

An algorithm for isoelectric point estimation

David L. Tabb

Created 7/10/01

Updated 6/28/03

1 Introduction

One of the most common techniques for separating mixtures of proteins is the two-dimensional polyacrylamide gel. In these gels, proteins are separated in one dimension based on their electro-migration speeds through the gel (roughly determined by their molecular weights) and in another dimension by the pH at which their side chains cumulatively amount to a neutral charge, called the isoelectric point. As a result, it is often useful to know the molecular weight and isoelectric point values for a protein. Both of these values can be estimated from the protein sequence. These estimates are not generally exact because many proteins are chemically modified after they are assembled by ribosomes, sometimes leading to mass or charge differences in the final protein product.

Molecular weights (MW) are simple to calculate. The masses for the amino acids in the protein are first summed, and then the mass of water is added. If any modifications to the protein structure are known, these masses should be added to the sum as well. One common modification clips the first amino acid from the protein (usually a methionine due to the way in which protein sequences are encoded by DNA). By summing together all the masses for the elements of the protein, the mass for the whole is calculated.

Isoelectric point (pI), on the other hand, is a more complex value to calculate for proteins. Generally, pI is defined as the pH at which a protein takes on a net negative charge. Proteins usually have many different ionizable groups in their structures, and so some parts may take on a negative charge while others take on a positive charge. At a particular pH, an individual amino acid's side chain will have some probability for adopting either a positive, neutral, or negative

charge. In groups of these amino acids, it simplifies matters to assess the percentage of copies for each amino acid which would adopt each of these charge states. Finding the pH at which all the charges in a protein sum to zero is the process of finding pI. This article suggests a simple but generally effective algorithm for determining the pI of a protein.

2 Algorithm

The following algorithm is implemented in the DTASelect algorithm for proteomic data mining. The first element of the program is a *calling function*, designed to count the numbers of copies of the amino acids which play a role in determining pI and to propose pH values for the other functions. The *charge determination function* determines the expected charge on the protein for a particular pH value. It, in turn, makes use of the *charge ratio determination functions*, which determine the expected proportion of charged and uncharged side chains for a particular amino acid, which are assessed from the supplied *pK values*.

2.1 The Calling Function

The two tasks to be performed by the calling function include assessing the numbers of charged groups in the protein and the proposal of pH values at which charge is to be assessed.

Several protein areas can take on a positive charge. Lysine, arginine, and histidine all possess basic side chains, and so these residues may result in positive charges on the protein. Aspartic and glutamic acids, cysteine, and tyrosine may take on negative charges. The calling function should count the numbers of each of these residues, preferably while calculating the molecular weight of the protein for efficiency. In addition, the N-terminus and C-terminus, if unmodified, may also make a contribution to the charge of the protein and thus the isoelectric point. DTASelect's pI calculator simply assumes that the termini are unmodified, but N-terminal acetylation is fairly common and could affect a protein's observed pI.

The most common way to propose pH values at which the protein's charge is calculated is to loop through a wide range of pH values in small steps. For each pH value, the protein charge is calculated, and a curve showing the charge at each pH is plotted. The pI is usually determined by interpolating between the values closest to zero charge. This approach is capable of producing a graph reminiscent of titration curves, though its accuracy for pI determination is suboptimal. This

approach is that used by Mark Southern's pICalculator, part of BioPerl.

An improved approach simulates an isoelectric focusing gel. First, the software calculates the charge of a protein at pH 7. Then next proposed pH moves 3.5 (half of 7) higher or lower depending on the charge. In the next jump, the proposal moves 1.75 (half of 3.5) in the appropriate direction. These jumps continue until the charge is appropriately close to zero. The search space is halved with each jump, resulting in an efficient search for the isoelectric point. This approach, rather than the one above, is employed by DTASelect.

2.2 The Charge Determination Function

The process of calculating the charge on the protein at a particular pH is handled by the charge determination function. In essence, the charge of the protein is equivalent to the sum of the fractional charges of the protein's charged groups:

$$Z = Nterm + Cterm + \alpha * K + \beta * R + \gamma * H + \delta * D + \epsilon * E + \zeta * C + \eta * Y,$$

where $Nterm$, $Cterm$, K , R , H , D , E , C , and Y are the charges these groups take on at a particular pH and the Greek letters before them are the count of each amino acid residue from the protein sequence. Z , the sum of these terms, is the charge on the entire protein.

This summation places some limitations on the accuracy of this algorithm. First, by adding together the charges in this way, the algorithm assumes that each group's charge is independent of all others; if the protein contains an arginine side chain and an aspartic acid side chain, for example, these two groups are assumed to take on a charge irrespective of their locations in the protein sequence. If a basic residue is adjacent to an acidic residue, each probably *does* change the other's ability to take on a charge, but this effect is ignored in this algorithm.

2.3 The Individual Charge Ratio Determination Functions

The structure and composition of a molecule determines its ability to take on a charge. Given thousands of copies of a particular ionizable chemical group, it is likely that some will and some will not take on a charge at a given pH. The proportion of molecular copies that take on a charge in response to pH changes is given by the pK values for that chemical group. The following are pK values corresponding to the side chains of amino acids:

The percentage of a group of molecules taking on a charge is computed by two functions. One determines the percentage of positive ions for the N-terminus, lysine, arginine, and histidine.

Group	DTASelect	Solomon's	Group	DTASelect	Solomon's
N-terminus	8.0	9.6*	C-terminus	3.1	2.4*
lys	10.0	10.5	asp	4.4	3.9
arg	12	12.5	glu	4.4	4.3
his	6.5	6.0	cys [ⓐ]	8.5	8.3
			tyr	10.0	10.1

Table 1: The above pK values can be used to determine the percentage of copies of a particular charged molecule which adopt a charge at a given pH. The values for DTASelect were drawn from a web page in French, now lost to history. The column marked as "Solomon's" gives the pK values shown in Solomon's *Organic Chemistry*, fifth edition. * The N-terminal and C-terminal pK values from Solomon's book were those given for leucine, typically the most common amino acid in protein sequences. When other amino acids are at the termini, they may cause the termini to ionize differently. [ⓐ] Cysteine residues have the pK values shown when they are *not* taking part in disulfide bridges. When these residues are connected to other such residues, they are cystines rather than cysteines, and cystines do not take on charge.

The other function determines the percentage of negative ions for the C-terminus, aspartic acid, glutamic acid, cysteine, and tryosine. Each function generates a concentration ratio (CR). For positive groups,

$$CR = 10^{pK-pH}.$$

Negative ions reverse this order:

$$CR = 10^{pH-pK}.$$

In other words, the further the pH is from the pK of the group, the more likely the group is to be positively charged. Once the CR is generated, the function can return the partial charge by returning this value:

$$\frac{CR}{CR + 1}.$$

In each case, the calculated partial charge is absolute; the function which calls this subfunction should assume the charges from the C-terminus, asp, glu, cys, and tyr to be negative, while the charges from the N-terminus, lys, arg, and his should be assumed to be positive.

3 Weaknesses

Different algorithms may use different pK values than those given in Table 1. Experiments yield varying pK values. The set of pK values chosen will affect the resulting isoelectric points yielded by this algorithm.

This algorithm assumes each residue is completely independent of the others in taking on a charge state. This is not the case; an arginine neighboring another charged residue is likely to be influenced by it.

If a protein has a modification which bears a charge (such as a phosphate group), there may be a major influence on the pI of the protein. Modifications are ignored in the described algorithm. Modifications on the N- or C-termini are sure to have some impact on the isoelectric point.

The appropriate pK value to use for each terminus will depend on the amino acids at those locations. As written, this algorithm ignores this variance.

It is not generally feasible to state which cysteine residues are taking part in disulfide bridges from the sequence alone. In many cases, however, these cross-links are disrupted before a protein is run through a gel. The reaction which breaks a disulfide link may also modify the cysteine residue such that it adopts charge much more strongly or weakly than unlinked cysteine.