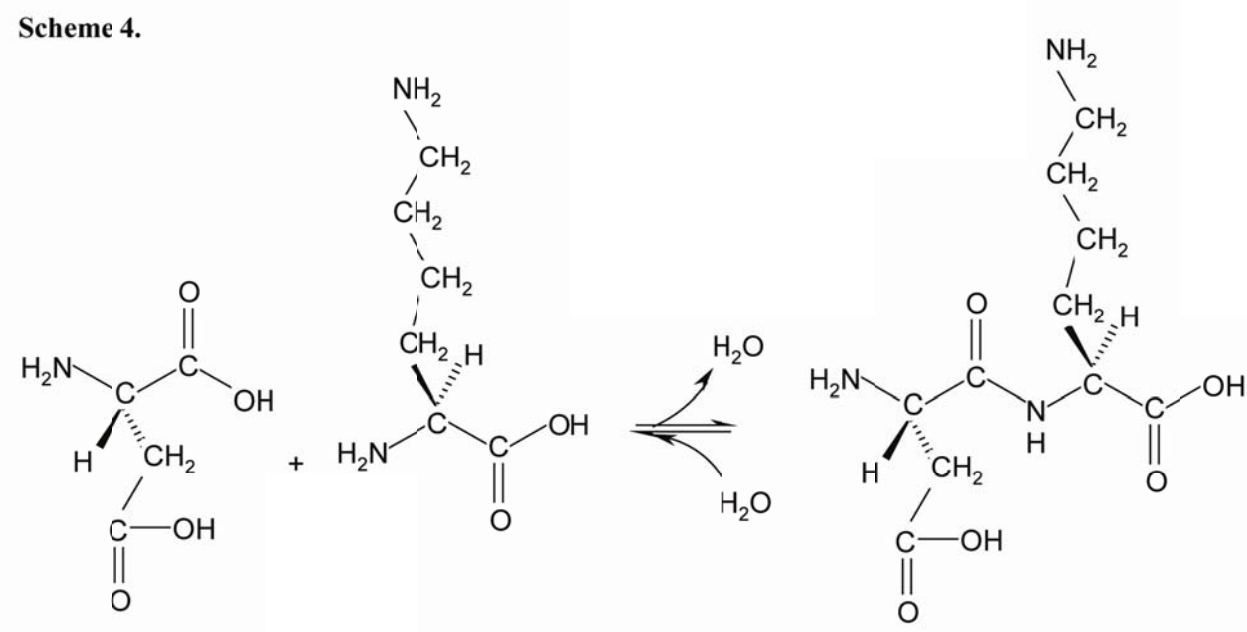


As before, the above considerations about the pI of amino acids that have basic side chains can be demonstrated and thus reinforced by use of a spreadsheet. Although we will not construct such a spreadsheet at this time, all of the basic tools have been provided above and it is left as an exercise for interested students. Now that we have discussed the fundamentals of using spreadsheets to help clarify acid-base concepts of simple, weak acids and of amino acids, we are ready to consider peptides and proteins.

Peptides and Proteins

To begin our discussion of peptides and proteins, we first consider the formation of a dipeptide as shown in Scheme 4.

Scheme 4.



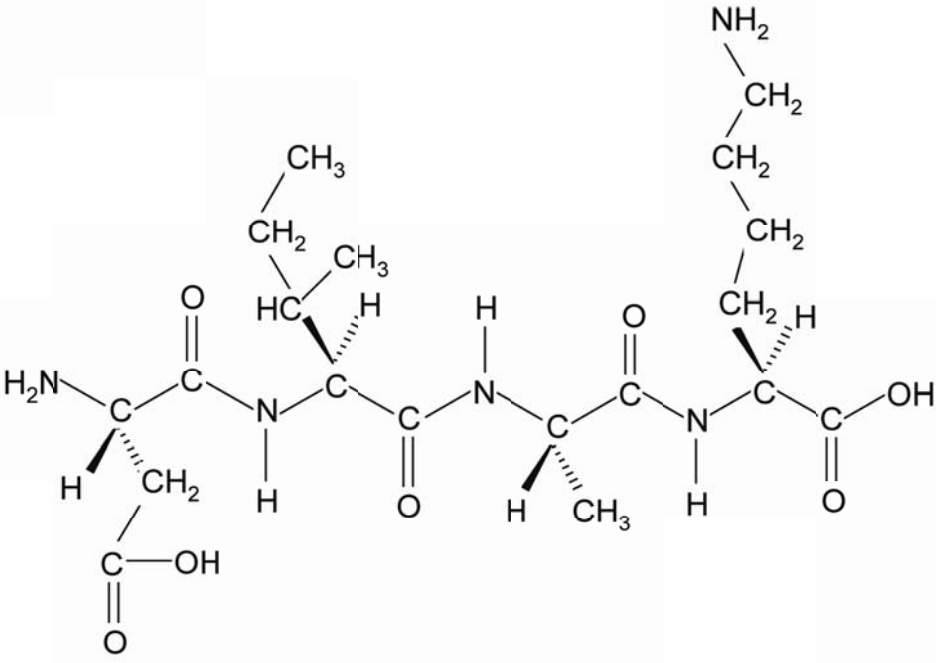
Specifically, Scheme 4 shows how the condensation of the amino acids aspartate and lysine forms a peptide (amide) bond that connects the two amino acids, although peptide bond formation in the cell does not proceed in so simple a manner. Since the elements of water are lost when a peptide bond is formed, we refer to amino acids that are part of a peptide as amino acid

residues. We can symbolize our dipeptide as Asp-Lys or as DK by using the three- and one-letter codes for the amino acids, as listed in Table 1. (Note that in Table 1 hints are provided to help one memorize these codes, which is something that all beginning students must do as part of learning the "language" of biochemistry.) It is important to note that we should not represent our dipeptide as Lys-Asp or KD because such a representation indicates a different dipeptide, as discussed below.

Note that in Scheme 4 the α-carboxyl group of aspartate becomes part of the peptide bond that links the two amino acids. Since this α-carboxyl group is "tied-up" in the peptide bond it is no longer free to ionize. Note also that the α-amino group of aspartate is not part of the peptide bond and thus is still free to ionize, depending upon the pH of the surrounding medium. The situation is reversed for the lysine residue. The α-amino group of the lysine becomes part of the peptide bond and thus is no longer free to ionize, whereas the α-carboxyl group of lysine is not part of the peptide bond and thus is free to ionize. By a convention that also reflects the way in which peptides are synthesized in cells, the end of a peptide with the free α-amino group is listed first and is referred to as the amino- or N-terminal amino acid residue; the end of a peptide with the free α-carboxyl group is listed last and is referred to as the carboxyl- or C-terminal amino acid residue. Since in a dipeptide the α-carboxyl group of the N-terminal residue and the α-amino group of the C-terminal residue are part of the peptide bond and thus not free to ionize, if we want to determine the pI of the dipeptide, we only have to consider the α-amino group of the N-terminal residue, the α-carboxyl group of the C-terminal residue, and the side-chains of the amino acids that make up the dipeptide (if these side-chains contain ionizable groups). This last statement applies to larger peptides such as tri-, tetra- and even polypeptides since the α-carboxyl and α-amino groups of all internal amino acid residues are not free to ionize, which can

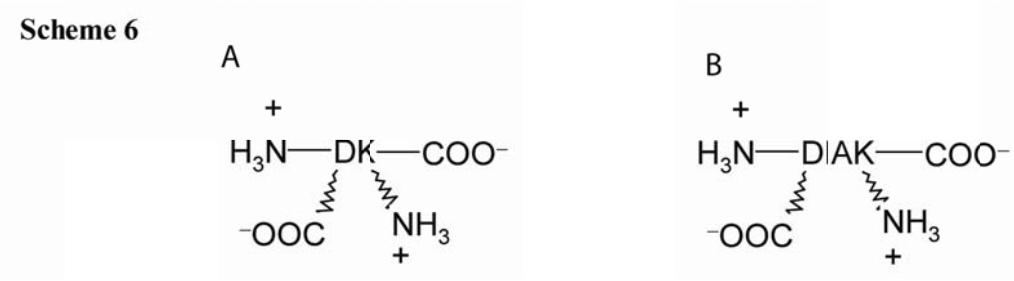
be seen in the case of the tetrapeptide shown in Scheme 5. This tetrapeptide is correctly represented as either Asp-Ile-Ala-Lys or DIAK using the three- or one-letter codes, respectively. Note that in Scheme 5 peptide bonds "tie-up" the α-carboxyl of aspartate, both the α-amino and the α-carboxyl of isoleucine, both the α-amino and the α-carboxyl of alanine, and the α-amino of lysine. Thus, to estimate the pI we have only to consider the free α-amino of aspartate (the N-terminal residue), the free α-carboxyl of lysine (the C-terminal residue), and the side-chains of aspartate and lysine.

Scheme 5.



One way to estimate the pIs of the above di- and tetrapeptides is to imagine the protonation states of the relevant groups at pH 7. We can facilitate this process by writing the peptides in a "shorthand" manner that emphasizes the relevant groups in their assumed protonation states as shown in Scheme 6 A and B.

Scheme 6



To decide on the protonation states it is important to realize that the pK_as of the relevant groups in a peptide might be different than the pK_as of the same groups in free amino acids. For example, the pK_a of the ε-ammonium group of Lys115 in the enzyme acetoacetate decarboxylase is estimated to be ~6 (*8-10*), which is about a 4.5 unit shift from the pK_a of the ε-ammonium group of free lysine. Despite the perturbed pK_as of the side-chains of some amino acid residues, we expect that on average most will have pK_as that are not too different from the respective pK_as of the free amino acids. The N- and C-terminal groups of peptides and proteins, however, are estimated to be considerably different from those of α-ammonium and α-carboxyl groups of free amino acids. Such differences are apparent in Table 2, which shows three sets of approximate pK_a values of ionizable groups in proteins.

Table 2. Different assumed pK_a values of groups in peptides and proteins.

Group	Assumed pK_a Values in Peptides and Proteins		
	Textbook	ExPASy*	EMBOSS*
N-terminal α-amino group	8.0	7.5†	8.6
C-terminal α-carboxyl group	3.1	3.55	3.6
β-carboxyl of Asp	4.1	4.05	3.9
γ-carboxyl of Glu	4.1	4.45	4.1
Imidazole of His	6.0	5.98	6.5
Thiol of Cys	8.3	9.0	8.5
Phenol of Tyr	10.9	10.0	10.1
ε-ammonium of Lys	10.8	10.0	10.8
δ-guanidinium of Arg	12.5	12.0	12.5

*These terms are discussed in the text.
†The actual values used depend on the identity of the N-terminal residue.

The first set of values in Table 2 is from a standard textbook of biochemistry (*11*). The second set of values in Table 2 is from (*12*, *13*); these values are used by the computer programs "pI/Mw" and "ProtParam" (*14*) on the Expert Protein

Analysis System (ExPASy) website (*15*). The third set of values in Table 2 is from the European Molecular Biology Open Software Suite (EMBOSS) (*16, 17*); these values are used by the programs "iep" and "pepstats" that are part of this suite of programs. If we use the first set of values in Table 2, we arrive at the protonation states shown in Scheme 6 for an assumed pH of 7. An inspection of Scheme 6 shows that both DK and DIAK appear to have net charges of zero at pH 7. The reason why we cannot conclude that the net charge *is* zero at pH 7 is that the value of the pK_a of the N-terminal α-amino groups is assumed to be 8.0, so at pH 7 the ratio of ammonium form to amino form of this group is 10:1 for both peptides. If we imagine that the pH is lowered below 7 then the ammonium form of the N-terminal α-amino groups of both peptides is favored. If the pH is lowered too much, the group with the next closest pK_a, which for both peptides would be the β-carboxylates of the aspartate residues, will start to become protonated. In fact, we can estimate that the pI is midway between the pK_as of the N-terminal α-amino groups and the β-carboxyl groups of the aspartate residues. The pI that we estimate is the same for both the di- and tetrapeptides since the side-chains of the two additional amino acid residues in the tetrapeptide do not ionize.

The above line of attack is very approximate, but it works reasonably well for small peptides. If we imagine that the protonation state of the peptide under consideration is such that the net charge is positive at pH 7, we know that we need to "move" to a higher pH to deprotonate a group and either expose a negative charge or remove a positive charge. If on the other hand we imagine the net charge of the peptide is negative at pH 7 then we know that we need to "move" to a lower pH to protonate a group to add a positive charge or neutralize a negative charge. Using this approach, we generally can identify the two residues whose pK_as most closely bracket the neutral species, and the pI usually is close to the average of the pK_as of these two residues. For

larger polypeptides, the above approach is somewhat cumbersome and time-consuming, and once again, a spreadsheet approach helps tremendously.

To use a spreadsheet to help us estimate the pI of a polypeptide or protein, we proceed in a manner similar to that described above for amino acids, but we add a few changes to make our spreadsheet more robust and flexible so that we can easily change the assumed pK_a values of amino acid residues and so we can use the same spreadsheet for different proteins, including proteins that are made up of more than one polypeptide or subunit – i.e., oligomeric proteins. An example of such a spreadsheet, applied to the tetrapeptide DIAK discussed above, is shown in Figure 8. Column B of Figure 8 is used to specify the number of N-terminal, C-terminal, or other ionizable residues indicated in column A. Column C of Figure 8 indicates the assumed pK_a values of the N-terminal, C-terminal, or other ionizable residues specified in columns A and B. The cells in columns B and C are "absolute references" for the calculations that will be performed in column E under the heading "Net Charge", as will be described below. The function statement in cell E2 of Figure 8 is: "=((1)*(B2)*(10^(C2–D2)/(10^(C2–D2)+1)) + (–1)*(B3)*(10^(D2–C3)/(10^(D2–C3)+1)) + (–1)*(B4)*(10^(D2–C4)/(10^(D2–C4)+1)) +

	A	B	C	D	E
1	Amino Acid	No. of Residues	pK_a value	pH	Net Charge
2	N-terminal	1	8	0	1.999126866
3	C-terminal	1	3.1	0.01	1.999106543
4	Asp	1	4.1	0.02	1.999085748
5	Glu	0	4.1	0.03	1.999064468
6	His	0	6	0.04	1.999042695
7	Cys	0	8.3	0.05	1.999020414
8	Tyr	0	10.9	0.06	1.998997616
9	Lys	1	10.8	0.07	1.998974287
10	Arg	0	12.5	0.08	1.998950416
.				.	.
.				.	.
609				6.07	4.08067E-05
610				6.08	-0.00049006

Figure 8. Portion of a spreadsheet used to calculate the pI of the peptide DIAK. This same spreadsheet can be used to calculate the pIs of other peptides or proteins as discussed in the text.

(–1)*(B5)*(10^(D2–C5)/(10^(D2–C5)+1)) +

(1)*(B6)*(10^(C6–D2)/(10^(C6–D2)+1)) +

(–1)*(B7)*(10^(D2–C7)/(10^(D2–C7)+1)) +

(–1)*(B8)*(10^(D2–C8)/(10^(D2–C8)+1)) +

(1)*(B9)*(10^(C9–D2)/(10^(C9–D2)+1)) +

(1)*(B10)*(10^(C10–D2)/(10^(C10–D2)+1)))". The use of dollar signs to specify various cells in the above function statement makes the cells absolute references, and the values in the cells are used repeatedly as the calculation of net charge is carried out for each of the pH values indicated in column D of Figure 8. Note that cell D2 in the above function statement is entered without the dollar signs so that this cell, as in our previous examples, is a relative reference, and the value used to calculate "Net Charge" in the adjacent cell(s) depends on the value of the relative reference, which increases as one moves down column D. For example, the first term in the above function statement indicates that the number of N-terminal residues (usually one for small peptides or monomeric proteins) is to be multiplied by positive one (the charge of the protonated form of the N-terminal residue) times the fraction of N-terminal groups in the protonated form (with the absolute reference to cell C2 indicating the assumed pK_a of the N-terminal group). The remaining terms of the function statement are analogous to this first term but refer (in Figure 8) to the number of C-terminal residues (cell B3), number of aspartate residues (cell B4) etc… in conjunction with their assumed pK_a values (column C). By constructing the spreadsheet in the manner shown in Figure 8 we can easily change the number of ionizable residues (column B) as would occur if we considered a different peptide or protein, or we could change the assumed pK_a values of the ionizable groups (column C) as will be discussed below. In either case, changing the values in columns B or C leads to concomitant

changes in column E (once we have filled in this column initially), and all we have to do to find the new pI estimate is to scroll down column E until we find the sign change.

Since, in our present example, a sign change occurs between cells E609 and E610 of column E, we conclude that the pI of the peptide DIAK is between 6.07 and 6.08, but likely closer to 6.07 since the corresponding net charge in cell E609 is closer to zero than is the corresponding net charge in cell E610. To know more precisely the value of the pI we could always decrease the incremental changes between cells in column D of Figure 8 (e.g., change 0.001 units between cells rather than 0.01 units), but such precision is unwarranted given the assumptions and approximations used in the calculations (e.g., all ionizable residues of a particular type have the same pK_a in a peptide or protein, etc…). Indeed, estimations of the pI beyond the first decimal point are overly optimistic but will be used to compare values obtained with the spreadsheet approach to those of other methods. For example, the value of 6.07 compares very well with the value of 6.05 that was obtained above by averaging the pK_as of the two residues that bracketed the neutral species. The slight difference is due to the influences of the other two other ionizable groups, which were ignored in the first method.

At this point, it is reasonable to consider how the spreadsheet approach of estimating pIs compares with other methods. A search with Google using the search terms “isoelectric point calculation” yields many possibilities. To avoid the risk of “link-rot”, however, we will compare the present method with two other methods that are likely to remain in use for some time. The first method is part of the ExPASy suite mentioned above in reference to column 2 of Table 2 (*14*, *15*). The ExPASy proteomics server has been in existence for over 15 years and therefore should be a good resource. The programs “pI/Mw” and “ProtParam” use the same algorithm to calculate pIs, but each has its own unique features. For example, the program “pI/Mw” accepts

entries as short as a single amino acid, and the output consists of the calculated pI and molecular weight. The program “ProtParam” requires that the input sequence contain at least five amino acid residues. The output for this program, however, provides more information than does the output for “pI/Mw”. This output includes estimates of the pI and molecular weight, but also the extinction coefficient, the amino acid composition, and many other parameters. The amino acid composition is very useful for longer polypeptides and proteins, and in fact a weakness of the spreadsheet approach is that in the absence of such a program one would have to visually inspect the sequence of a protein and count the number of ionizable residues of each type before entering this information into column B of Figure 8, which would be a very time-consuming task. The pK_a values used by the “pI/Mw” and “ProtParam” programs were primarily determined from experiments done at 25 °C and in 9.8 M urea since these conditions match the denaturing conditions that prevail during the first dimension of a 2-D electrophoresis experiment. A pK_a of 7.5 is used for most α-ammonium N-terminal groups, but different values are assumed when the following residues are at the N-terminus: Ala (7.59), Met (7.00), Ser (6.93), Pro (8.36), Thr (6.82), Val (7.44), Glu (7.70). In addition, the side-chains of Asp and Glu are assumed to have different pK_as (4.55 and 4.75, respectively) when they occur at the C-terminus (*12*, *13*). To run the “pI/Mw” and “ProtParam” programs, one has only to go to the ExPASy website (*15*), select the appropriate program, enter the amino acid sequence of the peptide or protein, and then select the “compute” tab.

The second internet-based method of estimating pK_as is, as mentioned, from the EMBOSS suite of programs and was established in 2000 (*16*, *17*). The data in the third column of Table 2 represent the pK_a values from the default “*Epk.dat*” file that is used by the programs “iep” and “pepstats”. Unlike the ExPASy programs mentioned above, to run “iep” and “pepstats”

one must first download and install the EMBOSS suite of programs (*17*). Once EMBOSS is installed, then one can select either program and proceed in a similar manner to that described above for “pI/Mw” and “ProtParam”. Both “iep” and “pepstats” can accept inputs consisting of a single amino acid.

When the pI of the tetrapeptide DIAK is estimated with “pI/Mw” the pK_a estimate is 5.84. If the same pK_a values that these programs use (second column of Table 2) are used in place of the values in column C of Figure 8, then the pK_a estimate determined by the spreadsheet approach also is 5.84. Similarly, when the programs “iep” or “pepstats” are used to estimate the pI of the tetrapeptide DIAK, the estimate is 6.34, which again agrees with the value obtained when the spreadsheet approach is used, provided that the pK_a values in column C of Figure 8 are replaced by the values in the third column of Table 2. Thus, we see that the spreadsheet approach compares favorably with the two methods freely available on the internet, at least in the case of the above tetrapeptide. Indeed, it appears to produce identical results to the two internet-based methods provided that the appropriate pK_a values are used.

What about the case of a longer polypeptide, say a full-length protein that has 124 amino acid residues? If we obtain the sequence of bovine pancreatic ribonuclease A (from any number of protein databases that are found on the internet; see, for example, (*18*)) and “paste” the amino acid sequence into the window provided by “ProtParam”, then we obtain a pI estimate of 8.64. (The sequence pasted includes residues 27 – 150 since the first 26 residues are not incorporated into the finished protein – i.e., these residues are post-translationally removed.) We also get the amino acid composition, which we can use in column B of Figure 8 and thus use our spreadsheet, with the appropriate pK_a values. When we do this, we also get an estimate of 8.64. If we use “iep” or “pepstats” we obtain a slightly lower pI estimate of 8.39. Use of the

spreadsheet approach (again, with the appropriate pK_a values) also yields a pI estimate of 8.39. These values are somewhat close to the experimental value of 9.6 (*19*), especially given the above-mentioned assumptions of uniform pK_a values for all residues of a particular type etc…, and thus would be useful in determining initial conditions for a possible ion-exchange chromatography step in the purification of the enzyme. But we can do better, however, if we recall that ribonuclease A has its eight cysteine residues involved in four disulfide bonds as discussed in most introductory biochemistry textbooks (for example, see, (*11*)). When we manually remove the eight cysteines from the sequence prior to running the programs we obtain estimates of 9.36 and 9.83 using "ProtParam" and "pepstats", respectively. We obtain the same results with the spreadsheet approach (using the appropriate pK_a values) by entering a zero into cell B7 of Figure 8 rather than entering the number 8. Thus, the calculated estimates of the pI agree more closely with the experimental value when we incorporate additional information about the protein prior to running the programs or using the spreadsheet.

Granted, information about a protein such as the number if disulfide bonds is not always available, but when it is we can take advantage of it. Again, if we wanted to purify ribonuclease A in the absence of such additional knowledge (but still knowing the sequence), then the initial estimates would still be useful in this regard. For example, since the initial pI estimate was close to 8.5, we might decide to use a cation-exchange resin (which carries a negative charge) equilibrated in a low-salt buffer at pH 7.5. Since this pH is less than the estimated pI of our protein, we anticipate that our protein would mostly be positively charged and thus would bind to the resin; any contaminating proteins that had isoelectric points lower than 7.5 would be negatively charged at this pH value and thus they would not bind to the resin, which would help to separate our protein from some of the contaminating proteins.

Our final example is a complex protein that is made up of more than one subunit. Such a protein, as mentioned, is called an oligomeric protein. The protein we will consider is the important oxygen-transporting protein hemoglobin. Specifically, hemoglobin is a heterotetramer; it is made up of two polypeptides of one type (called α-chains) and two polypeptides of a different type (called β-chains). The sequence of an α-chain is similar but not identical to that of a β-chain. How would we enter the sequence of such a protein into one of the above computer programs? There does not seem to be an immediately obvious answer. If one enters the sequence of the mature α-chain of bovine hemoglobin into the "ProtParam" program of ExPASy, a pI estimate of 8.19 is obtained. If one enters the sequence of the β-chain of bovine hemoglobin into the same program, a pI estimate of 7.02 is obtained. The experimental value of the pI is 6.77 (*20*), and only the latter of the above two estimates is somewhat close to this experimental value. The average of the two estimates is 7.60, but there is no reason to expect that this value would approximate the pI. Alternatively, one could "splice together" the sequence of the α-chain immediately followed by the sequence of the β-chain to form an αβ sequence that, in a sense, should be representative of an $\alpha_2\beta_2$ tetramer. When one does this splicing and runs the programs, a pI estimate of 8.00 is obtained. Again, there is no sound reason why this value should approximate the pI. In fact, this splicing together of the sequences neglects the fact that in an αβ heterodimer there are two N-termini and two C-termini.

The spreadsheet approach, however, is flexible enough to treat a complex protein such as hemoglobin. We start by indicating the appropriate number of N- and C-termini (cells B2 and B3, respectively, of Figure 8) and then adding the respective numbers of ionizable residues of, in this example, the α- and β-chains. When we do this we obtain a pI estimate of 7.62. This estimate is not an improvement over the above attempts. As discussed in most introductory

textbooks of biochemistry, some amino acid residues in hemoglobin have pK_a values that are significantly different from the "typical" values that we might expect. For example, at least two histidine residues (per αβ heterodimer) that form "salt-bridge" interactions have pK_a values close to 7 rather than the more typical value of 6 (*11*). If we subtract two histidine residues from our combined α- and β-sequences and then add a term to the function statement that includes the ionization of two imidazole groups with pK_as of 7, we obtain a new pI estimate of 7.75, which is even further away from the experimental value. But we also learn in introductory biochemistry texts that the amino groups of the N-termini of deoxygenated hemoglobin can form carbamates with CO_2 (*11*). These modified termini are estimated to have pK_a values of 5.3 (*21*). When we include one or two of these modified N-termini (along with their estimated pK_a values) into our spreadsheet, we obtain pI estimates of 7.25 and 6.84, respectively. This last estimate is fairly close to the above-mentioned experimental value of 6.77.

Although hemoglobin is a very complex protein, the spreadsheet approach proved to be amenable to the task of providing a satisfactory estimate of its pI. More importantly, such built-in flexibility in the hands of students would allow them to experiment with other complex proteins and see if they can satisfactorily estimate the pIs of these proteins as well. Such an exploration could form the core of assignments that require students to apply their knowledge of the acid-base chemistry of biomolecules and to become familiar with some of the programs and databases that are freely available on the internet.

Summary

We started with simple, weak acids and showed how a spreadsheet program could be used to facilitate understanding of the acid-base chemistry of these molecules. We then proceeded to amino acids and finally to peptides and proteins. Along the way, we saw that the spreadsheet approach was facile and even adaptable to working with complex proteins. Since the construction of these spreadsheets requires that the students enter their own equations, it is hoped that this “learning by doing” approach will help enhance the learning of these important concepts, which are essential for a fuller and deeper appreciation of many aspects of biochemistry in general.

Acknowledgment

The author gratefully acknowledges Paul F. Cook for many helpful comments and suggestions.

Literature Cited

1. D'Andrea, G.; Di Nicolantonio, G. *J. Chem. Educ.* **2002**, *79*, 972–975.

2. Sokolik, C. W. *J. Chem. Educ.* **1998**, *75*, 1500–1502.

3. National Research Council, *How people learn: brain, mind, experience, and school*, expanded edition, The National Academies Press, Washington D. C., 2000.

4. Dawson, R. M. C.; Elliot, D., C.; Elliot, W. H.; Jones, K. M. *Data for Biochemical Research*, 3rd ed., Oxford Science Publications, Oxford, 1986.

5. Hall, H. K. J., *J. Amer. Chem. Soc.* **1957**, *79*, 5441–5444.

6. Nelson, D. L.; Cox, M. M. *Lehninger Principles of Biochemistry*, 4th ed., Worth Publishers, New York, 2005.

7. Wood, W. B.; Wilson, J., H.; Benbow, R. M.; Hood, L. E. *Biochemistry: A Problems Approach*, 2nd ed., The Benjamin/Cummings Publishing Company, Menlo Park, 1981.

8. Frey, P. A.; Kokesh, F. C.; Westheimer, F. H., *J. Amer. Chem. Soc.* **1971**, *93*, 7266–7269.

9. Kokesh, F. C.; Westheimer, F. H., *J. Amer. Chem. Soc.* **1971**, *93*, 7270–7274.

10. Schmidt, D. E. J., and Westheimer, F. H., *Biochemistry* **1971**, *10*, 1249–1253.

11. Berg, J. M.; Tymoczko, J. L.; Stryer, L. *Biochemistry*, W. H. Freeman, New York, 2007.

12. Bjellqvist, B.; Basse, B.; Olsen, E.; Celis, J. E., *Electrophoresis* **1994**, *15*, 529–539.

13. Bjellqvist, B.; Hughes, G. J.; Pasquali, C.; Paquet, N.; Ravier, F.; Sanchez, J.-C.; Frutiger, S.; Hochstrasser, D., *Electrophoresis* **1993**, *14*, 1023–1031.

14. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M. R.; Appel, R. D.; Bairoch, A., *Protein Identification and Analysis Tools on the ExPASy Server*, Humana Press Totowa, NJ, 2005.

15. ExPASy Proteomics Server. http://br.expasy.org/ or http://ca.expasy.org/ (accessed Jan 2009).

16. Rice, P.; Longden, I.; Bleasby, A., *Trends in Genetics* **2000**, *16*, 276–277.

17. emboss. http://emboss.sourceforge.net/ (accessed Jan 2009).

18. UniProt. http://www.uniprot.org/ (accessed Jan 2009).

19. Tanford, C.; Hauenstein, J. D., *J. Amer. Chem. Soc.* **1956**, *78*, 5287–5291.

20. White, H. L.; Monaghan, B. R., *J. Biol. Chem.* **1936**, *113*, 371–374.

21. Roughton, F. J.; Rossi-Bernardi, L. *Studies on the pK and Rate of Dissociation of the Glycylglycine-Carbamic Acid Molecule.* In *CO_2: Chemical, Biochemical, and Physiological Aspects*, Forster, R. E.; Edsall, J. T.; Otis, A. B.; Roughton, F. J., Eds.; Special Publication

*This handout originally appeared as part of an online supplement that accompanied the article "Use of a Spreadsheet to Calculate the Net Charge of Peptides and Proteins as a Function of pH: an Alternative to Using "Canned" Programs to Estimate the Isoelectric Point (pI) of These Important Biomolecules" *J. Chem. Educ.* **2010**, *87*, 803 – 808.