J. Am. Soc. Mass Spectrom. (2012)
DOI: 10.1007/s13361-012-0449-0

RESEARCH ARTICLE

A Computational Model for Protein Ionization by Electrospray Based on Gas-Phase Basicity

Roberto Marchese,[1] Rita Grandori,[2] Paolo Carloni,[3] Simone Raugei[3,4]

[1]SISSA and INFM-DEMOCRITOS center, Trieste, Italy
[2]Department of Biotechnology and Biosciences, University of Milano-Bicocca, Milan, Italy
[3]Computational Biophysics, German Research School for Simulation Sciences, D-52425 Jülich, Germany and Institute for Advanced Simulation IAS-5, Computational Biomedicine, Forschungszentrum Jülich, D-52425, Jülich, Germany
[4]Fundamental and Computational Sciences Directorate, Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, WA 99352, USA

Abstract
Identifying the key factor(s) governing the overall protein charge is crucial for the interpretation of electrospray-ionization mass spectrometry data. Current hypotheses invoke different principles for folded and unfolded proteins. Here, first we investigate the gas-phase structure and energetics of several proteins of variable size and different folds. The conformer and protomer space of these proteins ions is explored exhaustively by hybrid Monte-Carlo/molecular dynamics calculations, allowing for zwitterionic states. From these calculations, the apparent gas-phase basicity of desolvated protein ions turns out to be the unifying trait dictating protein ionization by electrospray. Next, we develop a simple, general, adjustable-parameter-free model for the potential energy function of proteins. The model is capable to predict with remarkable accuracy the experimental charge of folded proteins and its well-known correlation with the square root of protein mass.

Key words: Electrospray ionization, Protein ionization, Gas-phase basicity, Monte-Carlo sampling, Molecular-dynamics, Simulations, Density functional theory calculations

Introduction

The interpretation of the data from electrospray-ionization mass spectrometry (ESI-MS) greatly benefits from uncovering the role of factors controlling the degree of protein ionization, which lead to the observed charge-state distributions (CSDs) [1, 2]. The current view of protein electrospray invokes different mechanisms to explain the ionization behavior of unfolded proteins, and folded globular proteins under non-denaturing conditions (the former may be either proteins under denaturing conditions or intrinsically disordered proteins). The degree of ionization of unfolded proteins is considered to be controlled by the apparent gas-phase basicity of protein ions (GB_{app}) relative to that of the solvent [3–5]. GB_{app} measures the propensity of the ionizable groups of desolvated protein ions to acquire a proton. It approaches the GB of the solvent for unfolded proteins in their highest observable charge state [4].

For folded proteins, the interpretation is less straightforward. The extent of protein ionization has been interpreted in terms of GB also in this case [6–9]. In particular, the GB_{app} of folded cytochrome *c* has been calculated from the crystallographic structure, accounting for Coulomb repulsions [9]. The derived value for the 11+ ion was little below the value for water, suggesting that the GB model could be extended to folded proteins. However, the most accredited hypothesis considers the charge of the precursor ESI droplet as the key factor determining the extent of protein ionization [10, 11]. In turn, the droplet charge is assumed to be close to

Electronic supplementary material The online version of this article (doi:10.1007/s13361-012-0449-0) contains supplementary material, which is available to authorized users.

Correspondence to: Rita Grandori; *e-mail:* rita.grandori@unimib.it, Simone Raugei; *e-mail:* simone.raugei@pnnl.gov

Published online: 20 September 2012

Received: 29 September 2011
Revised: 13 July 2012
Accepted: 14 July 2012

the Rayleigh limit. In support to this hypothesis, plots of average protein charge as a function of protein mass can be fitted well by the Rayleigh equation using the surface tension of water and a droplet radius equal to that of the globular protein structure [10]. This hypothesis would imply that the charge of the protein depends on solvent surface tension following the Rayleigh equation [12]. Such a dependence could not be reproduced by experiments on either folded or unfolded proteins [1, 2, 12–15], although solvent surface tension certainly plays an important role in the ESI process [1, 2], and has also been suggested to explain some effects of supercharging agents [11, 16].

Here, we use computational methods to investigate the relevance of GB_{app} for folded proteins under electrospray conditions, with no assumption on a role of the charge of the precursor ESI droplet. Our test systems are nine structurally diverse and well characterized proteins, spanning a range of molecular weight from 4.0 to 29.2 kDa and a wide range isoelectric point (i.e., from basic to acidic). We first calculate GB_{app} by introducing a Monte-Carlo/molecular-dynamics (MC/MD) scheme, which takes into account the ionization of basic (Arg, Lys, His, and N-terminus) and acidic (Asp, Glu, and C-terminus) groups. This procedure explicitly considers the influence of protein structure on the intrinsic gas phase basicity of each ionizable group, allowing for a combined exploration of the conformer and protomer space. This method leads to the identification of a set of lowest-energy protomers for each value of protein net charge. In most cases, self-solvation networks lead to maintenance of zwitterionic states. Next, we propose a simple mathematical model based on the results of these atomistic investigations. This model has no adjustable parameters and it reproduces the well-known correlations between the protein net charge, q, and both mass [10] and solvent-exposed surface [17]. Our model hints to GB_{app} as a key factor for the ionization of folded proteins, suggesting that protein ionization depends on intrinsic properties of the protein structure and on the GB of the solvent [4].

Computational Details

Proteins

Nine globular proteins of diverse size and fold were considered (Table 1): ragweed pollen allergen from *Ambrosia trifida* (40 residues, pdb entry 1BBG), bovine pancreatic trypsin inhibitor (56 residues, pdb entry 1UUA), C-terminal domain of the ribosomal protein L7/L12 from *E. coli* (68 residues, pdb entry 1CTF), α-amylase inhibitor tendamistat (74 residues, pdb entry 1OK0), human ubiquitin (76 residues, pdb entry: 1V80), Ribonuclease SA (96 residues, pdb entry 1C54), E60Q mutant of human FK506 binding protein-12 (107 residues, pdb entry 2PPP), hen-egg white lysozyme (129 residues, pdb entry 1LZT), human carbonic anhydrase II (260 residues, pdb entry 2CBA). The 1LZT, 1BBG, 1UUA, and 1OK0 proteins feature 4, 4, 3, and 2 disulphide bridges, respectively. Breaking of these bonds was not considered.

Available evidence suggests that, under mild ESI conditions, protons are mostly exchanged among few sites (i.e., mainly Arg, Lys, His, Glu, and Asp side chain and the N- and C-termini [18]. Thus, in order to keep the sampling problem tractable, only protonation and deprotonation of these residues was considered.

Protomer Space Exploration

Predicting the distribution of protonated sites within a protein is not trivial, since the number of possible protomers can be prohibitively large to be explored exhaustively by any theoretical approach, even for a small protein. To cope with this problem, several Monte-Carlo (MC) protocols have been proposed [19] (and references therein). These schemes suppose that the protein structure does not change with the protonation state and they usually assume that the (average) protein structure in aqueous solutions or the crystallographic structure is a good approximation of the gas-phase structure. This may be generally true for the protein backbone. However, it might not necessarily hold for side chains. To

Table 1. Proteins Studied in this Work: Ragweed Pollen Allergen (1BBG), Bovine Pancreatic Trypsin Inhibitor (1UUA), C-Terminal Domain of the Ribosomal Protein L7/L12 from *E. coli* (1CTF), α-Amylase Inhibitor Tendamistat (1OK0), Human Ubiquitin (1V80), Ribonuclease SA (1C54) E60Q Mutant of Human FK506 Binding Protein-12 (2PPP), Hen-Egg White Lysozyme (1LZT), and Human Carbonic Anhydrase II (2CBA). For Each Protein, the Table Lists the Number of Residues, the Mass (kDa), the Fold, the Number of Basic (Arg, Lys, His, and N-terminus) and Acidic Residues (Asp, Glu, and C-terminus) Considered in the Present Work and the Main Charge State Observed Experimentally for Ions Originated from Water along with the Charge Predicted in the Present Study

Protein	Residues	Mass	Fold	Basic sites	Acidic sites	Observed charge	Predicted charge
1BBG	40	4.3	α/β	6	5	5[a]	5
1UUA	56	6.3	α/β	10	5	6[b]	6
1CTF	68	6.9	α/β	12	14	6[c]	7
1OK0	74	8.0	Mainly β	7	9	7[d]	8
1V80	76	8.6	α/β	13	12	6[e]–7[f]	7
1C54	96	10.6	α/β	8	13	8[g]	8
2PPP	107	11.8	α/β	13	12	8[h]	9
1LZT	129	14.3	Mainly α	19	10	9[i]	8
2CBA	260	29.0	α/β	32	22	11[j]	12

[a]Ref. [43]; [b]Ref. [44]; [c]Ref. [45]; [d]Ref. [46]; [e]Ref. [47]; [f]Ref. [48]; [g]Ref. [49]; [h]Ref. [50]; [i]Ref. [51]; [j]Ref. [17]

tackle this issue, the present study employs a combined MC/MD sampling scheme using the OPLS/AA force field with GB corrections. Indeed, standard force fields for biomolecular simulations are unable to describe bond breaking and forming. This poses a serious limitation to the exploration of the protomer space using molecular-mechanics schemes. Here we propose to augment the standard force field energies, E_{FF}, with additional energy terms associated to the GB of ionizable residues, introducing the following corrected potential function:

$$E_{\mathrm{corr}} = -\sum_{i}{}^{\prime} (\mathrm{GB})_i \, \delta_i + E_{\mathrm{FF}}$$

where the summation runs over all of the ionizable residues and δ_i is 1 if the i-th residue is ionized and 0 otherwise.

We chose the OPLS/AA force field [20], because it offers the most complete set of base/conjugate acid pairs. The adopted correction was validated against density functional theory (DFT) calculations on the small ragweed pollen allergen protein. DFT calculations were performed using the Becke exchange [21] and Lee-Yang-Parr [22] correlation functionals (BLYP) within a hybrid Gaussian. A plane wave approach was adopted [23], along with norm-conserving pseudopotentials [24], to describe the core electrons. The TZV2P Gaussian basis set was used for valence electrons of all atoms, while the auxiliary electron density was expanded in plane waves up to a cutoff of 280 Ry. Interaction between periodic images in the reciprocal space was removed according to the decoupling scheme presented in Ref. [25]. Dispersion energy was included according to Ref. [26]. We will refer to the dispersion-corrected DFT energy as DFT + D. As previously described [27], the adopted DFT scheme has been validated against more accurate quantum-chemical calculations (DFT/B3LYP and MP2). The wave function has been optimized according to Ref. [23]. The calculations were carried out with the CP2K code [23].

The comparison between DFT + D and corrected force field calculations was performed over 35 selected protomers (over a total of about 460 possible protomers) of the ragweed pollen allergen protein at q=1+. For each protomer, conformational sampling was carried out according to the previously described protocol [27]. Several sets of GB values taken from the literature [9, 28, 29] were tested. The best agreement between DFT + D and corrected force field was obtained for the GBs calculated previously [27] for amino acids in an extended conformation, where the ionized groups do not make any short-range interaction such as hydrogen bonds or salt bridges. Indeed, this type of interaction is reasonably well described by the force field and there is no need to include it in the calculation of the amino-acid reference GB.

DFT energies do not correlate with non-corrected force-field energies (Figure SI-1A). The addition of the correction introduces a marked linear correlation between the two different energy evaluations (R^2=0.93, Figure SI-1B). The data dispersion indicates that the corrected force field allows one to discriminate between high and low energy protomers but not to identify small energy differences and, thus, to identify the single lowest-energy protomer. The standard error of the estimate using the linear correlation of Figure SI-1B is σ=35 kJ/mol. If we assume a normal distrubution of the DFT + D energies around the estimate obtained from the corrected force field, there is a confidence of 68.3 % and 99.7 % that the DFT + D energy is within σ and 3σ (about 100 kJ/mol) from the estimate, respectively. Indeed, all of the conformers located within 10 kJ/mol from the DFT + D minimum fall within 100 kJ/mol above the OPLS/AA. Similar discrepancies are found using the Amber [30] and GROMOS [31] force fields (data not shown). Hence, we carry out a statistical analysis of the protomer properties within a given energy cut off, which yields a high confidence to include the minimum-energy protomer. We will refer to these protomers as the most probable protomers. The discussion presented in this work is based on a 100 kJ/mol cut off. Only few structures are within this cut off (about 10 out of thousands). Different choices of the cut-off energy (from σ to 3σ) give comparable results (data not shown).

Using the OPLS/AA force field with GB corrections, protonation sites were randomly permuted and the total energy was calculated. At each MC step, a proton exchange is accepted or rejected according to the Metropolis criterion. The structure of each considered protomer was relaxed with the following simulated annealing-like procedure. First, a 400-ps, high-temperature (400 K) MD simulation was performed. This temperature was selected after several careful tests and it allows for an exhaustive sampling of side-chain conformations without disrupting, in the relatively short simulation time, the protein secondary structure. The resulting trajectory was split in 60 equally spaced time windows. In each of these windows, the geometry of the lowest-energy conformation was optimized. The optimized structure was then employed in the MC procedure. For each value of net charge, about 10^3 lowest-energy configurations were sampled. This procedure converges in a relatively small number of MC steps. Indeed, MC searches starting from different protomers substantially yielded the same final charge configurations, differing at the most in the position of one or two protons.

The initial structure for the MC procedure was extracted from a 2-ns MD simulation at ambient conditions of the protein in aqueous solution (with counter-ions added). These preparatory simulations were long enough to stabilize the protein dynamics, as deduced from the root mean square displacement (RMSD) of the backbone heavy-atoms. In all cases, the structure closest to the average one was taken. The initial charge configuration for the MC process, instead, was randomly generated allowing for positively charged basic residues and negatively charged acidic residues.

The time evolution of any lowest-energy protomer obtained from the MC/MD protocol was followed at 300 K for 20 ns (Table 2) and, in the case of the q=8+ ubiquitin ion, for 1 µs (Figure SI-9). All the calculations were carried

Table 2. Average Structural Properties at 300 K in Water and in Vacuo for the Proteins of Table 1. From Left to Right: Radius of Gyration in Water ($R_{g,\ wat}$ in Å); Radius of Gyration of the Backbone in Water ($R_{g,\ bb\ wat}$ in Å) Radius of Gyration in Gas-Phase ($R_{g,\ gp}$ in Å); Radius of Gyration of the Backbone in Gas-Phase ($R_{g,\ bb\ gp}$ in Å); Protein–Protein Hydrogen Bonds in Water (HB_{wat}); Protein–Protein Hydrogen Bonds in Gas-Phase (HB_{gp}); Total Surface Area in Water ($S_{tot,\ wat}$ in $Å^2$); Total Surface Area in Gas-Phase ($S_{tot,\ gp}$ in $Å^2$; Hydrophobic Surface in Water ($S_{pho,\ wat}$ in $Å^2$); Hydrophobic Surface in Gas-Phase ($S_{pho,\ gp}$ in $Å^2$); Root Mean Square Displacement of the Backbone in Water ($RMSD_{wat}$ in Å); Root Mean Square Displacement of the Backbone in Gas-Phase ($RMSD_{gp}$ in Å). The Proteins in Vacuo are in the Charge State Predicted in the Present Study, whereas the Proteins in Water have all of the Acidic and Basic Residues Ionized but Histidines, which are Assumed to be Neutral. Standard Deviations are Reported in Parentheses

Protein	$R_{g,\ wat}$	$R_{g,\ bb\ wat}$	$R_{g,\ gp}$	$R_{g,\ bb\ gp}$	HB_{wat}	HB_{gp}	$S_{tot,\ wat}$	$S_{tot,\ gp}$	$S_{pho,\ wat}$	$S_{pho,\ gp}$	$RMSD_{wat}$	$RMSD_{gp}$
1BBG	0.93 (0.01)	0.88 (0.01)	0.86 (0.01)	0.83 (0.00)	16.00 (2.56)	32.0 (2.5)	34.92 (1.22)	25.73 (0.51)	18.21 (0.63)	14.41 (0.94)	0.54 (0.20)	0.29 (0.01)
1UUA	1.15 (0.03)	1.11 (0.03)	1.01 (0.01)	0.98 (0.01)	22.3 (2.7)	42.8 (3.3)	53.54 (1.68)	42.24 (0.94)	28.99 (1.03)	28.13 (0.72)	0.47 (0.30)	0.24 (0.01)
1OK0	1.18 (0.01)	1.12 (0.01)	1.13 (0.01)	1.10 (0.01)	43.68 (2.4)	71.2 (3.4)	55.80 (0.79)	49.91 (1.08)	28.93 (0.57)	31.75 (0.84)	0.19 (0.01)	0.18 (0.01)
1V80	1.20 (0.01)	1.16 (0.01)	1.13 (0.00)	1.08 (0.01)	48.8 (3.8)	72.6 (3.8)	61.04 (1.29)	53.13 (1.16)	31.68 (1.19)	35.75 (0.93)	0.31 (0.02)	0.30 (0.01)
1C54	1.36 (0.02)	1.34 (0.02)	1.29 (0.01)	1.27 (0.01)	40.0 (4.7)	85.33 (4.1)	83.77 (1.62)	65.48 (1.25)	42.13 (0.96)	44.03 (0.93)	0.52 (0.02)	0.15 (0.01)
2PPP	1.34 (0.01)	1.30 (0.01)	1.34 (0.00)	1.30 (0.01)	81.9 (4.2)	103.4 (4.1)	79.46 (1.32)	74.22 (1.36)	41.57 (1.06)	50.69 (1.65)	0.16 (0.12)	0.32 (0.01)
1CTF	1.15 (0.01)	1.08 (0.01)	1.13 (0.01)	1.11 (0.01)	34.27 (2.8)	67.8 3.3)	57.54 (1.40)	48.73 (1.06)	30.90 (0.96)	36.19 (0.86)	0.54 (0.02)	0.52 (0.2)
1LZT	1.45 (0.01)	1.41 (0.01)	1.38 (0.01)	1.31 (0.01)	79.2 (4.5)	134.22 (5.2)	101.16 (1.77)	78.56 (1.59)	52.33 (1.31)	50.46 (1.20)	0.42 (0.01)	0.40 (0.01)
2CBA	1.99 (0.01)	1.95 (0.01)	1.76 (0.00)	1.74 (0.00)	160.8 (5.8)	249.46 (6.93)	190.01 (2.57)	154.13 (2.29)	104.30 (1.95)	102.59 (1.74)	0.86 (0.06)	0.23 (0.00)

out using the GROMACS [32] MD package. A time step of 1.5 fs for the integration of equations of motion was used in all of the simulations.

Calculation of the Apparent Gas-Phase Basicity

The GB_{app} is an extension of the concept of GB [4], and quantifies the ability of a protein, in a given conformation and charge state, to increase its charge state through the addition of a proton. The GB_{app} of a protein corresponds to the GB of the residue (embedded in the protein environment) with the highest gas-phase basicity. The $GB_{app,i}$ of the i-th residue in a protein with total charge q is defined as [4]

$$GB_{app,i} = GB_i - \left(E_{FF}^{(i,q)} - E_{FF}^{(i,q-1)}\right),$$

where GB_i is the GB of the i-th amino acid in the gas phase and $E_{FF}^{(i,q)}$ (or $E_{FF}^{(i,q-1)}$) is the energy of the protein with that residue protonated (or non protonated). In contrast to the original formulation, developed for a coarse-grained representation of an unfolded protein [4], we include in the GB_{app} calculation all of the classical energy terms considered by a force field. No vibrational correction was taken into account. The justification for this choice has been discussed previously in the literature [27, 33].

We stress that, in the present study, we do not report the GB_{app} of the lowest-energy protomer of a protein in a given charge state, which might be ill-defined as discussed in the previous section. Rather, we extrapolate trends of the average GB_{app} for a large set of proteins and charge states, and try to relate these trends to the experimentally observed CSDs and the GB of the solvent from which the ions have been originated.

Results and Discussion

Protein Structure in Vacuo

We have analyzed the conformational and protomer space of nine proteins featuring different size, fold, and pI (Table 1). Significant conformational rearrangements take place upon desolvation. However, the most probable protomers identified by the MC/MD procedure conserve their secondary and tertiary structure upon passing from the aqueous solution to the gas phase at room temperature (Table 2 and Figure SI-2). The gyration radius (R_g), calculated over the nanosecond time-scale, decreases in a similar way for all of the charge states considered here. This contraction involves the solvent-exposed side chains, which fold onto the protein surface and, to a lesser extent, also the backbone. These rearrangements lead to the formation of new intramolecular hydrogen bonds (HB_{pp}). The total surface area (A_{tot}) also decreases, whereas the hydrophobic portion of the total surface area (A_{phob}) increases, as already reported [34–36]. The proteins turn out to be more rigid in the gas phase, as indicated by the

RMSD of backbone atoms around their average positions. The most dramatic change is observed for the small pollen allergen, which has relatively flexible parts, pointing to the solvent in the native structure. Moreover, the protein structural rearrangements on passing from solution to gas phase lead to the formation of salt-bridges, as also reported in Ref. [34]. Further structural features (Table 2) are in close agreement to those found in previous simulation studies [34–36] and, hence, their detailed description is omitted.

Gas-phase Basicity and Protein Ionization

The number of ionized residues, n_{IR}, and the number of hydrogen bonds formed by ionized residues, n_{iHB}, per unit of protein surface is constant among our protein data set (Figure SI-3), with about 0.16 ionized residues and 0.43 hydrogen bonds per nm^2, for the most probable protomers of the proteins considered here, in their predicted most populated charge state: $n_{\mathrm{IR}}=0.148S+1.459$ ($R^2=0.976$) and $n_{\mathrm{iHB}}=0.434S-1.438$ ($R^2=0.989$), with S in nm^2. Both positive and negative ionized residues tend to form the maximum number of hydrogen bonds, compatible with the geometry of the gas-phase structures (Table 3). In general, protonated amino groups donate three hydrogen bonds (one for each N–H bond) and carboxylates receive four hydrogen bonds. The average number of hydrogen bonds per residue type is roughly constant across the proteins investigated. The differences are within the standard deviation (Table 3).

Zwitterionic states are mostly retained, especially for low charge states, as can be seen by plotting the number of ionized residues as a function of the protein total charge (Figure 1 and Figure SI-6). Indeed, intramolecular hydrogen bonds, salt-bridges, π-charge interactions, and long-range electrostatic interactions can compensate for the thermodynamic penalty of charge separation in vacuo, providing internal solvation [27]. This finding is consistent with recent experimental evidence [6, 37, 38]. It also supports the hypothesis that a higher propensity for zwitterionic states of folded versus unfolded proteins can lower the net charge of the former, contributing to conformational effects in ESI-MS [39]. In contrast, there is no clear dependence of the number of salt bridges on the protein size (Figure SI-4). This result suggests that the formation of salt bridges depends on peculiar features of the protein structure.

Table 3. Average Number of Hydrogen Bonds Formed by each Type of Ionized Residue in the Most Probable Protomers (Protomers Under 100 kJ/mol from the Energy Minimum) for the Proteins of Table 1. The Total Number of Ionized Hydrogen Bonds is Reported in the Last Column. The Standard Deviation from the Average is Given in Parentheses

Protein	LYS	ARG	ASP	GLU
1BBG	2.1(0.6)	-	3.0(1.0)	4.0(0.0)
1UUA	2.3(0.9)	2.7(0.6)	3.2(0.3)	3.0(0.0)
1CTF	2.1(0.7)	-	3.4(0.3)	3.6(0.5)
1V80	2.2(0.7)	2.6(1.2)	-	3.5(0.7)
2PPP	2.2(0.7)	3.2(1.2)	3.6(0.9)	4.0(0.0)
1C54	-	3.4(0.9)	3.0(0.0)	3.0(0.0)
1LZT	2.4(0.6)	2.7(1.1)	3.4(0.5)	-
2CBA	1.9(0.7)	2.7(0.6)	2.5(1.0)	2.3(0.9)

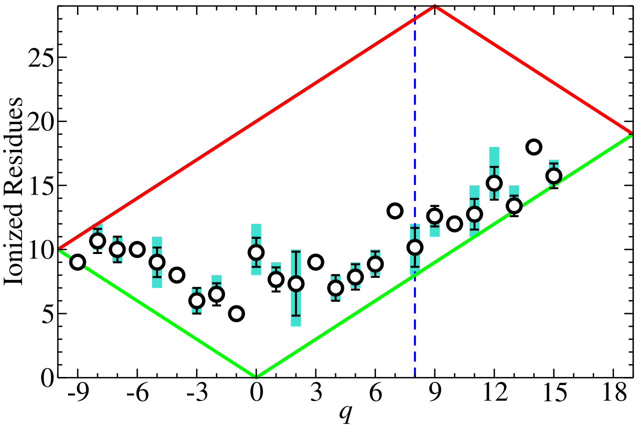

Figure 1. Average number of ionized residues (circles) in the most probable protomers of the hen-egg white lysozyme as a function of the protein net charge (q). The data for the other proteins considered in the present study are reported in Figure SI-6. Standard deviation from the average is given as error bar. The interval of ionized residues spanned by the most probable protomers for each charge is reported as a cyan bar. The minimum and the maximum number of possible ionized residues for each total charge are indicated by the green and the red line, respectively. The vertical dashed blue line indicates the main charge state starting from aqueous solutions

The GB$_{\mathrm{app}}$ values decreases linearly as the protein net charge increases (Table SI-1, $R^2 \geq 0.99$ for all of the nine proteins) [4, 9, 40]. As expected, the slope of the line depends on the specific protein. Most notably, the intersection of the GB$_{\mathrm{app}}$ fitted line with the line of solvent GB corresponds with remarkable agreement to the experimental main (most abundant) charge state under mild ESI conditions [7, 10] (Figure 2 and Figure SI-7). Instead, GB$_{\mathrm{app}}$ turns out to be systematically underestimated when the calculation is performed constraining the non-hydrogen atoms in the position of the NMR or X-ray structures (Figure 2). By reproducing the experimental charge for proteins of different size, our calculations are also consistent with the well-known charge-to-mass empirical relation $q \propto \sqrt{M}$ [10, 41].

It has been previously shown that such an empiric charge-to-mass relation reflects a linear log-log charge-to-surface relation, which has been proven to hold both for folded [41, 42] and unfolded [43] proteins. The relation holds when comparing the surface of hydrated proteins or protein complexes with different surface-to-mass relation [41]. Our results also support a linear log–log correlation between charge and desolvated protein

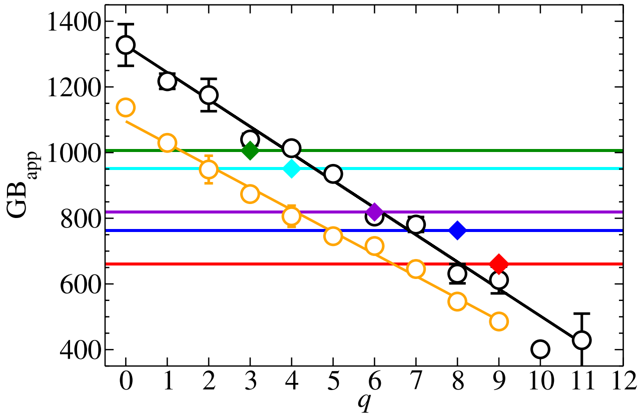

Figure 2. Average GB_{app} (in kJ/mol) calculated for the most probable protomers of hen-egg white lysozyme. Black circles and orange circles represent, respectively, values calculated from optimized and the non-optimized (pdb) structures. The black line and orange line are the result of a linear fitting (in both cases the correlation coefficient, R^2, is around 0.995). Standard deviation from the average is given as error bar (when not visible, the standard deviation is smaller than the symbol size). The data for the other proteins considered in the present study are reported in Figure SI-7. The horizontal lines indicate the GB of various solvents: water (red line), isopropanol (blue line), ammonia (purple line), triethylammonium bicarbonate (cyan line), 1,5-diazabicyclo[4.3.0]-5-ene (green line). The experimental main charge states observed from these solvents are shown by symbols colored accordingly. Comparison with the main charge state is made, instead of maximum or average charge state, because of its less ambiguous determination from literature data [39, 51]

surface. Indeed, a linear correlation seems to be respected when the area calculated from the gas-phase structures is employed (Figure SI-5). However, more datapoints will be needed to explore larger molecular weights.

A Simple Model for Proteins in the Gas Phase

We now extend the correlation between charge and mass to any folded protein, by introducing a simple and general model energy function, without adjustable parameters. This model is based on a plausible assumption: the experimentally observed charge-to-mass relation can be interpreted as a linear combination of energetic contributions due to electrostatic repulsion and internal solvation. The latter is considered to be proportional to the protein surface and it is based on the above observation that the number of ionized residues and the number of hydrogen bonds they present per surface unit is constant.

We start by modeling a protein in the gas phase as a sphere of radius R and density ρ, with a net charge q uniformly distributed on its surface. The electrostatic energy of the protein can be expressed as:

$$U(q) = \frac{1}{2}\frac{1}{4\varepsilon_0\pi}\frac{q^2}{R} - 4\pi\xi R^2. \qquad (1)$$

The quadratic dependence of the electrostatic energy on the total protein charge (first term) accounts for the linear change in GB_{app} described above (see also Figure SI-8). The second term takes into account the stabilization by intermolecular interactions. It is proportional to the surface area $4\pi R^2$ via a parameter to be determined by fitting our computational results ($\xi\approx0.994$ N/m with ρ from Ref. [52]). The protein is unstable when $U(q)\geq0$. Hence, the maximum charge attainable is

$$q = 4\pi\sqrt{2\varepsilon_0\xi R^3} \qquad (2)$$

or

$$q = 2\sqrt{\frac{6\pi\varepsilon_0\xi}{\rho}}\sqrt{M}, \qquad (3)$$

being M=4/3$\pi\rho R^3$.

A numerical model based on a charged ellipsoid shows that, also in such a case, $q \propto \sqrt{M}$ (see Supplementary Information). Thus, the $q \propto \sqrt{M}$ relation is rather independent on the specific shape being used. More generally, it is easy to see from a simple dimensional analysis that any stabilizing contribution ($-\xi S$) that depends on the surface area (S) yields a $\sqrt{M}$ dependence of the maximum charge possible. Indeed, the electrostatic potential is dimensionally proportional to charge^2×length^{-1}, whereas

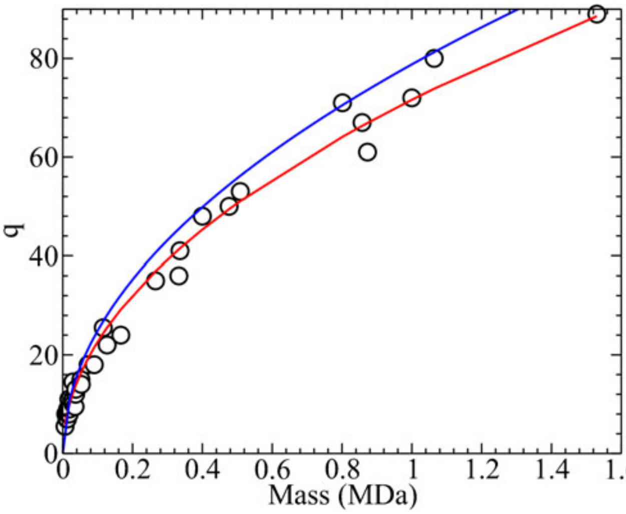

Figure 3. Experimental average charge state as a function of the protein mass [10]. The curves predicted by the Rayleigh-charge hypothesis [10] (blue line, reduced χ^2=0.51) and the spherical model introduced here (red line, reduced χ^2=0.25) are also shown

the stabilizing contribution is proportional to length^2. Thus, the square of maximum charge is proportional to a volume, and consequently the charge must be proportional to the square root of the mass.

Substituting our fitted ξ value in Equation (3) yields, with remarkable accuracy, not only the main charge state displayed by the proteins considered in this study, but also the experimental q versus M curve for a large variety of other proteins (Figure 3). Notice that we compare our predicted maximum charges q with the reported experimental main charge. This approximation is justified by the fact that experimental main and maximum charges are very similar for folded proteins [10] and they are related to each other.

Our simple model Equation (3) yields formally a result similar to that obtained with the Rayleigh-charge hypothesis. Both predict a $\sqrt{M}$ dependence of the protein charge. However, in contrast to the Rayleigh-charge hypothesis, the present model does not involve adjustable parameters and does not introduce any dependence on the solvent surface tension [12]. The only parameter entering in Equation (3) is the stabilization term due to internal solvation and it is obtained from the quantities calculated from the set of proteins considered here. This model eliminates the explicit dependence on surface tension that derives from the application of the Rayleigh equation to the prediction of protein ionization by electrospray. At the same time, the simplifying assumptions of the model do not allow to capture the indirect role that the solvent surface tension might play on protein ionization via its effect on the electrospray mechanism. Surface tension is the limiting factor for droplet charging during electrospray, as indicated by the Rayleigh equation, and conditions might be found in which it becomes the limiting factor also for protein ionization [1, 2].

It should also be underscored that the results of this study do not help discriminating between the ion-evaporation model and the charged residue model, concerning the mechanism of production of gas-phase ions during ESI [1, 2]. The present model is not meant to test those hypotheses and is compatible with both mechanisms.

Conclusions

The present computational study establishes the relevance of GB_{app} for folded proteins under electrospray conditions. The calculations do not assume a role for the charge of the precursor ESI droplet. Our proposed model reproduces the well-known relationship between observed charge and protein mass based only on intrinsic protein features and solvent nature. Hence, the liquid medium of the precursor droplet provides or accepts protons according to its GB relative to the GB_{app} of the protein. Along with previous studies [2, 4], these results support a model in which the same principle (i.e., the GB_{app} of the protein relative to the GB of the solvent) is applied to folded and unfolded proteins, in order to explain the experimentally observed charge values.

Acknowledgment

The authors gratefully acknowledge CNR/INFM and Juelich Super Computing for a large allocation of computer time, respectively, at the CINECA and JUGENE supercomputing centers.

References

1. Verkerk, U.H., Kebarle, P.: Ion–ion and ion–molecule reactions at the surface of proteins produced by nanospray. Information on the number of acidic residues and control of the number of ionized acidic and basic residues. *J. Am. Soc. Mass Spectrom.* **16**, 1325–1341 (2005)
2. Kebarle, P., Verkerk, U.H.: Electrospray: From ions in solution to ions in the gas phase, what we know now. *Mass Spectrom. Rev.* **28**, 898–917 (2009)
3. Konermann, L.: A minimalist model for exploring conformational effects on the electrospray charge state distribution of proteins. *J. Phys. Chem. B* **111**, 6534–6543 (2007)
4. Schnier, P.D., Gross, D.S., Williams, E.R.: On the maximum charge-state and proton-transfer reactivity of peptide and protein ions formed by electrospray-ionization. *J. Am. Soc. Mass Spectrom.* **6**, 1086–1097 (1995)
5. Gronert, S.: Coulomb repulsion in multiply charged ions: a computational study of the effective dielectric constants of organic spacer groups. *Int. J. Mass Spectrom.* **187**, 351–357 (1999)
6. Peschke, M., Verkerk, U.H., Kebarle, P.: Prediction of the charge states of folded proteins in electrospray ionization. *Eur. J. Mass Spectrom.* **10**, 993–1002 (2004)
7. Isabel Catalina, M., van den Heuvel, R.H.H., van Duijn, E., Heck, A.J.R.: Decharging of globular proteins and protein complexes in electrospray. *Chem.-Eur. J.* **11**, 960–968 (2005)
8. Touboul, D., Jecklin, M.C., Zenobi, R.: Investigation of deprotonation reactions on globular and denatured proteins at atmospheric pressure by ESSI-MS. *J. Am. Soc. Mass Spectrom.* **19**, 455–466 (2008)
9. Schnier, P.D., Gross, D.S., Williams, E.R.: Electrostatic forces and dielectric polarizability of multiply protonated gas-phase cytochrome *c* ions probed by ion/molecule chemistry. *J. Am. Chem. Soc.* **117**, 6747–6757 (1995)
10. Fernandez de la Mora, J.: Electrospray ionization of large multiply charged species proceeds via dole's charged residue mechanism. *Anal. Chim. Acta* **406**, 93–104 (2000)
11. Iavarone, A.T., Williams, E.R.: Mechanism of charging and supercharging molecules in electrospray ionization. *J. Am. Chem. Soc.* **125**, 2319–2327 (2003)
12. Samalikova, M., Grandori, R.: Testing the role of solvent surface tension in protein ionization by electrospray. *J. Mass Spectrom.* **40**, 503–510 (2005)
13. Samalikova, M., Grandori, R.: Protein charge-state distributions in electrospray-ionization mass spectrometry do not appear to be limited by the surface tension of the solvent. *J. Am. Chem. Soc.* **125**, 13352–13353 (2003)
14. Lomeli, S.H., Yin, S., Ogorzalek Loo, R.R., Loo, J.A.: Increasing charge while preserving noncovalent protein complexes for esi-ms. *J. Am. Soc. Mass Spectrom.* **20**, 593–596 (2009)
15. Lomeli, S.H., Peng, I.X., Yin, S., Ogorzalek Loo, R.R., Loo, J.A.: New reagents for increasing ESI multiple charging of proteins and protein complexes. *J. Am. Soc. Mass Spectrom.* **21**, 127–131 (2010)
16. Sterling, H.J., Cassou, C.A., Trnka, M.J., Burlingame, A.L., Krantz, B.A., Williams, E.R.: The role of conformational flexibility on protein supercharging in native electrospray ionization. *Phys. Chem., Chem. Phys.* **13**, 18288–18296 (2011)
17. Prakash, H., Mazumdar, S.: Direct correlation of the crystal structure of proteins with the maximum positive and negative charge states of gaseous protein ions produced by electrospray ionization. *J. Am. Soc. Mass Spectrom.* **16**, 1409–1421 (2005)
18. Jarrold, M.F.: Peptides and proteins in the vapor phase. *Annu. Rev. Phys. Chem.* **51**, 179–207 (2000)

19. Patriksson, A., Adams, C.M., Kjeldsen, F., Zubarev, R.A., van der Spoel, D.: A direct comparison of protein structure in the gas and solution phase: The trp-cage. *J. Phys. Chem. B* **111**, 13147–13150 (2007)
20. Jorgensen, W.L., Tirado-Rives, J.: Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6665–6670 (2005)
21. Becke, A.D.: Density-functional exchange-energy approximation with correct asymptotic-behavior. *Phys. Rev. A* **38**, 3098–3100 (1988)
22. Lee, C.T., Yang, W.T., Parr, R.G.: Development of the Colle-Salvetti correlation-energy formula into a functional of the electron-density. *Phys. Rev. B.* **37**, 785–789 (1988)
23. VandeVondele, J., Krack, M., Mohamed, F., Parrinello, M., Chassaing, T., Hutter, J.: Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Comp. Phys. Comun.* **167**, 103–128 (2005)
24. Hartwigsen, C., Goedecker, S., Hutter, J.: Relativistic separable dual-space Gaussian pseudopotentials from h to rn. *Phys. Rev. B* **58**, 3641–3662 (1998)
25. Martyna, G.J., Tuckerman, M.E.: A reciprocal space based method for treating long range interactions in ab initio and force-field-based calculations in clusters. *J. Chem. Phys.* **110**, 2810–2821 (1999)
26. Grimme, S.: Semiempirical gga-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **27**, 1787–1799 (2006)
27. Marchese, R., Grandori, R., Carloni, P., Raugei, S.: On the zwitterionic nature of gas-phase peptides and protein ions. *PLOS Comput. Biol.* **6**, 1–11 (2010)
28. Li, Z., Matus, M.H., Velazquez, H.A.: D. A Dixon, C. J. Cassady. Gas-phase acidities of aspartic acid, glutamic acid, and their amino acid amides. *Int. J. Mass Spectrom.* **265**, 213–223 (2007)
29. Harrison, A.G.: The gas-phase basicities and proton affinities of amino acids and peptides. *Mass Spectrom. Rev.* **16**, 201–217 (1997)
30. Case, D.A., Cheatham, T., Darden, T., Gohlke, H., Luo, L., Merz, K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.: *J. Comp. Chem.* **26**, 1668–1688 (2005)
31. Oostenbrink, C., Villa, A., Mark, A.E., van Gunsteren, W.F.: A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53a5 and 53a6. *J. Comput. Chem.* **25**, 1656–1676 (2004)
32. Lindahl, E., Hess, B., van der Spoel, D.: Gromacs 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **7**, 306–317 (2001)
33. Berka, K., Laskowski, R., Riley, K.E., Hobza, P., Vondrekásek, J.: Representative amino acid side chain interactions in proteins. a comparison of highly accurate correlated ab initio quantum chemical and empirical potential procedures. *J. Chem. Theory Comput.* **5**, 982–992 (2009)
34. Patriksson, A., Marklund, E., van der Spoel, D.: Protein structures under electrospray conditions. *Biochemistry* **46**, 933–945 (2007)
35. Meyer, T., de la Cruz, X., Orozco, M.: An atomistic view to the gas phase proteome. *Structure* **17**, 88–95 (2009)
36. D'Abramo, M., Meyer, T., Bernado, P., Pons, C., Fernandez-Recio, J., Orozco, M.: On the use of low-resolution data to improve structure prediction of proteins and protein complexes. *J. Chem. Theory Comp.* **5**, 3129–3137 (2009)
37. Heck, A.J.R., van den Heuvel, R.H.H.: Investigation of intact protein complexes by mass spectrometry. *Mass Spectrom. Rev.* **23**, 368–389 (2004)
38. Krusemark, C.J., Frey, B.L., Belshaw, P.J., Smith, L.M.: Modifying the charge state distribution of proteins in electrospray ionization mass spectrometry by chemical derivatization. *J. Am. Soc. Mass Spectrom.* **20**, 1617–1625 (2009)
39. Grandori, R.: Origin of the conformation dependence of protein charge-state distributions in electrospray ionization mass spectrometry. *J. Mass Spectrom.* **38**, 11–15 (2003)
40. Gross, D.S., Schnier, P.D., Rodriguez-Cruz, S.E., Fagerquist, C.K., Williams, E.R.: Conformations and folding of lysozyme ions in vacuo. *Proc. Natl. Acad. Sci. USA* **93**, 3143–3148 (1996)
41. Kaltashov, I.A., Mohimen, A.: Estimates of protein surface areas in solution by electrospray ionization mass spectrometry. *Anal. Chem.* **77**, 5370–5379 (2005)
42. Halgand, F., Leprévote, O.: Mean charge state and charge state distribution of proteins as structural probes. an electrospray ionization mass spectrometry study of lysozyme and ribonuclease A. *Eur. J. Mass Spectrom.* **7**, 433–439 (2001)
43. Testa, L., Brocca, S., Grandori, R.: Charge-surface correlation in electrospray ionization of folded and unfolded proteins. *Anal. Chem.* **83**, 6459–6463 (2011)
44. Fenaille, F., Nony, E., Chabre, H., Lautrette, A., Couret, M.-N., Batard, T., Moingeon, P., Ezan, E.: Mass spectrometric investigation of molecular variability of grass pollen group 1 allergens. *J. Proteome Res.* **8**, 4014–4027 (2009)
45. Konermann, L., Douglas, D.J.: Equilibrium unfolding of proteins monitored by electrospray ionization mass spectrometry: Distinguishing two-state from multi-state transitions. *Rapid Commun. Mass Spectrom.* **12**, 435–442 (1998)
46. Benjamin, D.R., Robinson, V.C., Hendrick, J.P., Hartl, F.U., Dobson, C.M.: Mass spectrometry of ribosomes and ribosomal subunits. *Proc. Natl. Acc. Sci. U.S.A.* **95**, 7391–7395 (1998)
47. Douglas, D.J., Collings, B.A., Numao, S., Nesatyy, V.J.: Detection of noncovalent complex between a-amylase and its microbial inhibitor tendamistat by electrospray ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **15**, 89–96 (2001)
48. Hoerner, J.K., Xiao, H., Kaltashov, I.A.: Structural and dynamic characteristics of a partially folded state of ubiquitin revealed by hydrogen exchange mass spectrometry. *Biochemistry* **44**, 11286–11294 (2005)
49. Smith, R.D.: Evolution of ESI–mass spectrometry and Fourier transform ion cyclotron resonance for proteomics and other biological applications *Int. J. Mass Spectrom.* **200**, 509–544 (2000)
50. Samalikova, M., Grandori, R.: Role of opposite charges in protein electrospray ionization mass spectrometry. *J. Mass Spectrom.* **38**, 941–947 (2003)
51. Wang, H., Zhang, X., Xiao, J., Yang, S., Nie, A., Zhao, L., Li, S.: ESI-MS study on noncovalent bond complex of rhfkbp12 and new neurogrowth promoter. *Sci. China Ser. C* **46**, 286–292 (2003)
52. Fischer, H., Polikarpov, I., Craievich, A.F.: Average protein density is a molecular-weight-dependent function. *Protein Sci.* **13**, 2825–2828 (2004)