UPTEC X 03 015
MAY 2003
ISSN 1401-2138

ULRIKA SKARP

Relating surface potentials to cation and anion exchange retention of proteins

Master’s degree project

UPPSALA UNIVERSITET

Molecular Biotechnology Programme

Uppsala University School of Engineering

UPTEC X 03 015	Date of issue 2003-05

Author

Ulrika Skarp

Title (English)

Relating surface potentials to cation and anion exchange retention of proteins

Title (Swedish)

Abstract

This work reports the development of a new program (SCARP) for the calculation of electrostatic potentials in proteins. The program is based on the linearized Poisson-Bolzmann equation and takes protein structure (PDB) files as input. Several electrostatic potentials are calculated by the program. Among these, the average surface potential has been used as a descriptor for the modelling of ion exchange chromatography. For cation exchange, previously published results describing a clear relation between average surface potentials and retention times have been reproduced and extended to a wider set of proteins. For some proteins, the same principles seem to apply also for anion exchange chromatography, whereas for other proteins this is not the case.

Keywords

Electrostatic potential, Poisson-Bolzmann equation, ion-exchange chromatography, proteins

Supervisors

Dr. Enrique Carredano

R&D protein separations, Amersham Biosciences

Scientific reviewer

Dr. Gerard Kleywegt

Dept. of Cell and Molecular Biology, Uppsala University

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **39**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala

Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

Relating surface potentials to cation and anion exchange retention of proteins

ULRIKA SKARP

Sammanfattning

Proteiner har på senare år börjat användas allt oftare, bland annat i läkemedelsindustrin. Det är därför viktigt att kunna rena proteiner. En vanligt förekommande metod för detta är jonbyteskromatografi, vilken bygger på att proteiner i de allra flesta fall är elektriskt laddade och att olika proteiner har olika laddning. För att separera en blandning av olika proteiner får de passera genom en kolonn (ett rör), fylld med ett elekriskt laddat, poröst material. De proteiner som har en laddning motsatt den som finns i kolonnen kommer att fördröjas på sin väg genom kolonnen. Genom att variera pH och saltkoncentration kan man få ut proteinerna i en serie efter varandra.

I detta arbete har försök gjorts att skapa en modell som förklarar fördröjningen hos de olika proteinerna. Modellen bygger på beräkningar av den elektrostatiska potentialen på proteinernas ytor. Detta beskriver hur laddningarna hos proteinet förväntas påverka en annan laddning på proteinytan och kan användas som ett mått på hur starkt proteinet påverkas av laddningarna i kolonnen. I beräkningarna tas hänsyn till proteinets form och laddningsfördelning, pH och saltkoncentration, samt att proteinerna befinner sig i vattenlösning.

Examensarbete 20 p
Civilingenjörsprogrammet i molekylär bioteknik
Uppsala universitet, maj 2003

Table of contents

1. Introduction

Ion exchange chromatography is a widely used method in protein separation. The method is based on retention of charged protein molecules in the chromatographic column by adsorbents of opposite charge. The degree of protein retention depends on protein-specific features that are affected by experimental conditions such as pH and salt concentration. Changing the pH changes the charge of the protein, which usually affects the retention. Salt ions in the buffer may occupy some of the binding sites at the stationary adsorbents, making fewer binding sites available for protein binding (Figure 1). Ions may also bind to charged protein residues, neutralizing some of the protein charges. With all experimental parameters fixed, some proteins would not be eluted from the chromatographic column in reasonable time. Thus, the experimental procedure normally includes varying at least one parameter, typically the salt concentration (Scopes, 1994).

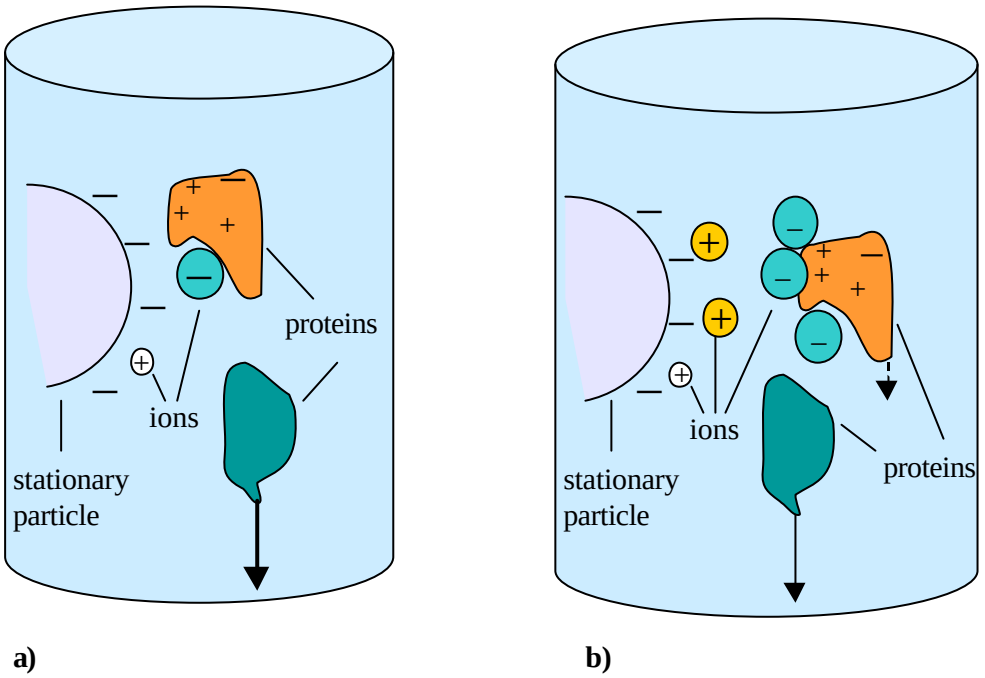

Figure 1 A schematic picture of an ion-exchange column: a) when only a few ions are present, the charges of the protein interact strongly with the charged column particles, causing retention. **b)** with more ions in the solvent buffer, the charges are shielded, causing a loss of retention.

Optimal separation conditions need to be found for each ion exchange experiment, which is very time-consuming and expensive. A great deal of time could be saved, however, by making a theoretical estimate of the adsorption behavior of the protein of interest before carrying out any experiments. Ideally, one would like to establish a computational model that outputs the chromatographic retention time given protein sequence or structure and experimental conditions and that is valid for all proteins. Such experimental conditions might be pH, salt concentration, type of salt, temperature, etc.

Early attempts to model the retention process theoretically were based on the net charge of the protein (Kopaciewicz *et al.*, 1983). The protein charges depend on pH and protein structure and can be approximated fairly well based on amino acid sequences and standard pK_a values of free amino acids. However, the retention does not depend only on the existence of charges but also on how the free energy of the protein molecules is affected by the chromatographic environment. The total free energy can be separated into different contributions such as electrostatic interactions, van der Waals interactions, hydrogen bond contributions, internal bond energies and kinetic energy caused by the flow pressure.

One possible model of this system is obtained using molecular dynamics. Here, all molecules, or even atoms, in the system are considered as being separate units that interact through the energies mentioned above. However, due to the enormous number of separate units, this would require unreasonably long calculation times given the computer resources available today. This has encouraged scientists to try to neglect some of the contributions to the total energy. One way of finding protein characteristics that are important in a particular case is to calculate a large amount of energy-related characteristics (descriptors) for a set of proteins. Such descriptors might be anything that is protein-specific and, in some way, measurable or possible to calculate from measurable data. Examples of possible descriptors are the molecular weight, the isoelectric point and the fraction of hydrophobic groups on the protein surface. Data-mining techniques (*e. g.*, partial least squares) can then be used to select the most significant descriptors for predicting the behavior of the system. In further studies, only these descriptors need to be taken into account. Results using this approach to predict anion exchange retention times suggest that descriptors connected to charge and electrostatic properties are of high importance in this application (Song et al, 2002). However, in these calculations, the protein has been considered as being surrounded by vacuum, which has different electrostatic properties than water.

The results of Song *et al.* suggest that every term except the electrostatic contribution can be neglected in predictions of ion exchange chromatography retention. All terms except the electrostatic potential can also be regarded as being either protein independent (such as the kinetic contributions), identical in the stationary and mobile phases (for instance internal bonds) or very small compared to the electrostatic term (*e. g.*, hydrogen bonds and van der Waals interactions in ion exchange conditions). Thus, the electrostatic potential on or around the protein surface could be used to obtain a protein-specific measure of the degree of retention in ion exchange chromatography. In this model it is assumed that there are always stationary charges in excess. This would be valid at certain pH and salt concentrations for each type of stationary particle.

In order to calculate the electrostatic potential, a mathematical expression needs to be derived. Depending on any assumptions that are reasonable in the system and on the level of detail chosen, a more or less complex expression can be used. A frequently used model is to consider one part of the system (usually the solvent) as a continuum with constant properties such as a dielectric constant. Using this model to calculate electrostatic potentials for proteins in a water-based salt solution, the Poisson-

Bolzmann equation (PBE) is generally held to give reliable results. This equation takes into account the dielectric differences between water and protein as well as the statistical positioning of buffer ions at the protein surface. The PBE is rather complex and can only be solved analytically for a limited number of model geometries. Thus, one possible way of solving the PBE is to model the proteins as planes, cylinders or spheres. Calculations using this approach have predicted a correlation between the retention factor and the logarithm of the salt concentration in ion-exchange chromatography of proteins (Ståhlberg *et al.*, 1991). However, these models do not account for differences among proteins that depend on their structure.

Another approach is to use the atomic structure of the protein and to employ numerical schemes to solve the PBE. These calculations could be rather computationally demanding. However, the time requirements for the calculations can be reduced by linearization of the equation. It can be shown mathematically that the solution to the linearized PBE is similar to the solution of the original equation, provided that the value of the potential is of an appropriate order of magnitude. This is not always true for the electrostatic potential of proteins. Yet, linearization of the PBE has been shown not to give rise to significant errors when calculating the electrostatic potential of moderately charged molecules like proteins, even when the potential is locally quite large (Fogolari *et al.*, 1999). For highly charged molecules, however, the original equation or the linearized equation with a correction factor should be used. An algorithm to solve the linearized Poisson Bolzmann equation (LPBE) on proteins in aqueous solutions was elaborated in 1986 (Klapper *et al.*, 1986). This procedure was further developed and implemented in the software DelPhi (Nicholls and Honig, 1990). DelPhi and other programs built on the same concept have been widely used in different biophysical applications and areas such as the study of pK_a shifts (Yang *et al.*, 1993), calculations of solvation free energies and enzyme kinetics (Honig and Nicholls, 1995, and references therein).

The solution of the LPBE can also be used in theoretical predictions of protein ion exchange retention times. The mean electrostatic potential of the protein surface has been shown to correlate linearly with cation exchange retention times in a set of eight different proteins (Haggerty and Lenhoff, 1991). However, no such correlation has yet been found for anion exchange data.

Recent studies using confocal microscopy has provided a deeper insight to the protein uptake in stationary ion exchange particles (Ljunglöf, 2002). This method makes it possible to visualize how proteins penetrate into the stationary column particles, which is an important factor for protein retention. These studies using have revealed the formation of radial concentration rings, which shows that proteins do not penetrate into the particles in a uniform way. Rather, rings (or shells) of high protein concentration followed by shells with almost no protein move into the particles. This phenomenon cannot be explained by diffusion or related approaches. However, it has been suggested (Ljunglöf, 2002) that the electrostatic potential of the stationary particles and proteins could offer a reasonable explanation to the formation of radial concentration rings.

The aim of this work is to further study the correlation between cation exchange data and electrostatic potential in proteins found by Haggerty and Lenhoff. Today there are more structure and retention data available that could be used to estimate the applicability of the electrostatic potential model. Moreover, several ways of improving this model by taking further aspects into account will be evaluated. This will hopefully yield results that can also account for the behavior of proteins in anion retention processes.

2. Theory

2.1. ELECTROSTATIC POTENTIALS

For a homogeneous, uncharged system with a uniform dielectric constant, the electrostatic potential is described by the Poisson equation (mathematically a Laplace equation).

$$\nabla(\nabla\Phi(\mathbf{r}))=0 \qquad \text{(eq1)}$$

where Φ is the electrostatic potential and ∇ is the gradient operator.

Proteins have, in almost all cases, several charged groups. Furthermore, the buffer used in chromatographic experiments contains charged ions. When charges are present in the system, the right-hand side of eq1 is no longer zero. Instead, the system is described as

$$\nabla(\nabla\Phi(\mathbf{r}))=\rho(\mathbf{r})/\varepsilon \qquad \text{(eq2)}$$

where $\rho(\mathbf{r})$ is the charge distribution and ε is the material-dependent permittivity ($\varepsilon = \varepsilon_r * \varepsilon_0$, where ε_r is the relative dielectric constant and ε_0 is the permittivity of vacuum). Eq2 is only applicable in a homogeneous system, *i. e.*, when the dielectric value can be regarded as a constant. Proteins and buffer solutions have very different dielectric properties, which must be taken into account. Thus, the dielectric constant ε_r is regarded as a variable and the Poisson equation is written

$$\nabla\bullet\varepsilon(\mathbf{r})\nabla\Phi(\mathbf{r})=\rho(\mathbf{r}) \qquad \text{(eq3)}$$

When considering a charged macromolecule in a solvent containing salt, the charge density can be separated into two components, one arising from the macromolecule and one arising from the salt. The charge distribution of the salt can be modeled by a Bolzmann distribution:

$$\rho(\mathbf{r})=\rho_{macro}(\mathbf{r})+\rho_{salt}(\mathbf{r})=\rho_{macro}(\mathbf{r})+\Sigma cN_A qe^{(-q\Phi(\mathbf{r})/kT)} \qquad \text{(eq4)}$$

where ρ_{macro} is the charge distribution in the protein, c is the bulk salt concentration, q is the charge on each ion, N_A is the Avogradro constant, k is the Bolzmann constant and T is the temperature. The sum is calculated over all salt species present in the solution.

Combining eq3 and eq4 gives the Poisson-Bolzmann equation

$$\nabla\bullet\varepsilon(\mathbf{r})\nabla\Phi(\mathbf{r})= \rho_{macro}(\mathbf{r})+ \Sigma cN_A qe^{(-q\Phi(\mathbf{r})/kT)} \quad \text{(eq5)}$$

If there are only two types of buffer ions, one positive and one negative, both being monovalent, q=+*e* or –*e*. Then the sum on the right hand side of eq5 can be rewritten

$$cN_A\,(ee^{(-e\Phi(\mathbf{r})/kT)}-eqe^{(e\Phi(\mathbf{r})/kT)})= -2cN_A e\sinh(e\Phi(\mathbf{r})/kT) \quad \text{(eq6)}$$

Taylor expansion around the point Φ=0 gives

$$\rho_{salt}(\mathbf{r})= -2cN_A e\sinh(e\Phi(\mathbf{r})/kT)=$$
$$=-2cN_A e\,((e\Phi/kT)+(e\Phi/kT)^3/3!+(e\Phi/kT)^5/5!+\ldots.) \quad \text{(eq7)}$$

where all terms except the first (linear) term can be neglected if $(e\Phi/kT) \ll 1$

Proceeding in this way, the remaining term becomes

$$\rho_{salt}(\mathbf{r})= -2cN_A e(e\Phi/kT)=-2(cN_A e^2/kT)\Phi \quad \text{(eq8)}$$

Thus, the linearized PBE can be rewritten as

$$\nabla\bullet\varepsilon(\mathbf{r})\nabla\Phi(\mathbf{r})= \rho_{macro}(\mathbf{r})-\varepsilon_0\varepsilon_r\kappa^2\Phi \quad \text{(eq9)}$$

where

$$\kappa^2=2e^2N_A c/kT\varepsilon_0\varepsilon_r \quad \text{(eq10)}$$

In order to get an explicit formula for Φ, all terms are integrated over a small, cubical volume segment with side length h. Integrating the right-hand side gives:

$$???(\rho_{macro}(\mathbf{r}) - \varepsilon_0\varepsilon_r\kappa^2\Phi)d\mathbf{r}= q - h^3\varepsilon_0\varepsilon_r\kappa^2\Phi \quad \text{(eq11)}$$

where q is the charge inside the volume segment.
The left-hand side of eq9 can be rewritten using Gauss' theorem

$$???\nabla\bullet\varepsilon(\mathbf{r})\nabla\Phi(\mathbf{r})d\mathbf{r}=??\varepsilon(\mathbf{r})\nabla\Phi(\mathbf{r})d\mathbf{n} \quad \text{(eq12)}$$

where the double integral is calculated over the surface of the cubical volume segment and d**n** is the normal vector element to the surface. $\varepsilon(\mathbf{r})$ can be approximated by the value of ε at the center of each area segment. The normal component of the gradient can be rewritten using a standard central difference scheme:

$$\nabla\Phi(\mathbf{r_1})*\mathbf{n}=(\Phi(\mathbf{r_0})-\Phi(\mathbf{r_2}))/h \quad \text{(eq13)}$$

where r_1 is the midpoint between r_2 and r_0. Integration of these constant terms on each surface element gives a factor of h^2. Since h is small, ε may be regarded as a constant on the surface element and $\varepsilon(\mathbf{r})$ can be written as $\varepsilon_0\varepsilon_r$. Thus, the right hand side of eq12 becomes

$$??(\varepsilon_0\varepsilon_r(\Phi(\mathbf{r_0})-\Phi(\mathbf{r_2}))/h)(d\mathbf{r})^2=\Sigma(\varepsilon_0\varepsilon_r(\Phi(\mathbf{r_0})-\Phi(\mathbf{r_i}))*h) \qquad \text{(eq14)}$$

where the sum is calculated over all six quadratic surface areas. $\mathbf{r_0}$ refers to the central grid point in the cubical volume element, $\mathbf{r_i}$ refers to each of the six neighbor grid points. Combining eq9, eq11 and eq14 gives

$$\Sigma(\varepsilon_0\varepsilon_r(\Phi(\mathbf{r_0})-\Phi(\mathbf{r_i}))*h)= q- h^3\varepsilon_0\varepsilon_r\kappa^2\Phi \qquad \text{(eq15)}$$

which can be rearranged to

$$\Phi(\mathbf{r_0}) = \frac{\Sigma\,\varepsilon_r\Phi(\mathbf{r_i})+ q/(h\varepsilon_0)}{\Sigma\,\varepsilon_r + \varepsilon_r(\kappa h)^2} \qquad \text{(eq16)}$$

Eq16 can be solved using a stationary iterative method (Heath, 1997). Using the Jacobi iterative method, eq16 would be solved using the scheme:

$$\Phi(\mathbf{r_0})_{i+1} = \frac{\Sigma\,\varepsilon_r\Phi(\mathbf{r_i})_i + q/(h\varepsilon_0)}{\Sigma\,\varepsilon_r + \varepsilon_r(\kappa h)^2} \qquad \text{(eq17)}$$

However, faster convergence is achieved using the Gauss-Seidel method:

$$\Phi(\mathbf{r_0})_{i+1} = \frac{\Sigma\,\varepsilon_r\Phi(\mathbf{r_i})_k + q/(h\varepsilon_0)}{\Sigma\,\varepsilon_r + \varepsilon_r(\kappa h)^2} \qquad \text{(eq18)}$$

where k is either i or i+1 depending on whether $\phi(\mathbf{r_i})_{i+1}$ has yet been calculated. This scheme has been shown to converge to the same solution independently of the initial estimate Φ_0 (Nicholls and Honig, 1990).

2.2. CHARGES

The charges, q in the above formulae, originate from charged points in the protein, *i.e.*, acidic or basic side chains, backbone terminals and bound ions. All charges except the bound ion contributions are pH-dependent. This is described by the Henderson – Hasselbach equation:

$$pH = pK_a + \log([A^-]/[HA]) \qquad \text{(eq19)}$$

The side-chain pK_a values can be approximated by standard pK_a values of the corresponding free amino acids. In order to obtain the fraction of the charged form given pH and pK_a, eq19 can be rewritten

$$\frac{[A^-]}{[A]_{tot}} = \frac{[A^-]}{[A^-]+[HA]} = \frac{1}{1+10^{(pH-pKa)}} \qquad \text{(eq20)}$$

In the finite difference scheme, q relates to the total charge inside the volume segment. Thus, one possible model would be to assign the charge obtained in eq20 to the nearest grid point. This would correspond to the charged atom being entirely inside the volume segment surrounding this grid point. However, the charged group is not confined to a point in space, but has an extent. This can be taken into account using a trilinear weighting formula (Klapper *et al.*, 1986) to divide the charge between the eight closest grid points.

$$q_{gridpoint}=q_{atom}(1-a/h)(1-b/h)(1-c/h) \qquad \text{(eq21)}$$

where a, b and c are the distances in the x, y and z directions, respectively, between the charged atom and the grid point and h is the grid spacing.

This is a fairly good approximation when the grid spacing is rather big (about an atom radius). However, when the distance between two grid points is much smaller, another formula (*e.g.*, a Gaussian distribution) might be a better approximation.

3. Materials and methods

3.1. Outline

Electrostatic potentials were calculated using a three dimensional grid based finite difference scheme (eq18). This formula uses known values of charge, dielectric constants and salt concentration at discrete locations in space (grid points). Thus, in order to calculate the electrostatic potentials, these three parameters need to be known at each grid point before solving the Poisson-Bolzmann equation.

The parameters depend on whether the grid point is inside the protein and if there are any nearby charged groups. This information can be derived from PDB-files containing protein structures. All protein structures used in this work were obtained from the Protein Data Bank (PDB, 2003), some with a few modifications to the original structure.

The implementation was done using C++ Builder v. 5 (Borland Inprise). This resulted in the program SCARP (relating Surface potentials to Cation and Anion Retention of Proteins). All calculations with the new program were carried out on a Pentium II PC (Compaq Deskpro) with 128 MB RAM.

3.2. BOUNDARIES

To solve differential equations appropriate boundary conditions are needed. Since the potential should vanish at an infinite distance from the protein, setting the potential to zero at a long distance from the protein would be a reasonable boundary condition. However, this would require an enormous number of grid points, which would be very demanding computationally. Using a sparser grid, on the other hand, would give a large discretization error to the solution. In this work, a process known as focusing (Klapper *et al.*, 1986) has been used to set the boundary values, thereby avoiding the problems mentioned above. Using focusing, the electrostatic potential is first solved on a wide-stretched sparse grid. This solution is used to set the boundary values of a smaller grid, lying entirely within the first grid and with a much smaller value of h. Thus, an adequate solution can be obtained within limited computational time. An extra benefit of this method is that the solution of the sparse grid can be used as an initial estimate of the solutions on interior grid points in the smaller, denser grid. This is also known as a multigrid method (Heath, 1997) and speeds up the calculations further.

3.3. GRIDS

The extent of the grid and the location of the grid points are determined by the size of the protein and the grid spacing ($h=h_{init}$, obtained as an input parameter, typically ~1 Å). An initial volume, box1, is defined as the smallest possible box including the protein, one grid spacing, h, and one radius R_{max}, defined as the largest of the water and salt radii. (Figure 2). In order to implement the focusing procedure discussed above, all the electrostatic calculations are performed using two separate grids sequentially. The first grid is defined from box1 but is scaled up in all directions using an integer focusing parameter (f) obtained from input data, typically ~4. The grid spacing h is modified accordingly.

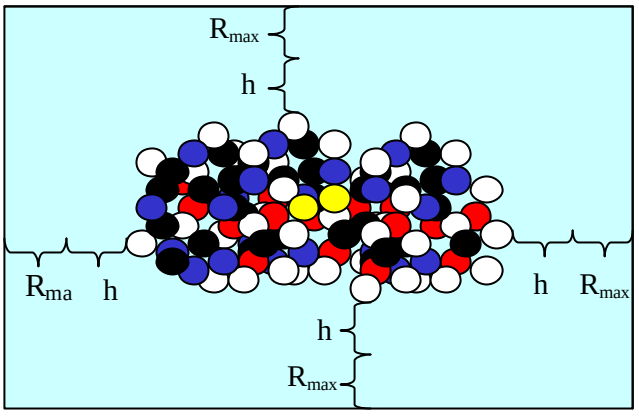

Figure 2. Two-dimensional projection of box1: box1 is the basis for setting the grid extensions and grid point locations. R_{max} is the largest of the water and salt radii and h is the grid spacing.

The second grid is defined in a similar way as the first grid with two exceptions. First, the parameter maxShell (obtained from input data, typically ~1-10 Å) is used instead of R_{max}. Second, all boundary grid points are set so as to coincide with grid points in the first grid, in order to obtain adequate boundary values. As grid spacing, h, in the

second grid, the unmodified input h (h_{init}) is used. The positioning of the two grids in relation to each other and to the protein is schematically shown in Figure 3.

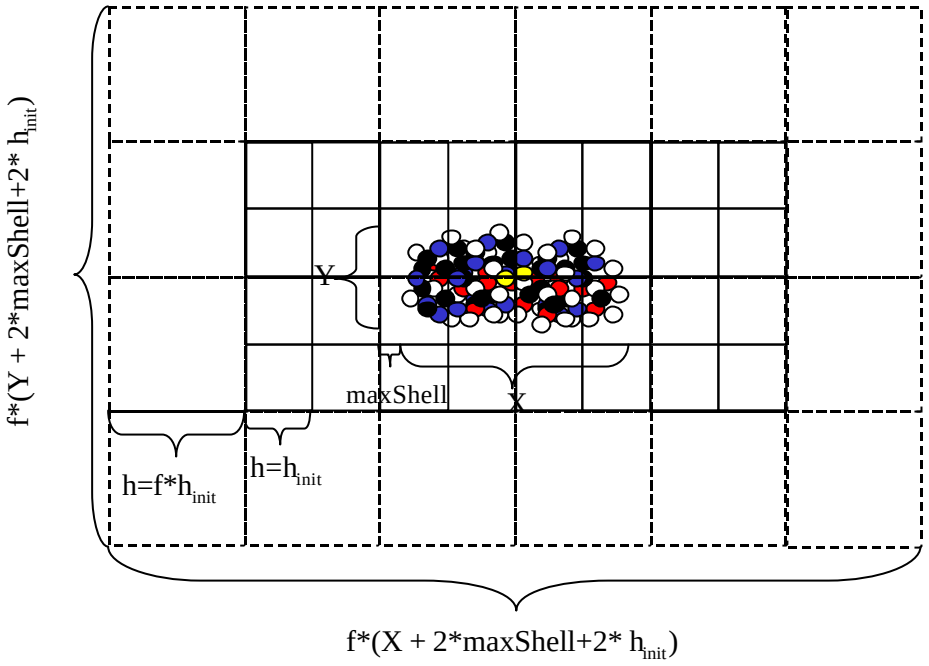

Figure 3. The location of the two grids: h_{init} is the input grid size, h is the grid size, f is the input focusing parameter, maxShell is an input parameter, X is the length of the van der Waals volume of the protein in the x-direction, Y is the corresponding length in the y-direction. The grids used in the calculations are much denser than those in the figure.

These two grids are used as a basis to find variable values (charge, dielectric constants, salt concentrations), and calculate the electrostatic potential.

3.4. IDENTIFICATION OF SOLVENT AND SALT ACCESSIBLE GRID POINTS

All grid points lying outside of box1 are considered to be accessible to both salt and solvent. All grid points lying within any of the protein atoms (defined from their van der Waals radii) are considered to be inaccessible to both salt and solvent. The atomic van der Waals radii can be found in Appendix 1.

For all other grid points, a more thorough accessibility assignment is carried out. As initial estimate, all these grid points are considered to be accessible to salt, but not to solvent. For each of the grid points, the minimal distance (d_{min}) to the protein van der Waals surface is calculated. If d_{min} is longer than the probe radius (the van der Waals radius of water), a sphere centered in the grid point and with a radius of d_{min} is considered. All grid points lying within this sphere are set to water accessible (Figure 4a). If d_{min} is shorter than the specified salt van der Waals radius, the grid point is set to salt inaccessible (Figure 4b).

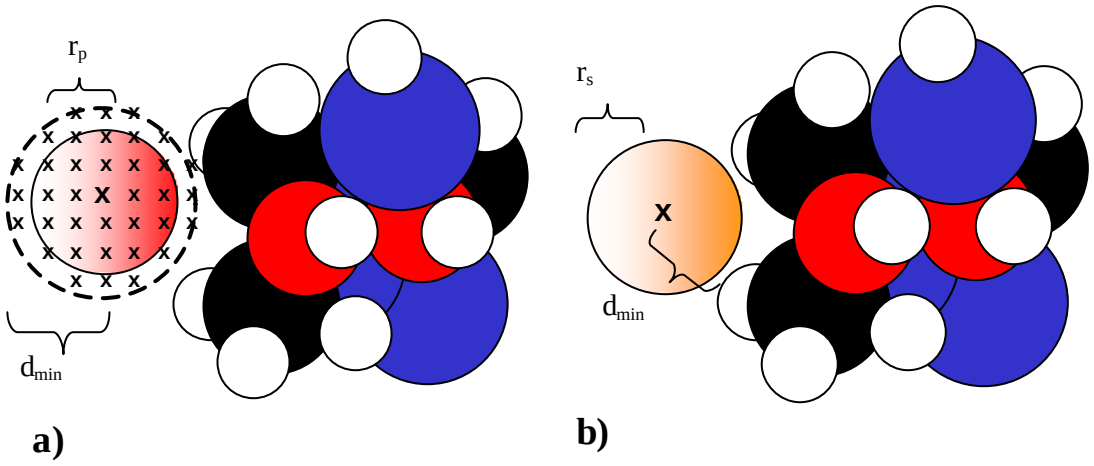

Figure 4. Definitions of accessibilities: d_{min} is the smallest distance from the grid point to the protein van der Waals surface. **a)** Accessibility of water: the dotted circle is centered in the grid point and has a radius of d_{min}. If d_{min} is longer than r_p, the probe radius, all grid points inside the dotted circle (sphere) are considered to be water accessible. **b)** Accessibility of salt: if d_{min} is longer than r_s, the van der Waals radius of the salt ion, the grid point is considered to be salt accessible

The reason for using different definitions of salt and solvent accessibility is that the behavior of the salt and water molecules is described differently by the Poisson-Bolzmann equation. The salt is regarded as a set of point charges that cannot lie too close to the protein surface. This is supposed to reflect the fact that the salt ion has a radius, which makes it impossible for the ion center to lie close to the protein surface. The solvent is regarded as a continuum. All points in space that could lie within a water molecule are assumed to have the dielectric properties of water.

3.5. IDENTIFICATION OF CHARGED GRID POINTS

Charges are calculated using standard pK_a values for amino acids in free solution (Appendix 2) and the Henderson – Hasselbach formula (eq19) for calculating charge given pH and pK_a. Charges on residues with delocalized electrons (*e. g.*, aspartic acid) are divided among the atoms involved. Cysteins are considered to be non-ionizable if there is another cystein in the vicinity. The charges on amino and carboxy terminals are also considered. All atom charges are divided among their eight closest grid points using a trilinear weighting formula (eq 21).

3.6. ELECTROSTATIC POTENTIAL

The electrostatic potential is calculated using the finite difference scheme (eq18) with the grid point charges set as described above. The salt concentration is set to the buffer salt concentration for salt accessible grid points and to zero otherwise.

The dielectric constants in eq18 are defined at the midpoints of the central grid point and the neighbor grid points. This is approximated using three possible values for the dielectric constant. When both grid points (the central grid point and its neighbor) are solvent accessible, a dielectric value of water (ε_r =80, Nordling & Österman, 1996) is used. When none of them are solvent accessible an approximate dielectric value of protein (ε_r =3, Fogolari *et al.*, 1999) is used. If one of the grid points (the central grid

point or its neighbor) is solvent accessible and the other is not, the average of the dielectric values of water and protein (ε_r =41.5) is used.

Values for the constants *e*, N_A, k and ε_0 were obtained from the Physics Handbook for Science and Engineering (Nordling & Österman, 1996). The temperature was set to 298 K.

Using a finite difference scheme, all grid points must be given a start value, the initial estimate, before calculating the electrostatic potential. In the first grid, the start value for the electrostatic potentials is set to zero in each grid point. In the second grid, the solution on the first grid can be used to interpolate a better initial estimate. After setting the initial estimate, values are updated according to eq18.

The potential is considered as having converged when the mean square difference of its value in two subsequent iterations is less than 10^{-5} kT/e. For a discussion on why this value was chosen, see Appendix 2.

3.7. EXTRA FOCUSING STEP

In order to increase the accuracy in certain points of interest, an extra focusing step was implemented, as has been previously reported (Yang et al, 1993). Here, the second grid as described above, is used to provide boundary conditions for a third grid, centered at an arbitrarily chosen point and with grid spacing and dimensions specified by input parameters. The electrostatic potentials are then calculated as described above and the resulting potential in the specified point of interest is interpolated from its neighboring grid points. This procedure may, however, give rise to inappropriate values of the dielectric constant for some points close to the edge of the grid. These points cannot be covered by any water probe centered inside the grid, and will thus always be set to water inaccessible, even if a water probe centered outside the grid could reach them.

The extra focusing step was not used except for validation purposes.

3.8. OUTPUT

SCARP outputs:

- A list of all grid points with charge, accessibility and electrostatic potential.
- A list of all protein atoms with electrostatic potentials in the atomic centra (interpolated from the eight nearest grid points).
- Surface points and their electrostatic potential. The surface is defined using two input parameters, minShell and maxShell. All grid points lying at a distance d (minShell < d < maxShell) from the protein van der Waals surface are considered to be part of the surface.
- The theoretical protein net charge and surface statistics such as average surface potential, Bolzmann weighted average surface potential, minimum and maximum surface potential.

As an option, the first three lists can be saved as molecular spreadsheets in order to visualize them with the modeling software Sybyl.

3.9. VALIDATION OF ELECTROSTATIC POTENTIAL CALCULATIONS

The program was validated partly using one of the worst possible scenarios. The electrostatic potential was calculated at the atomic centra (*i.e.*, including charged atoms for which the potential is, in theory, infinite).

In one part of the validation, the electrostatic potential was calculated in all atom positions for several rotated structures of the same protein, β-Purothionin (PDB code 1BHP). A grid spacing of 0.7 Å was used. The same analysis was done using the extra focusing step on all charged atoms. The average surface potential was calculated for each rotated structure and for different pH values.

The electrostatic potentials calculated by SCARP were then compared to results from the commercial software DelPhi. This was done both for atom center potentials of the rotated structures and for a set of 2000 random points on the original structure.

3.10. CORRELATION OF ELECTROSTATIC POTENTIALS AND EXPERIMENTAL ION EXCHANGE DATA

Retention data (curves of retention time against pH, obtained from experiments with salt gradient) were obtained from earlier published results (Kopaciewicz *et al.*, 1983). Data from three of the fourteen studied proteins (bacterial α-amylase, almond β-glucosidase and equine immunoglobulin G) were discarded. For these proteins, the information available was not sufficient to find a reliable structure or model representing the protein. The remaining eleven proteins are listed in Table 1, together with information on their structures.

Table 1. Structures used in correlation of between electrostatic potentials and experimental ion exchange data.

Protein	**MW (kDa)**[*)]	**PDB code**	**Source**	**Structure determination method**
Cytochrome C	12	1AKK	Equine	NMR
	12	1HRC	Equine	X-ray
Ribonuclease A	13	7RSA	Bovine	X-ray
Lysozyme	13	1AKI	Hen	X-ray
α-Chymotrypsinogen A	25	1CGI	Bovine	X-ray
Serum Albumin	69	hom[**)]	Bovine	
Ovalbumin	44	1OVA	Hen	X-ray
Ovotransferrin (conalbumin)	77	1OVT	Hen	X-ray
Trypsin Inhibitor	20	1AVU	Soybean	X-ray
β-Lactoglobulin A	35	1BEB	Bovine	X-ray
	35	1CJ5	Bovine	NMR (5 structures)
Myoglobin	18	1DWR	Equine	X-ray
Carbonic anhydrase	30	hom[**)]	Bovine	X-ray

[*)] molecular weights according to Kopaciewicz *et al.*, 1983.
[**)] these models were made using homology modeling, amino acid sequences obtained from Expasy (Expasy, 2003) and protein structures from related species, obtained from the Protein Data Bank (PDB, 2003).

Net charge and average surface potential were calculated for all eleven proteins. Since the surface is defined as all grid points lying within a certain distance range from the protein van der Waals surface, this distance range needs to be optimized. Four distance ranges (0-0.7 Å, 1.5-2.5 Å, 1.5-6 Å, 1.5-10 Å) were tested. For each distance range and for seven of the proteins, the average surface potential (surfPhi) was plotted against the retention time. Four of the proteins (carbonic anhydrase, conalbumin, myoglobin and β-lactoglobulin) were used for validation of the selected shell and not included in these plots. Data points with retention times of less than 2 minutes were omitted, as these points most likely represent unretained proteins.

In Appendix 4, some other descriptors that have been investigated are listed. In Appendix 5, attempts have been made to find a model that predicts whether or not a protein will be retarded in the chromatographic column.

4. Results and discussion

4.1. TIME REQUIREMENTS

The time for calculating the electrostatic potential of a protein varied from about two minutes for a small protein (~5 kDa) to about twenty minutes for a big protein (~80 kDa) using the input values given in Appendix 2.

4.2. ROTATED STRUCTURES

Figure 5 shows calculated values of the electrostatic potential at the atomic centra of the rotated structures plotted against the values corresponding to the original structure. The results imply that points with high absolute value of the electrostatic potential are very sensitive to the exact grid location. Points with lower potentials seem to be almost unaffected when rotating the grid. These values are plotted at the atom positions, which are sometimes the exact locations of charges. Since the potential peaks at charged points, this set is likely to show a higher spreading than would another set of points (for instance, the grid points).

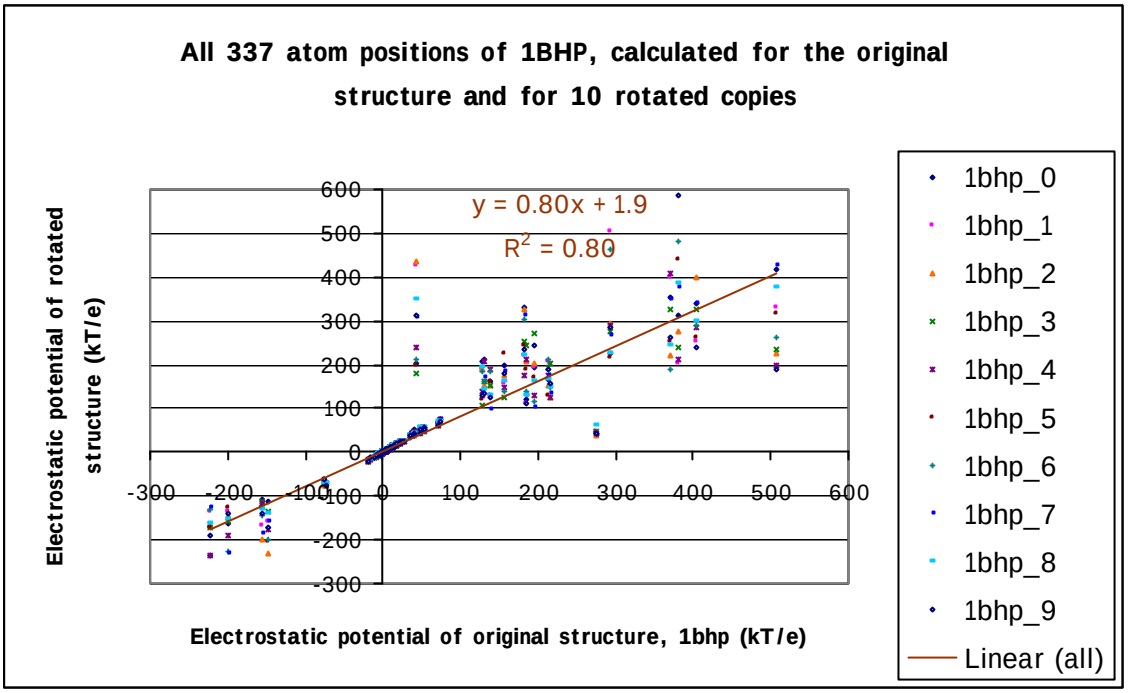

Figure 5. Electrostatic potentials in the atomic centra of β-Purothionin, calculated for different rotated copies of the same structure.

The corresponding results using the extra focusing step are shown in Figure 6. The figure shows that the correlation improves (R^2 is higher and the slope closer to 1) despite the inappropriate accessibility assignment to some points mentioned above. Also, the absolute values of the calculated potential at the charged atom positions are about one order of magnitude larger. This is to be considered more accurate since the corresponding theoretical values are infinite.

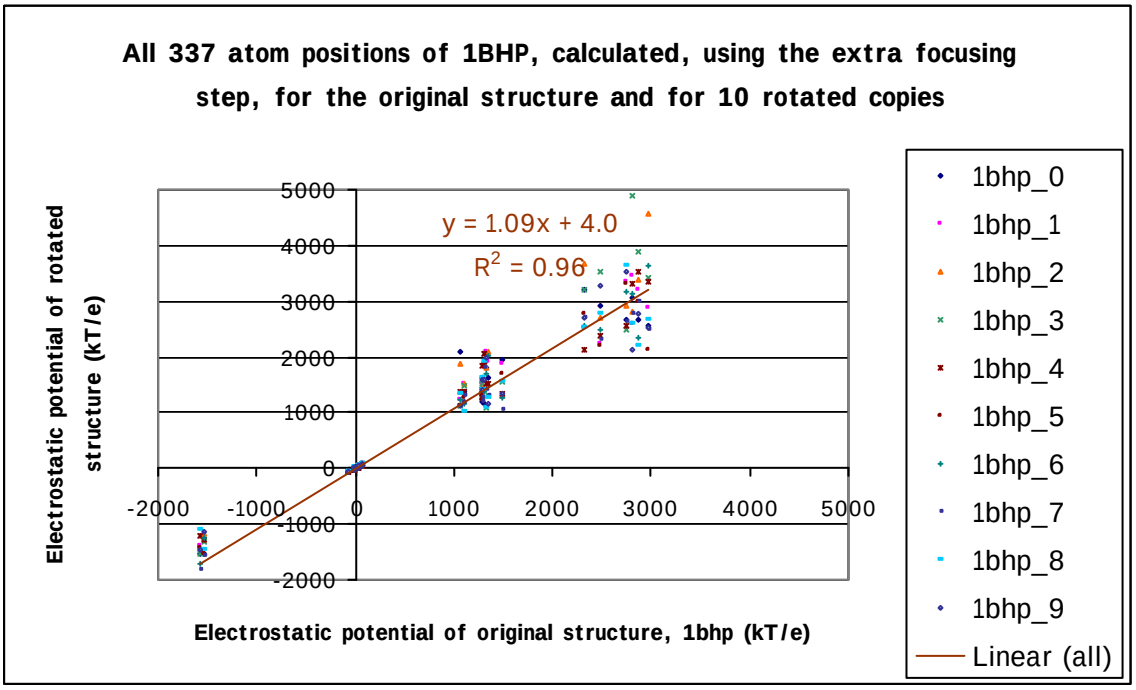

Figure 6. Electrostatic potentials in the atomic centra of β-Purothionin, calculated for different rotated copies of the same structure. An extra focusing step has been performed at each charged atom.

Average surface potentials for each of the rotated structures are listed in Table 2. These results give an idea of the magnitude of the discretization error in calculations of the average surface potential.

Table 2. Average surface potentials for the 11 differently oriented copies of β-Purothionin. The row titled "average" contains the average of the average surface potentials.

		pH6	pH7	pH8	pH9	pH10	pH11	pH12	pH13
net charge (e)		9.02	8.90	8.43	7.45	4.47	1.27	-0.93	-2.63
average surface potential (kT/e)	1bhp	0.87	0.85	0.77	0.63	0.25	0.09	-0.06	-0.20
	1bhp0	1.01	0.98	0.91	0.74	0.32	0.09	-0.08	-0.24
	1bhp1	0.86	0.85	0.77	0.63	0.26	0.06	-0.10	-0.23
	1bhp2	0.92	0.91	0.85	0.72	0.35	0.09	-0.06	-0.20
	1bhp3	0.89	0.88	0.82	0.67	0.29	0.07	-0.08	-0.23
	1bhp4	0.88	0.85	0.79	0.65	0.26	0.05	-0.10	-0.23
	1bhp5	0.89	0.88	0.81	0.68	0.28	0.05	-0.11	-0.25
	1bhp6	0.82	0.81	0.75	0.63	0.30	0.08	-0.06	-0.19
	1bhp7	0.85	0.84	0.78	0.65	0.28	0.08	-0.07	-0.21
	1bhp8	1.09	1.07	1.00	0.86	0.46	0.14	0.00	-0.12
	1bhp9	0.86	0.85	0.78	0.65	0.27	0.06	-0.09	-0.23
average		**0.91**	**0.89**	**0.83**	**0.69**	**0.31**	**0.08**	**-0.08**	**-0.21**
standard deviation		*0.08*	*0.08*	*0.08*	*0.07*	*0.06*	*0.03*	*0.03*	*0.04*

4.3. COMPARISON WITH DELPHI

The plot corresponding to Figure 5 but calculated using DelPhi is shown in Figure 7. This plot again illustrates the discretization error, which is most noticeable at points of higher potential. Thus, it seems that the large spread at high potentials is a feature of the method itself.

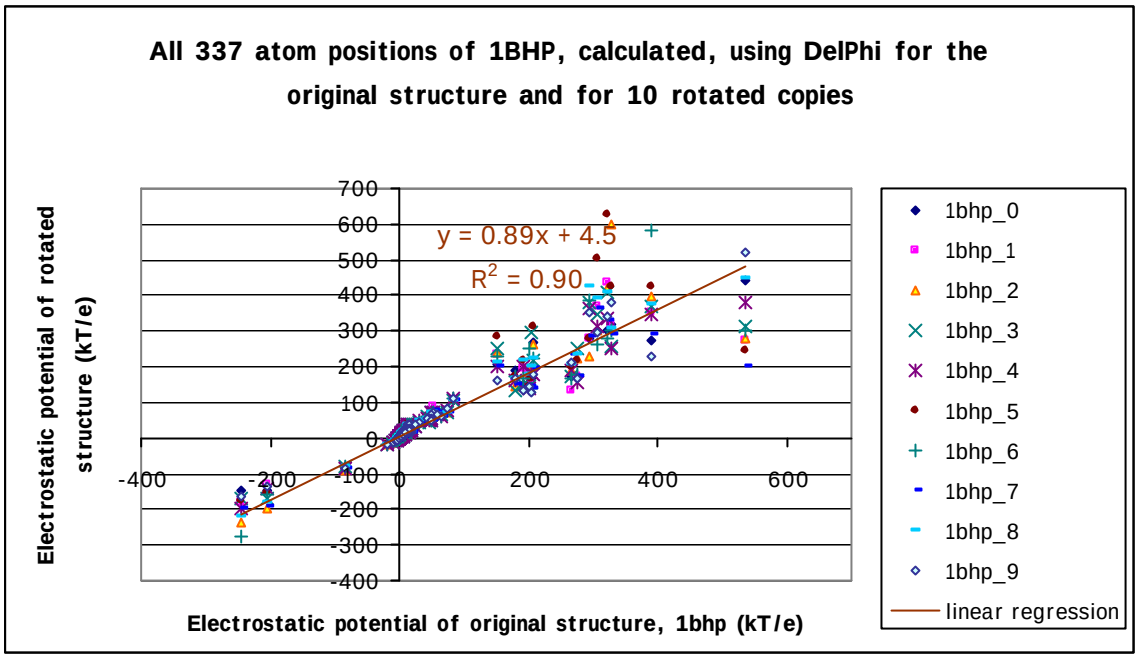

Figure 7. Electrostatic potentials in the atomic centra of β-Purothionin, calculated for different roteted copies of the same structure using DelPhi.

For each of the rotated structures as well as for the original structure, the value calculated by Delphi was plotted against the value calculated by SCARP. The plot for the original structure is shown in Figure 8. The other plots showed similar behavior. Linear regression of these plots gave R^2-values ranging between 0.85 and 0.98 and slopes ranging between 0.92 and 1.46. This plot also illustrates the lower accuracy of values calculated at points of high potentials.

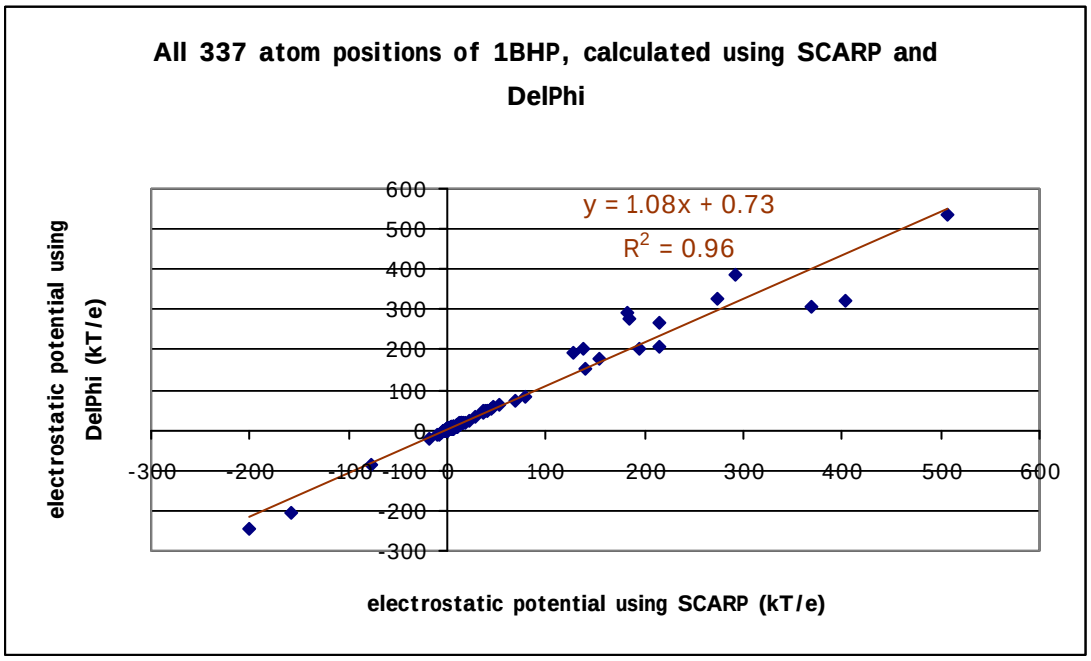

Figure 8. Electrostatic potentials in the atomic centra of β-Purothionin, the results using DelPhi against the results using SCARP.

Since the potentials in the atomic centra are extremely sensitive to the exact grid location, a more relevant comparison between DelPhi and SCARP was carried out using a set of 2000 randomly chosen points within the grid (not necessarily coinciding with grid points). This plot (Figure 9) shows that the results from SCARP are highly correlated with the results from DelPhi. Slight variations are obtained for points at the boundary and inside the protein.

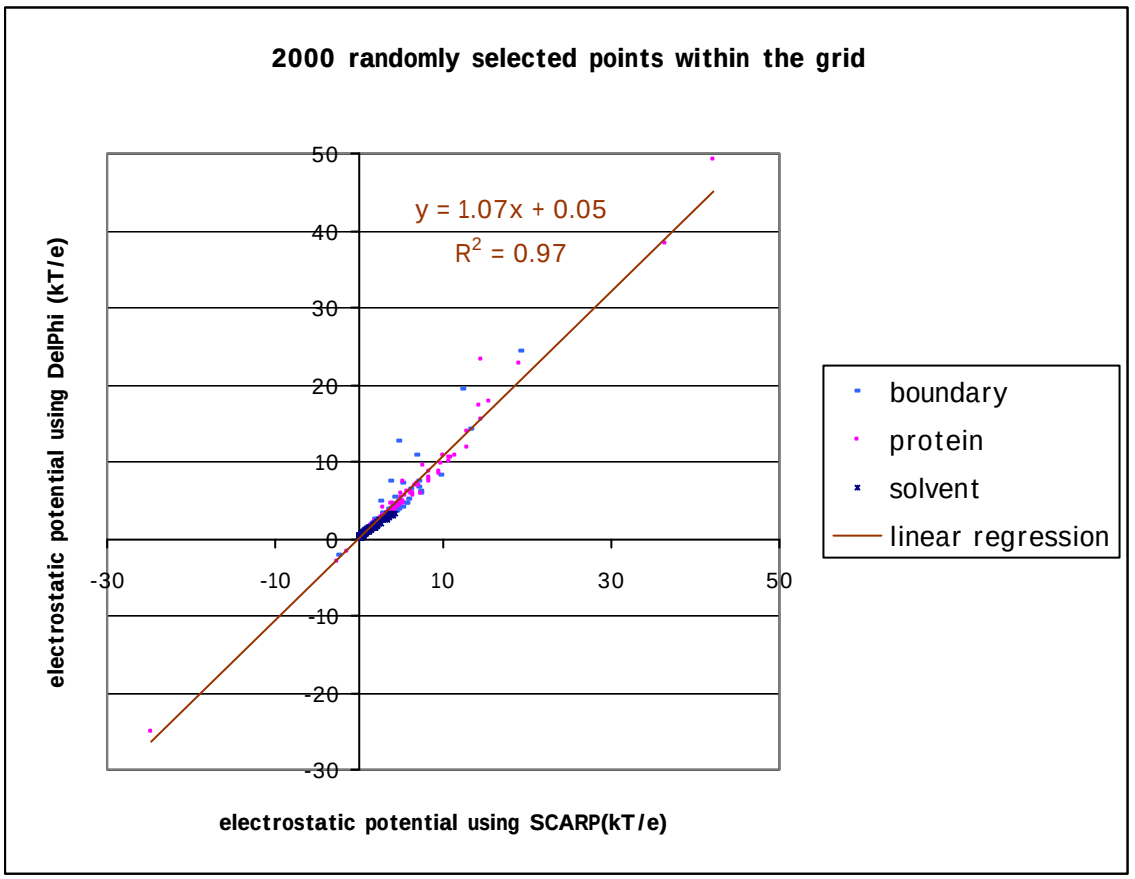

Figure 9. Electrostatic potentials in 2000 randomly selected points inside and outside a protein molecule (β-Purothionin). The results using DelPhi are plotted against the results obtained with SCARP. The boundary points are points whose nearest eight neighbor grid points do not lie entirely inside or outside the protein.

4.4. Correlation with experimental cation exchange data

As explained in section 3.10, for several shell sizes, the correlation with the retention times was studied. By inspection of these results (data not shown), the shell size 1.5-6 Å was chosen as being the optimal one among the four studied. This suggests that the potential in the vicinity of the protein also contributes to the interaction with the stationary column particles. Moreover, this suggests that the potential at a very small distance (less than ~1.5 Å) from the molecular surface of the protein does not affect the retention. A possible interpretation of this observation is that the charged groups in the stationary column particles cannot access the molecular surface of the protein.

Figure 10 shows plots of Q, Q/MW and surfPhi against retention time for cation exchange. The different data point within one protein correspond to different pH values.

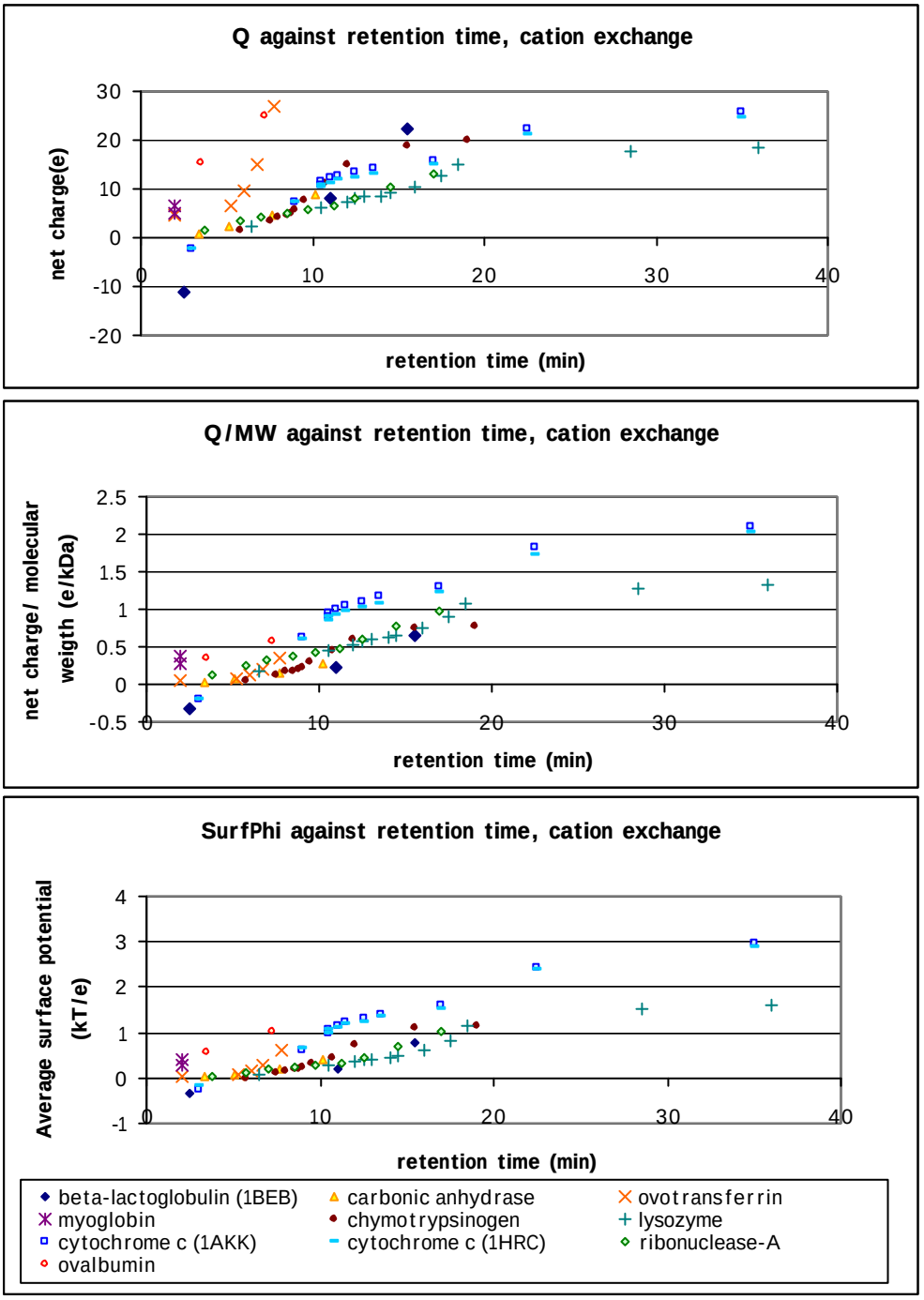

Figure 10. Protein net charge (Q), protein net charge divided by the molecular weight (Q/MW) and average surface electrostatic potential (surfPhi) plotted against experimentally determined cation exchange retention times for several pH values.

The plots of net charge (Q) and average surface potential (surfPhi) in Figure 10 are similar to the previously published results (Haggerty & Lenhoff, 1991). The differences can be explained by the fact that different protein structures were used in the published report and in this work. In both reports, retention data are obtained from the same paper (Kopaciewicz *et al.*, 1983) in which the source of each protein is specified. For some proteins, Haggerty and Lenhoff used the structure of a different

species than that from which the retention data were obtained. For some of the proteins, the information on the proteins used in the chromatographic experiments is not sufficient to decide which protein structure should be used in the calculations. Table 3 shows a compilation of the proteins for which the structures used in this report differ from the ones used in the report of Haggerty and Lenhoff.

Table 3 A comparison of protein structure used in different reports

Protein	Source species of retention data[*)]	Source species of previous report[**)]	Source species in this report
myoglobin	equine	sperm whale	equine
serum albumin	bovine	not used	bovine (homology model)
carbonic anhydrase	bovine	human	bovine (homology model)
α-amylase	bacterial	A. oryzae	not used[***)]
immunoglobulin G	equine	human	not used[****)]
β-lactoglobulin	bovine	not used	bovine
ovotransferin	hen	not used	hen
cytochrome c	equine	not used	equine
ovalbumin	hen	not used	hen
trypsin inhibitor	soybean	not used	soybean
chymotrypsinogen	bovine	bovine	bovine
ribonuclease A	bovine	bovine	bovine
lysozyme	hen	hen	hen

[*)] Kopaciewicz *et al*., 1983
[**)] Haggerty & Lenhoff, 1991
[***)] The source specification (bacterial) is insufficient to identify an unambiguous protein sequence.
[****)] Immunoglobulin G contains variable parts that are not specified for the retention data.

Comparing only the data corresponding to the three cases where the same structures were used in both studies reveals an excellent agreement with the previously published results (Figure 2 in Haggerty & Lenhoff, 1991). This is shown in Figure 11 and proves that it has been possible to reproduce these results for cation exchange data. This also proves the strong correlation between the experimentally obtained retention times and the surface potential for this data set. The plot for all the proteins (Figure 10) shows that the correlation is somewhat reduced as reflected by a wider spread in the data points. It appears that the curves of surface potential for individual proteins have different intercepts, although having very similar slopes. This difference may be due to chromatographic properties not related to ionic interactions. In the plot of Haggerty and Lenhoff it is apparent that the curves have been modified with respect to the intercept, since null retention times are present in the plot. In this work the unmodified retention values (Kopaciewicz *et al.*, 1983) have been used.

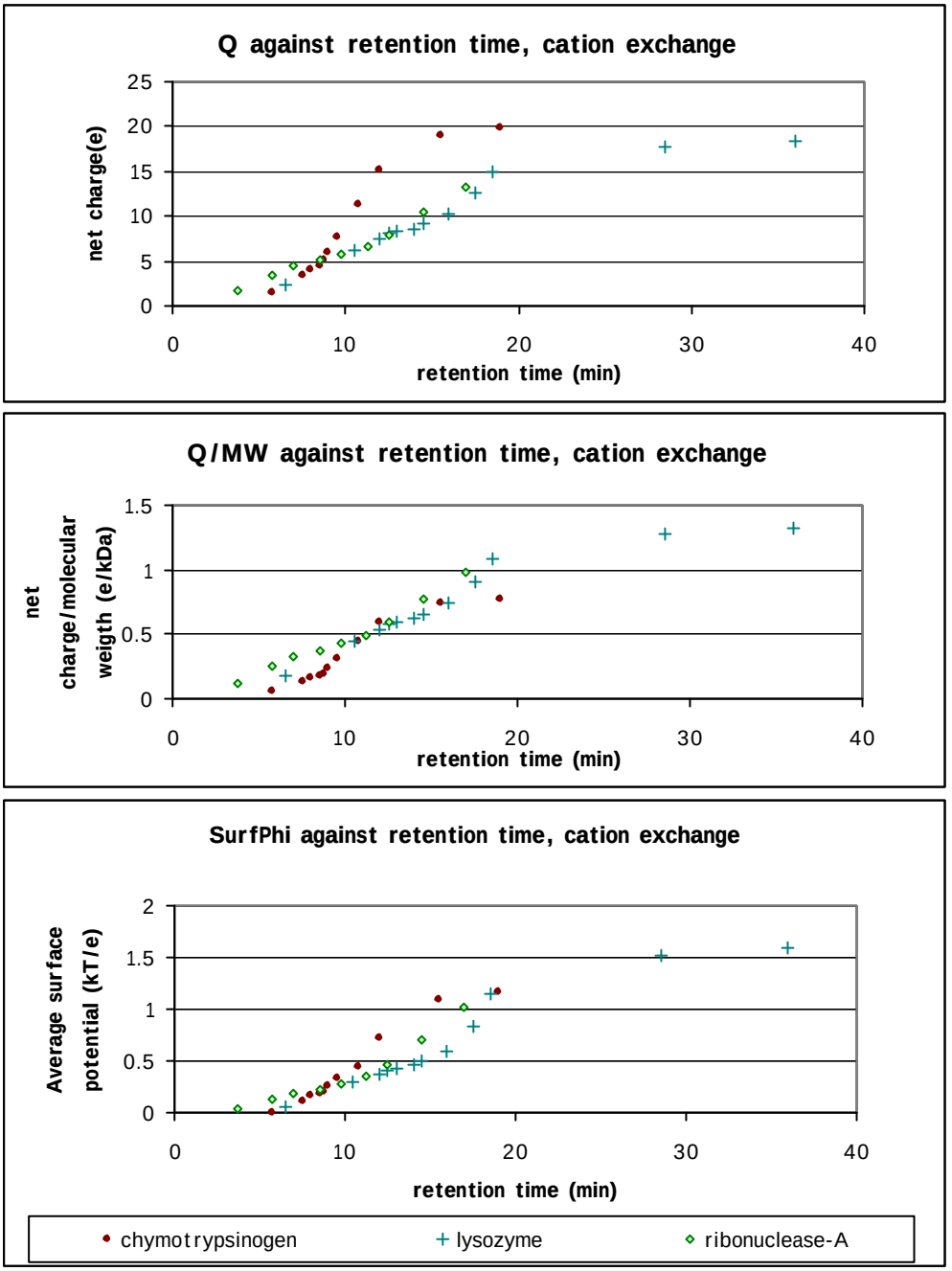

Figure 11. Protein net charge (Q), protein net charge divided by the molecular weight (Q/MW) and average surface electrostatic potential (surfPhi) plotted against experimentally determined cation exchange retention times for the proteins previously studied (Haggerty and Lenhoff, 1991).

These plots show that SurfPhi, which is a property calculated from first principles provides a better description of cation exchange behavior than does the net charge. This is in agreement with previous results (Haggerty & Lenhoff, 1991). Interestingly, a simple normalization of the net charge by the molecular weight (Q/MW) produces plots that are very similar to those of the average surface potential (surPhi). This suggests that the net charge normalized by the molecular weight could be used as an easy-to-obtain descriptor, without the need for a protein 3D-structure, and that is almost as valuable as the average potential for modeling cation exchange. Both these descriptors have in common that they represent an electrostatic property of the protein related to its size. Thus, it is apparent that the size of the protein should be taken into consideration.

4.5. CORRELATION WITH EXPERIMENTAL ANION EXCHANGE DATA

Figure 12 shows plots of Q, Q/MW and surfPhi against retention time for anion exchange chromatography. To our knowledge, such plots have not been reported previously. Even in this case, there is a relation between the calculated property and the experimentally obtained values.

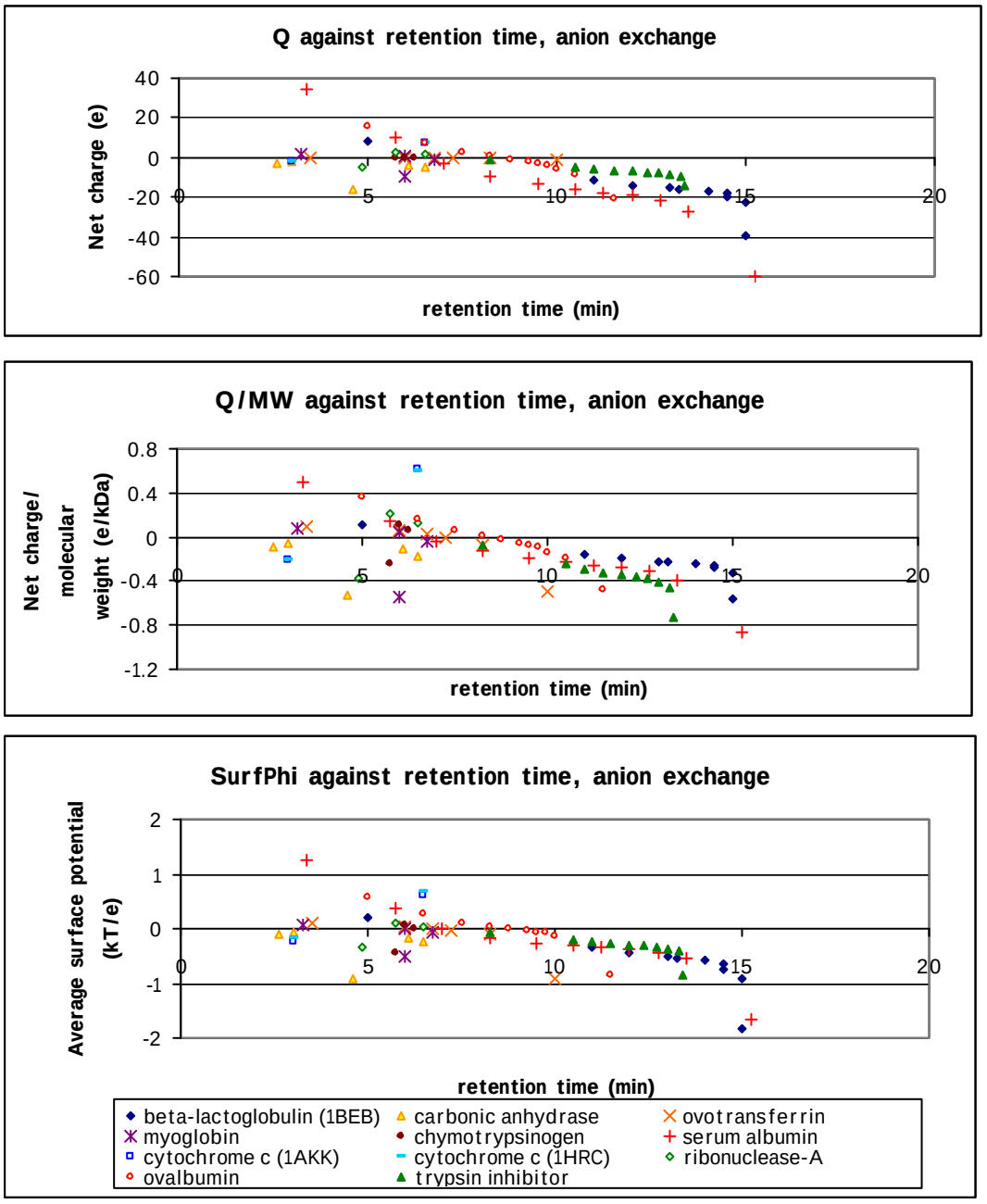

Figure 12. Protein net charge (Q), protein net charge divided by the molecular weight (Q/MW) and average surface electrostatic potential (surfPhi) plotted against experimentally determined anion exchange retention times for different pH values.

However, for some of the proteins (chymotrypsinogen, ribonuclease, carbonic anhydrase, myoglobin and cytochrome c) the correlation is poor. None of these proteins are retained in a significant part of the studied pH range and they have in common that shorter retention times are obtained when increasing pH above a certain

value (about 9). This effect is not captured in any of the descriptors studied. As has been suggested (Kopaciewicz *et al.*, 1983) the observed behavior could be due to protein structure modifications such as denaturation and aggregation.

Excluding these proteins from the plots in Figure 12 gives a better correlation (Figure 13). In this case the net charge plot is rather linear and there is not much improvement in the Q/MW plot. On the other hand, the curves for the average surface potential lie closer together.

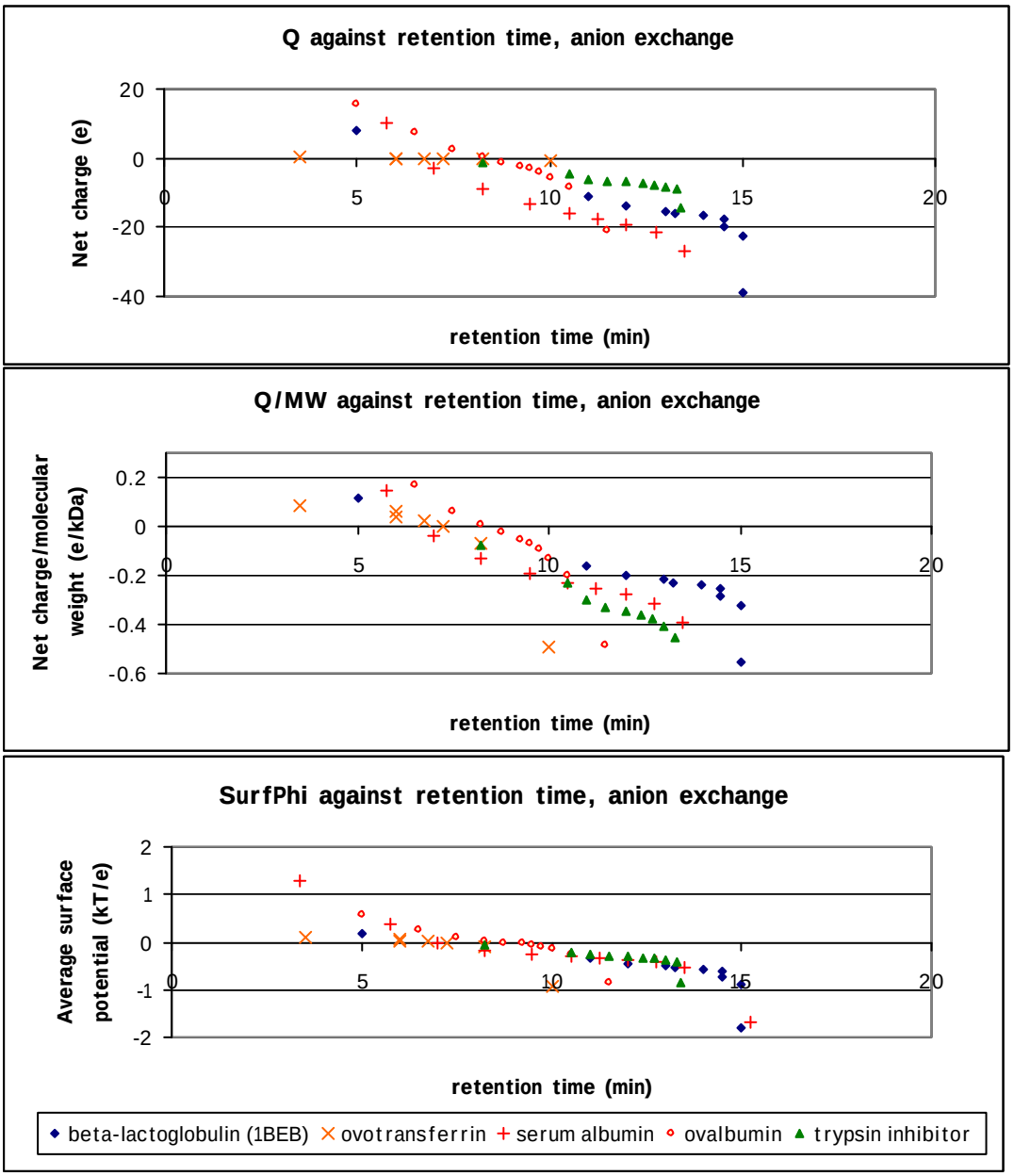

Figure 13. Protein net charge (Q), protein net charge divided by the molecular weight (Q/MW) and average surface electrostatic potential (surfPhi) plotted against experimentally determined anion exchange retention times. Outliers are omitted.

4.6. COMPARISON OF ANION AND CATION EXCHANGE RESULTS

The absolute value of the slopes of the curves in the two plots of surfPhi (bottom of Figures 10 and 12) are of the same order of magnitude (about 0.1 kT/(e min)). The

two plots differ in that, for cation exchange, there are data points available for retention times of about 30 min, whereas for anion exchange, no proteins are retained for more than about 15 min. This difference could be due to a difference in the two types of adsorbents, probably related to their pH stability ranges and the pH ranges used in the experiments.

Some proteins with a descriptor value over zero (for anion exchange) or below zero (for cation exchange) are retained in the column, which implies that a protein can be retained in a column of the same charge. This behavior seems to be more common in the anion exchange data set than in the cation exchange data set. However, the data sets may be too small to decide whether this reflects a fundamental difference between the two types of ion exchange.

Studying one protein at a time suggests that the relation between Q/MW or surfPhi and the retention time is not exactly linear. Instead, the curves appear to be somewhat sigmoid-shaped.

In this work, the same salt concentration (10mM) has been used in all the calculations that have been used in comparisons with experimental retention times. This might seem inappropriate, since the retention data are obtained from experiments using a linear salt gradient. However, calculations using different salt concentrations (0-140 mM, data not shown) indicated that the choice of salt concentration does not affect the results significantly, provided that it is set within reasonable limits as compared to the experimental conditions.

4.7. CAN THE SURFACE POTENTIAL BE USED TO PREDICT RETENTION TIMES?

The results show that the surface potential is definitely a relevant descriptor in ion interaction processes since the plots for diverse proteins are very similar in shape and quite linear. This is a satisfactory result since this property is calculated from information on the protein structure alone, using elementary electrostatic theory. The number of variable parameters has been kept to a minimum, avoiding the risks of overfitting.

There are, however, two main difficulties in attempting to use this property directly as a predictor for retention time. One is that the intercepts of the plots seem to be protein dependent. Thus, a possible further development of this work would be to develop a model for this difference based on other properties than electrostatic ones. For instance, the difference could possibly be explained partly by the time it takes for the protein to pass through the column in a pH or salt range where the electrostatic interactions are very small. This time is not necessarily the same for all proteins and might depend on, for instance, the size or the shape of the molecule. The second difficulty arises from the fact that a few proteins show a behavior that is incompatible with the correlation described. This was illustrated for several proteins in the anion exchange data set considered. For these proteins, the predictions will in any case not work and the only way to identify them seems to be from the experimental data itself. In Appendix 5, an attempt is made to identify these outliers computationally by means of electrostatic descriptors.

5. Conclusions

This work has shown that it is possible to reproduce the previously published results on the relation between average surface potentials and cation exchange retention times for a certain set of proteins. It has also been possible to extend this set and still see the same pattern, *i.e.*, an approximately linear correlation between the average surface potential and the retention time in cation exchange chromatography. However, another descriptor, the protein net charge divided by the molecular weight, seems to contain almost the same information as does the average surface potential.

Moreover, it has been shown that the same principles apply, to some extent, to anion exchange modeling. However, some proteins have been shown not to follow this pattern in anion exchange chromatography.

Thus, the average surface potential has been shown to be a highly relevant descriptor for ion exchange phenomena, although it does not account for all differences among various proteins. The remaining differences most likely depend on other factors than electrostatics and it might be possible to model them in some other way.

It has also been shown that, for the cation exchange data set studied studied, the net charge divided by the molecular weight provides almost the same information as does the average surface potential.

6. Acknowledgements

I thank all the people at Amersham Biosciences that have, in some way, contributed to this work. In particular, I would like to thank my supervisor, Enrique Carredano, for his support. I would also like to thank Jinyu Zou, who has contributed to this work in several ways, and Maria Strömgren, for providing me with useful source code for this program.

Furthermore, I would like to thank Gerard Kleywegt at the Dept. of Cell and Molecular Biology at Uppsala University. Besides being scientific reviewer of this work, he has also provided me with rotated protein structures for validation of the program.

Finally, I thank my family. Without their help and support, this work would not have been possible.

7. References

Amersham Pharmacia Biotech (1999). Ion Exchange Chromatography. Principles and Methods.

ExPASy, Expert Protein Analysis System molecular biology server: http://www.expasy.org/ (April 2003).

Fogolari, F., Zuccato, P., Esposito, G.,Viglino, P.(1999) Biomolecular electrostatics with the linearized Poisson-Bolzmann equation. *Biophys. J.*, **76**, 1-16.

Gilson, M. K., Introduction to continuum electrostatics, with molecular applications. August 25, 2000. http://gilsonlab.umbi.umd.edu/ce_www1a.pdf (February 14, 2002).

Haggerty, L.,Lenhoff, A. M.(1991) Relation of Protein Electrostatic Computations to Ion-Exchange and Electrophoretic Behaviour. *J. Phys. Chem.*, **95**, 1472-1477.

Heath, M.,H. Scientific Computing: An Introductory Survey, McGraw-Hill Book Co (1997).

Honig, B., Nicholls, A. (1995) Classical Electrostatics in Biology and Chemistry. *Science*, **268**, 1144-1149.

Klapper, I., Hagstrom, R., Fine, R., Sharp, K.,Honig, B. (1986) Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino-acid modification. *Proteins*, **1**, 47-59.

Kopaciewicz, W., Rounds, M. A., Fausnaugh, J., Regnier, F. E. (1983) Retention model for high-performance ion-exchange chromatography. *J. Chromatogr.*, **266**, 3-21.

Ljunglöf, A. (2002).Direct Observation of Biomolecule Adsorption and Spatial Distribution of Functional Groups in Chromatographic Adsorbent Particles. Ph D thesis, Uppsala University.

Nicholls, A., Honig, B. (1990) A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Bolzmann equation. *J. Comp. Chem.*, **12**, 435-445.

Nordling, C., Österman., J. Physics Handbook for Science and Engineering, 5th ed., Studentlitteratur (1996).

PDB, the Protein Data Bank: http://www.pdb.org/ (Jan-May 2003).

Scopes, R. K., Protein Purification, 3rd ed., Springer-Verlag (1994).

Song, M., Breneman, C. M., Bi, J., Sukumar, N., Bennet, K. P., Cramer, S., Tugcu, N. (2002) Prediction of Protein Retention Times in Anion-Exchange Chromatography Systems Using Support Vector Regression. *J. Chem. Inf. Comput. Sci*. **42**, 1347-1357.

Ståhlberg, J., Jönsson, B., Horváth, C. (1991) Theory for Electrostatic Interaction Chromatography of Proteins. *Anal. Chem*., **63**, 1867-1874.

Yang, A-S., Gunner, M. R., Sampogna, R., Sharp, K., Honig, B. (1993) On The Calculation of pK_as in Proteins. *Proteins*, **15**, 252-265.

Appendix 1: Overview of the program

The program SCARP is outlined as a flowchart in Figure 14.

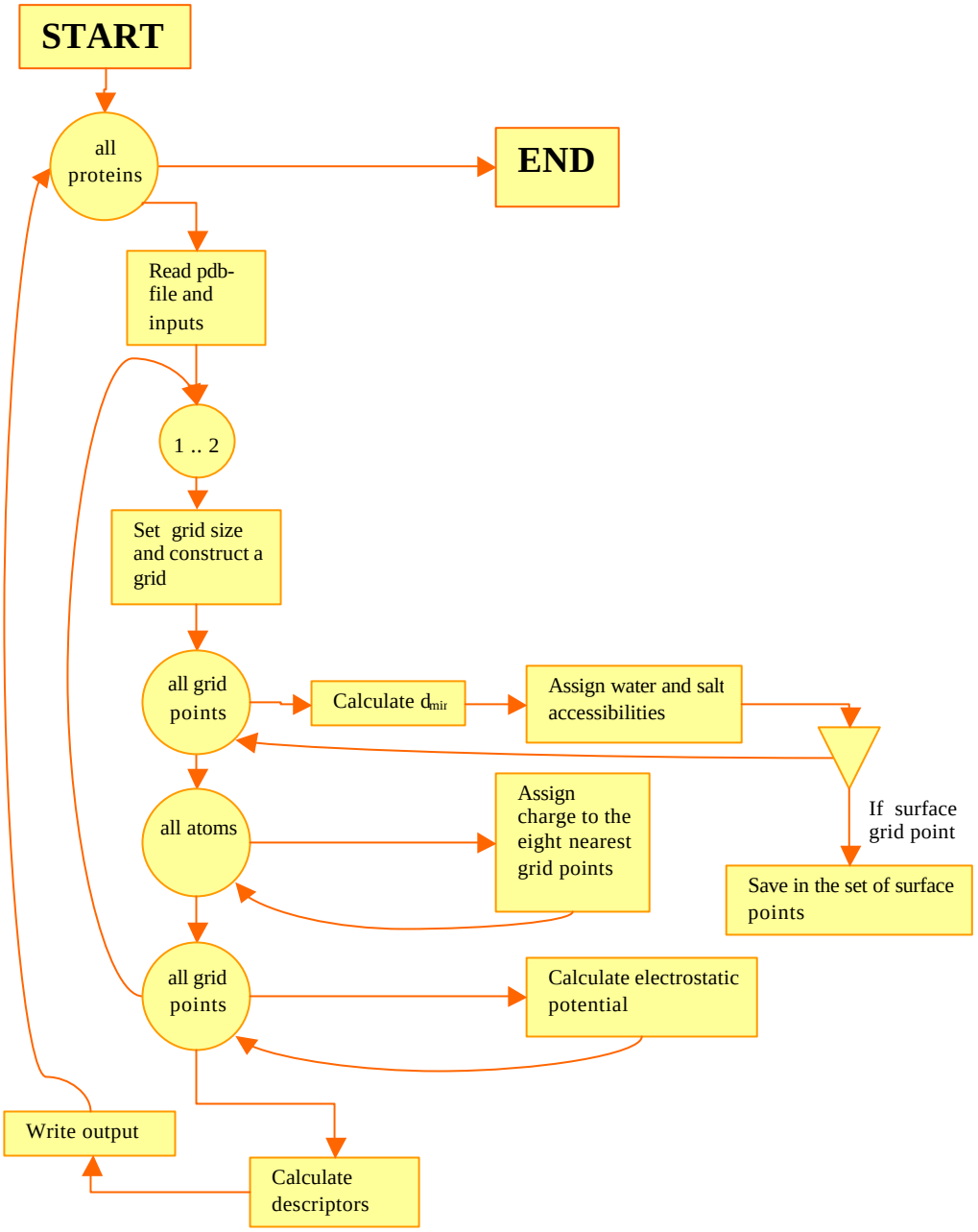

Figure 14. Flowchart of SCARP.

Appendix 2:Variable values used in this work

Tables 3 and 4 show van der Waals radii and pK_a values used in this work. The input grid size was set to 0.7 Å and the focusing parameter f was set to 4. The salt concentration was set to 140 mM in the validation of the electrostatic potential calculations and to 10 mM in the calculations for comparisons with retention data.

Table 3 .van der Waals radii

Entity	**van der Waals radius (Å)**
C	1.90
O	1.75
N	1.65
S	1.90
Fe	0.70
Na, Cl[*)]	1.38
H_2O	1.40

*) Approximate average of Na^+ and Cl^- radii.

Table 4. pK_a values of side chain residues and backbone terminals

residue	**pK_a**
cystein	8.5
aspartate	4.4
glutamate	4.4
histidine	6.5
lysine	10.0
arginine	12.0
tyrosine	10.0
amino terminal	8.0
carboxy terminal	3.1

Appendix 3: Convergence criterion

The electrostatic calculations are carried out using an iterative method. Using this method, the mean square difference between the electrostatic potential in two subsequent iterations decreases with each step. When it gets below a certain threshold (the convergence criterion), the calculations finish.

To find an appropriate convergence criterion, calculations were carried out for a series of different convergence criteria. For comparison, calculations were carried out with an extremely low convergence criterion (10^{-15} kT/e). This calculation was compared to each of the calculations in the series. The correlation as a function of convergence criterion is shown in Figure 15. An inverse proportionality between speed and accuracy was observed. Based on this plot, the convergence criterion 10^{-6} kT/e was chosen as an appropriate trade-off between accuracy and speed.

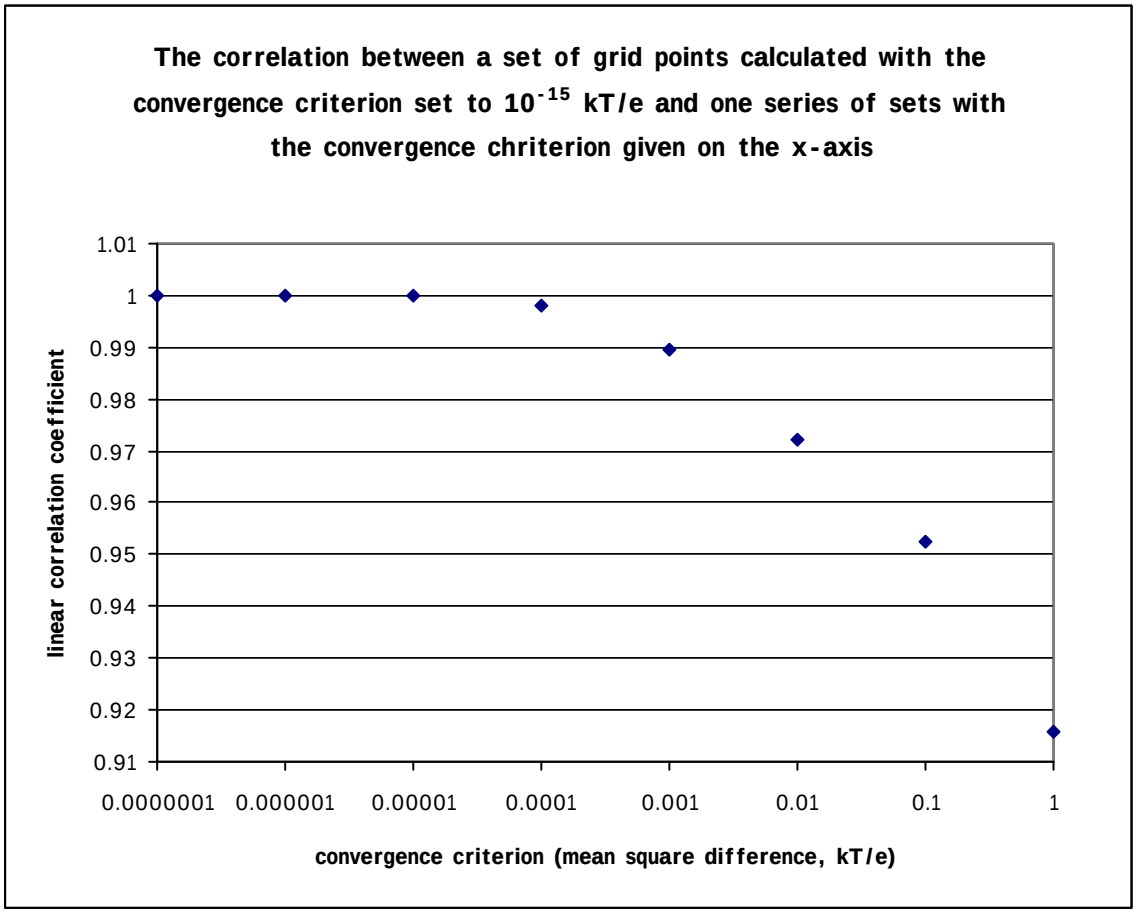

Figure 15. The correlation between a set of low convergence criterion and sets with varied convergence criterions.

Appendix 4: Other descriptors based on the electrostatic potential

The electrostatic potential is a continuos function in space. To be able to compare different proteins, one would like to incorporate the important properties of this three-dimensional property into a single value, a descriptor. In this work, several different descriptors were studied. One of them, the average surface electrostatic potential (surfPhi) is discussed in sections 3.10 and 4.4-4.7. The other descriptors studied were:

- the largest value of the electrostatic potential in any grid point on the protein surface
- the smallest value of the electrostatic potential in any grid point on the protein surface
- Bolzmann-weighted average using a positive probe ($=\Sigma(q\Phi_i e^{(-q\Phi i)})/ \Sigma(\Phi_i q)$)
- Bolzmann-weighted average using a negative probe ($=\Sigma(-q\Phi_i e^{(q\Phi i)})/ \Sigma(-\Phi_i q)$)
- the total number of surface points
- the number of surface points with positive potential
- the number of surface points with negative potential
- the fraction of surface points with positive potential
- the fraction of surface points with negative potential
- the average potential in all surface points with a positive potential
- the average potential in all surface points with a negative potential
- the average potential in all surface points with a positive potential times the number of surface points with positive potential
- the average potential in all surface points with a negative potential times the number of surface points with negative potential
- the average potential in all surface points with a positive potential times the fraction of surface points with positive potential
- the average potential in all surface points with a negative potential times the fraction of surface points with negative potential

These descriptors are closely correlated. Consequently, most of them showed similar behavior when plotted against experimentally determined retention times. Since the two models of cytochrome c gave very different results of the Bolzmann-weighted averages, these descriptors appeared to be very sensitive to small changes in structures. Most of the descriptors did not seem to be better correlated to the retention time than were Q/MW or surfPhi.

The fraction of positive or negative surface (fracPos or fracNeg), however, seemed to be linearly correlated to the retention time up to about 20 minutes, when it reaches its maximum value of 1. Here, all surface points have the same sign on the electrostatic potential. This descriptor was subjected to the same analysis as was done on Q/MW and surfPhi, which is described earlier in this report (section 3.10).

Figure 16 shows fracPos against cation exchange retention time. Figure 17 shows fracNeg against anion exchange retention time. These plots show some linearity between fracPos/fracNeg and retention time.

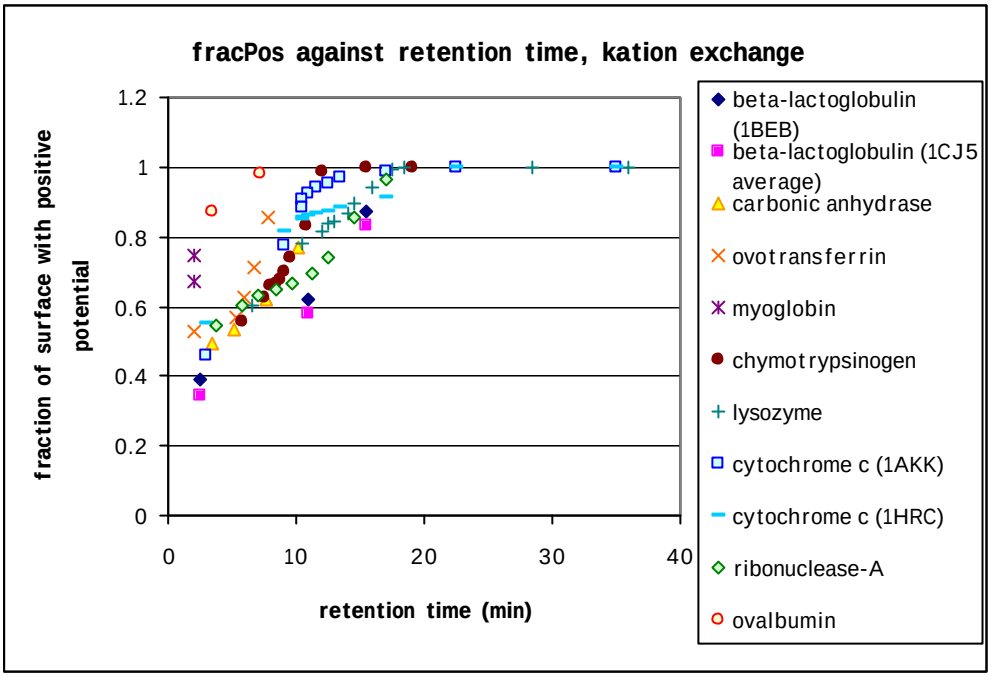

Figure 16. Fraction of positive surface potential (fracPos) plotted against experimentally determined kation exchange retention times.

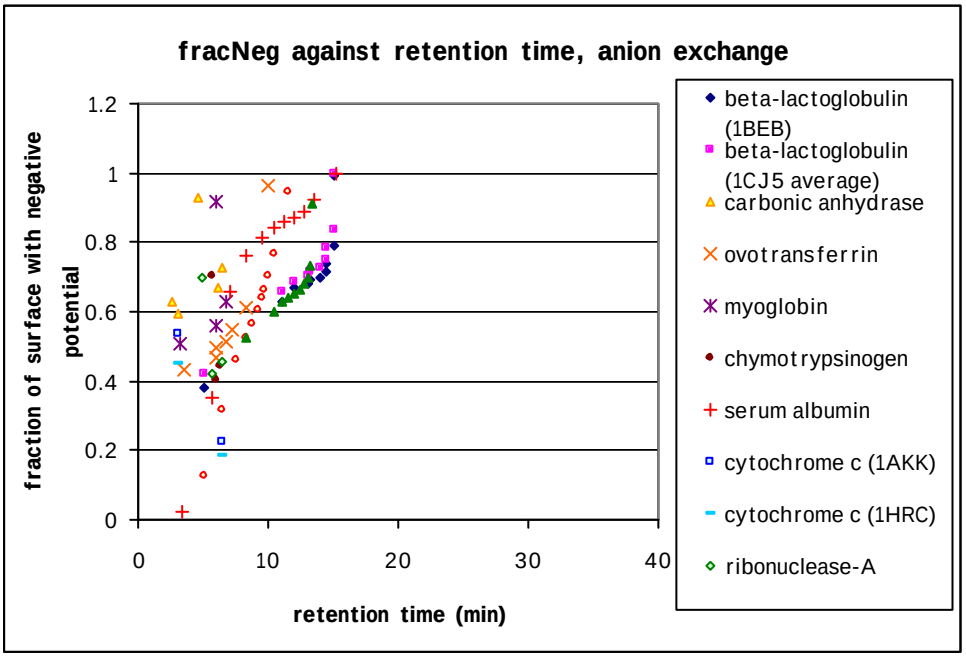

Figure 17. Fraction of negative surface potential (fracNeg) plotted against experimentally determined kation exchange retention times.

Appendix 5: Prediction of whether or not a protein is retained in an ion-exchange column

For anion exchange as well as for cation exchange, the retention data could, by visual inspection, be separated into two sets. The first set consisted of all data points with a retention time of about two minutes. These points were interpreted as not being retained by any electrostatic interactions in the chromatographic column. The other set consisted of all other data points, and was interpreted to contain all proteins retained in the column. These two sets were used as a basis to study the possibility to predict, based on electrostatic descriptors, to which one of the two sets a protein belongs. Three descriptors, Q/MW, surfPhi and fracPos/fracNeg, were studied.

For each descriptor, histograms over the two sets were used to visualize the separation of the two sets. This was done both for anion- and for cation exchange. For anion exchange, a third set was identified from the curves of retention time against pH. This set consisted of all proteins in a pH range where a higher pH gave rise to shorter retention times. This third set was included in the histograms for anion exchange.
In these histograms, only seven of the proteins were used. The remaining four proteins (carbonic anhydrase, conalbumin, myoglobin and β-lactoglobulin) were saved for validation of the models.

Figure 18 shows cation exchange histograms for Q/MW, surfPhi and fracPos, respectively. Figure 19 show anion exchange histograms for Q/MW, surfPhi and fracNeg, respectively.

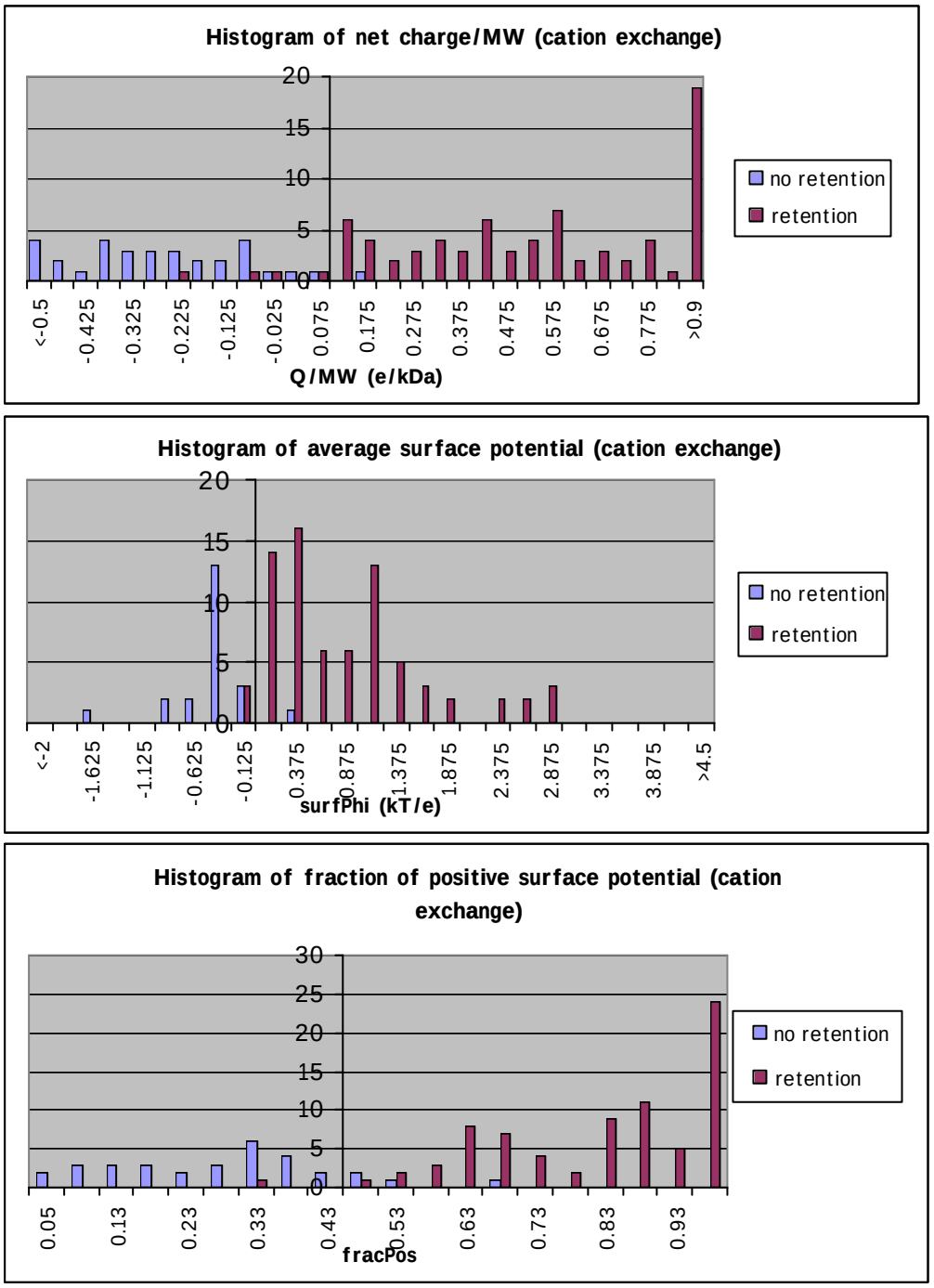

Figure 18. Histogram showing the distribution of Q/MW, surfPhi and fracPos in the test set, cation exchange. The y-axis symbolizes the discrimination rules: proteins with values to the left of the y-axis would be predicted not to be retained and those to the right would be predicted to show retention.

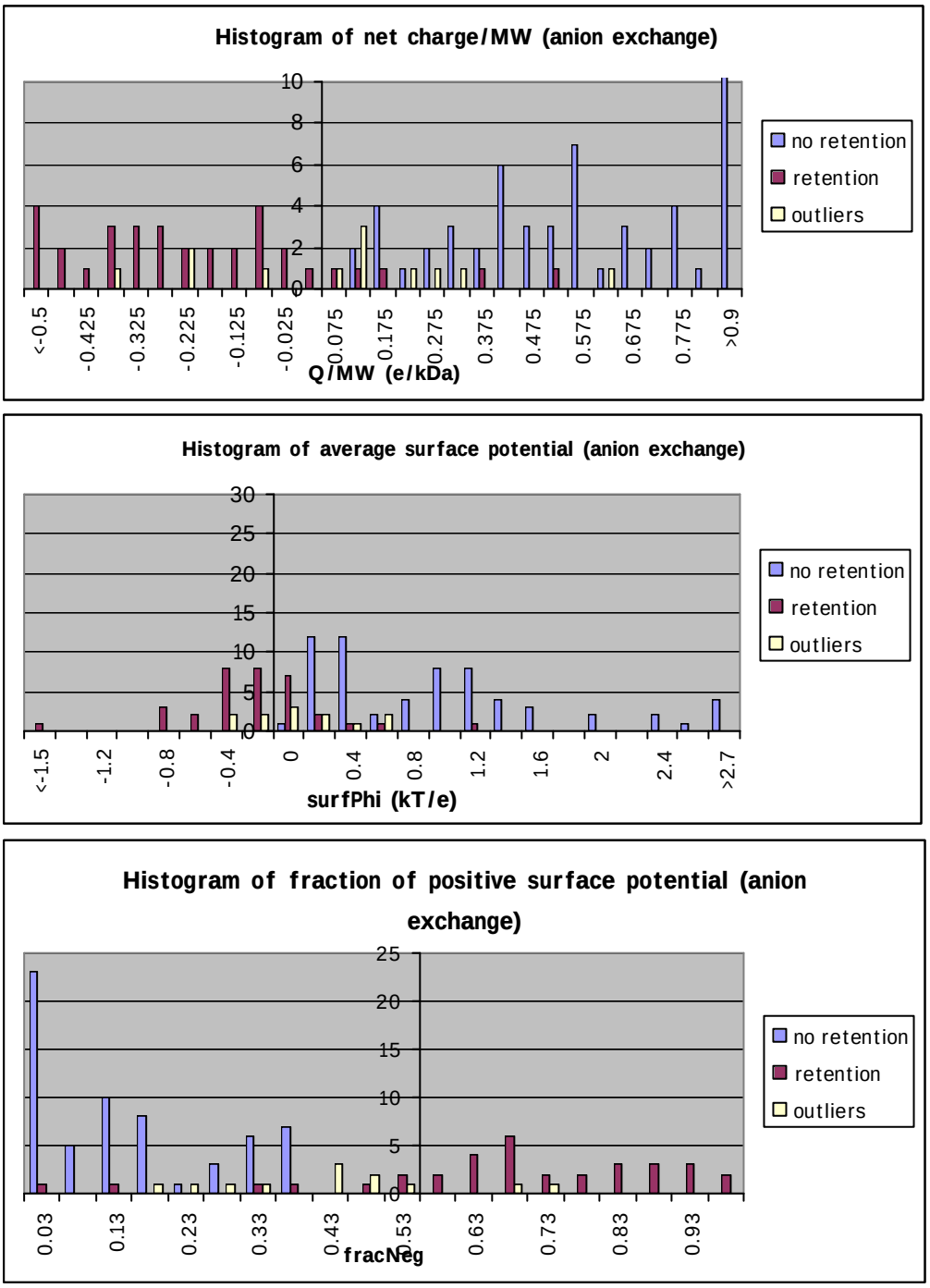

Figure 19. Histogram showing the distribution of Q/MW, surfPhi and fracNeg in the test set, anion exchange. The y-axes symbolizes the discrimination rules: proteins with values to the right (for surfPhi and for Q/MW) or to the left (for fracNeg) of the y-axis would be predicted not to be retained and those to the left (for surfPhi and for Q/MW) or to the right (for fracNeg) would be predicted to show retention.

Based on the histograms, discrimination rules were found for each of the three descriptors. These rules were used to predict whether or not the proteins in the validation set would be retained on each type of column. The predictions are listed in Table 5. The percentage of correct predictions is similar for all three descriptors and

the percentage of false positives is slightly higher for Q/MW. However, the data set is not big enough for any general discrimination rules to be established.

Table 5. The frequency of correct and incorrect predictions

		fracPos	Q/MW	surfPhi
kation exchange	true neg	10	8	8
	true pos	10	13	13
	false pos	1	3	3
	false neg	4	1	1
anion exchange	true neg	15	14	16
	true pos	28	31	25
	false pos	1	4	0
	false neg	8	3	11
	accuracy *)	0.76	0.92	0.76
	reliability **)	0.95	0.86	0.93

*) The accuracy is defined as the number of true positives divided by the number of true positives plus the number of false negatives (*i.e.*, all positives in the original set)

**) The reliability is defined as the number of true positives divided by the number of true positives plus the number of false positives (*i.e.*, all data point predicted as positives)