[g012]

Correction of Charge-Transfer Indices for Multifunctional Amino Acids: Application to Lysozyme

Francisco Torrens* **and Gloria Castellano**

Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P. O. Box 22085, E-46071 València, Spain. http://www.uv.es/~uiqt/index.htm. Tel. +34 963 544 431, fax +34 963 543 274

* Author to whom correspondence should be addressed. E-mail: francisco.torrens@uv.es

Abstract: Valence topological charge-transfer (CT) indices are applied to the calculation of pH at the pI isoelectric point. The combination of CT indices allows the estimation of pI. The model is generalized for molecules with heteroatoms. The ability of the indices for the description of molecular charge distribution is established by comparing them with the pI of 21 amino acids. Linear correlation models are obtained. The CT indices improve multivariable regression equations for pI. The variance decreases by 95%. No superposition of the corresponding G_k–J_k and G_k^V–J_k^V pairs is observed in most fits, which diminishes the risk of collinearity. The inclusion of heteroatoms in π-electron system is beneficial for the description of pI, owing to either the role of the additional p orbitals provided by heteroatom or role of *steric* factors in π-electron conjugation. The use of only CT and valence CT indices $\{G_k,J_k,G_k^V,J_k^V\}$ gives limited results for modelling pI of amino acids. Furthermore, the inclusion of the numbers of acidic and basic groups improves all models. The effect is specially noticeable for amino acids with more than two functional groups. The fitting line obtained for the 21 amino acids can be used to estimate the isoelectric point of lysozyme and its fragments, by only replacing $(1+\Delta n/n_T)$ with $(M+\Delta n)/n_T$. For lysozyme, the results of smaller fragments can estimate that of the whole protein with 1–13% errors.

Keywords: charge distribution, valence topolgical charge-transfer index, isoelectric point, amino acid, lysozyme.

Introduction

During the simulation of pH at the pI isoelectric point of $n = 21$ amino acids, indices D and D^v were introduced together with the concept of fragmentary molecular connectivity indices, *i.e.*,

indices that were mainly determined by the characteristics of the secondary functional groups in amino acids [1,2]. As this property is highly dependent on the type of side chain an amino acid has, the normal connectivity indices of set eight achieved a totally unsatisfactory modelling. The construction of the first fragmentary molecular connectivity indices was awkward. An entirely new and sound set of fragmentary molecular connectivity terms was proposed, which were derived with an easy trial-and-error procedure [3–5]. These terms are defined in the following way

$$X_{\mathrm{p}I} = \frac{\chi}{{}^{0}\chi^{v}}\left(1+\frac{\Delta n}{n_T}\right) \tag{1}$$

where $\Delta n = n_{\mathrm{A}} - n_{\mathrm{B}}$, n_{A} = number of acidic groups (two for Asp and Glu, one for all others), n_{B} = number of basic groups (two for His and Lys, three for Arg, as well as one for all others), and $n_T = n_{\mathrm{A}} + n_{\mathrm{B}}$ (total number of functional groups); notice that for $n_T = 2$, $\Delta n = 0$. Clearly there are eight such terms following the type of index that enters in numerator χ. The nomenclature for such terms can be defined in the following way for $\chi = D^{v} \rightarrow X \equiv {}^{D}X^{v}$, *etc*. The best single descriptor for pI is ${}^{0}X^{v}$ with $Q = 2.12$, $F = 267$, $r = 0.966$, $s = 0.46$, $\mathbf{u} = (16,28)$. The statistics, specially the utility statistic, seem quite satisfactory. Now statistic Q can be improved at the expenses of statistics F and $\mathbf{u}$, with the following linear combination of X terms made up of connectivity indices, which can be derived by the aid of both forward and full combinatorial techniques

$$\left\{{}^{D}X^{v}, {}^{0}X, {}^{0}X^{v}, {}^{1}X\right\}: \quad Q = 2.53, \quad F = 95, \quad r = 0.980, \quad s = 0.39, \quad \mathbf{u} = (3.1,2.8,4.7,2.8,26)$$

Average <**u**> drops from 22.4 to 7.9, the utility of ${}^{0}X^{v}$ drops dramatically, and only the unitary index maintains a good utility.

To improve these utilities and detect possibly dominant descriptors, use is made of the following vector of orthogonalized terms: $\Omega = ({}^{1}\Omega, {}^{2}\Omega, {}^{3}\Omega, {}^{4}\Omega, U_0)$, where ${}^{1}\Omega \equiv {}^{0}X^{v}$, ${}^{2}\Omega \leftarrow {}^{D}X^{v}$, ${}^{3}\Omega \leftarrow {}^{1}X$, ${}^{4}\Omega \leftarrow {}^{0}X$. The orthogonalized vector shows the following utilities: $\mathbf{u} = (19,1.3,1.0,2.8,33)$. The utility vector indicates that only the first ${}^{1}\Omega \equiv {}^{0}X^{v}$ and last $U_0 \equiv \Omega^{0} \equiv 1$ parameters are important descriptors. We are thus back to the single-term description but with an enhanced utility for ${}^{1}\Omega$ and U_0: 19 and 33 instead of 16 and 28. Notice that the statistical score of the molar masses for pI is $Q = 0.002$ and $F = 0.14$. An inspection of the interrelation between the eight terms confirms their small interrelation as $\langle r_{\mathrm{IM}}(\mathrm{p}I:\{X\})\rangle = 0.560$, $r_w({}^{D}X, X_t) = 0.004$ and $r_s({}^{D}X, {}^{1}X) = 0.975$, where r_w and r_s stand for the weakest and strongest interrelations, respectively. A critical analysis of the ${}^{0}X^{v}$ term lets us notice that this term is trivial, as it is nothing other than $(1 + \Delta n/n_T)$ [6]. Now as the best description is given by a relation consisting of only this term, this means that molecular connectivity indices are not needed to simulate this property. Let us resort to a deeper trial-and-error search, discovering the following not-at-all trivial term

$$X'_{\mathrm{p}I} = \frac{\left({}^{1}\chi^{v}\right)^{0.5} - 180\chi_t^{v}}{D}\left(0.04\chi_t^{v} + \frac{\Delta n}{n_T}\right) \tag{2}$$

The modelling power of this dominant term is remarkable: $Q = 3.41$, $F = 693$, $r = 0.987$, $s = 0.29$, $\langle \mathbf{u} \rangle = 58$, $\mathbf{u} = (26,90)$, and the correlation vector $\mathbf{C} = (77.99429,5.75382)$. Thus the final modelling equation can be written as $pI = 5.75 + 77.99X'_{pI}$. Not only is the improvement in F and $\mathbf{u}$ more than expected but, furthermore, this term is a highly dominant *dead-end* term, as it does not allow any better combination with any other index or term. The term like the preceding ${}^0X^v$ term is mainly based on valence-type molecular connectivity indices, an expected result as side-chain functional groups in amino acids are rich in double bonds and lone-pair electrons.

The generation and decomposition of amino-acid and peptide radicals are processes of great biological importance, due to their connection to the oxidative damage caused by ionizating radiation or oxidizing agents [7,8]. Moreover, several experimental studies showed that amino-acid and peptide radical cations can be generated by the electrospray technique and peptide cationization using Cu^{2+} [9]. The mass spectra obtained in these cases are rich and differ considerably from those of protonated systems, which can provide useful information in peptide sequencing. In order to shed some light on the properties of amino-acid and peptide radical cations, the group of Sodupe performed quantum chemical calculations on nine amino acids and the smallest *N*-glycylglycine peptide [10,11]. They discussed the influence of intramolecular hydrogen bonds and amino-acid side chain on the localization of the electron hole upon oxidation and subsequent fragmentation process. They showed that for systems involving aromatic amino acids, oxidation is mainly produced at the side chain, whereas for non-aromatic ones oxidation is produced either at the basic NH_2 or CO groups, the nature of the electron hole depending on the existent intramolecular hydrogen bonds. In earlier publications, topological charge-transfer (CT) indices were applied to the calculation of the molecular dipole moment of hydrocarbons [12], valence-isoelectronic series of benzene, styrene [13,14] and cyclopentadiene [15], as well as phenyl alcohols [16] and 4-alkylanilines [17]. In the present report, the valence CT indices have been applied to the calculation of pH at the pI isoelectric point of 21 amino acids. Section 2 presents the CT indices and their generalization for heteroatoms. Section 3 presents and discusses the calculation results. Section 4 summarizes the conclusions.

Results and Discussion

The molecular CT indices G_k, J_k, G_k^V and J_k^V (with $k < 6$) are reported in Table 1 for 21 amino acids. Hydroxyproline (4-hydroxypyrrolidine-2-carboxylic acid, Hyp) differs from proline (Pro) by the presence of a hydroxyl (–OH) group attached to the C_γ atom. The G_k indices contain both CT and size effects, *e.g.*, $G_k(\text{Pro}) < G_k(\text{Hyp})$. The size effect is eliminated in the J_k, *e.g.*, $J_2(\text{Pro}) > J_2(\text{Hyp})$. The effect of heteroatoms is included in both G_k^V and J_k^V, *e.g.*, $G_4^V(\text{Pro}) > G_4^V(\text{Hyp})$.

Table 1. Values of the G_k and J_k charge-transfer indices up to fifth order for 21 amino acids (AA).

AA	*N*	G_1	G_2	G_3	G_4	G_5

Ala	6	2.5000	2.6667	0.5000	0.0000	0.0000
Arg	12	4.2500	6.2222	1.2500	0.4622	0.2639
Asn	9	4.0000	4.7778	0.9375	0.3689	0.1250
Asp	9	4.0000	4.7778	0.9375	0.3689	0.1250
Cys	7	2.5000	2.8889	0.6875	0.1111	0.0000
Gln	10	4.0000	5.0000	1.0000	0.3422	0.2708
Glu	10	4.0000	5.0000	1.0000	0.3422	0.2708
Gly	5	2.0000	2.3333	0.2500	0.0000	0.0000
His	11	3.7500	7.5000	1.2569	0.5211	0.3472
Hyp	9	3.0000	2.8889	0.8750	0.3422	0.0625
Ile	9	3.0000	3.3333	1.0000	0.3422	0.0625
Leu	9	3.5000	2.8889	0.9375	0.3511	0.1250
Lys	10	2.5000	2.8889	0.8125	0.3111	0.2014
Met	9	2.5000	2.8889	0.8125	0.3111	0.1458
Phe	12	3.2500	9.2222	1.2500	0.6844	0.4306
Pro	8	2.0000	2.6667	0.6528	0.2222	0.0000
Ser	7	2.5000	2.8889	0.6875	0.1111	0.0000
Thr	8	3.0000	3.1111	0.8750	0.2222	0.0000
Trp	15	4.2500	13.3889	2.0278	0.9989	0.6425
Tyr	13	4.5000	10.5556	1.6250	0.9867	0.4861
Val	8	3.0000	3.1111	0.8750	0.2222	0.0000

AA	J_1	J_2	J_3	J_4	J_5
Ala	0.5000	0.5333	0.1000	0.0000	0.0000
Arg	0.3864	0.5657	0.1136	0.0420	0.0240
Asn	0.5000	0.5972	0.1172	0.0461	0.0156
Asp	0.5000	0.5972	0.1172	0.0461	0.0156
Cys	0.4167	0.4815	0.1146	0.0185	0.0000
Gln	0.4444	0.5556	0.1111	0.0380	0.0301
Glu	0.4444	0.5556	0.1111	0.0380	0.0301
Gly	0.5000	0.5833	0.0625	0.0000	0.0000
His	0.3750	0.7500	0.1257	0.0521	0.0347
Hyp	0.3750	0.3611	0.1094	0.0428	0.0078
Ile	0.3750	0.4167	0.1250	0.0428	0.0078
Leu	0.4375	0.3611	0.1172	0.0439	0.0156
Lys	0.2778	0.3210	0.0903	0.0346	0.0224

Met	0.3125	0.3611	0.1016	0.0389	0.0182
Phe	0.2955	0.8384	0.1136	0.0622	0.0391
Pro	0.2857	0.3810	0.0933	0.0317	0.0000
Ser	0.4167	0.4815	0.1146	0.0185	0.0000
Thr	0.4286	0.4444	0.1250	0.0317	0.0000
Trp	0.3036	0.9563	0.1448	0.0713	0.0459
Tyr	0.3750	0.8796	0.1354	0.0822	0.0405
Val	0.4286	0.4444	0.1250	0.0317	0.0000

AA	G_1^V	G_2^V	G_3^V	G_4^V	G_5^V
Ala	4.5000	3.3222	1.2333	0.0000	0.0000
Arg	7.1500	6.9306	1.9722	0.6922	0.3497
Asn	6.8000	6.0889	1.5458	0.5058	0.2130
Asp	7.9000	6.0889	1.6625	0.5408	0.1250
Cys	4.5000	3.3222	1.4181	0.3861	0.0000
Gln	6.8000	5.7611	1.8472	0.5147	0.2062
Glu	7.9000	6.3111	1.9694	0.5835	0.2942
Gly	4.5000	2.9361	0.4944	0.0000	0.0000
His	8.6500	7.6583	1.8625	0.4696	0.3174
Hyp	7.8000	4.3583	1.1556	0.4597	0.0625
Ile	5.0000	3.5444	1.6028	0.8810	0.2385
Leu	5.5000	3.3222	1.4236	0.6036	0.4770
Lys	5.1000	3.3750	1.4153	0.4836	0.2663
Met	4.5000	3.3222	1.4181	0..4949	0.3103
Phe	5.2500	9.6556	1.7361	0.5642	0.3519
Pro	5.6000	3.7028	1.1361	0.5222	0.0000
Ser	6.2000	3.4278	1.2875	0.1111	0.0000
Thr	6.2000	3.9778	1.4722	0.4972	0.0000
Trp	6.9500	13.1028	2.5111	0.8436	0.4239
Tyr	7.2000	10.5444	1.8611	0.7289	0.4399
Val	5.0000	3.3222	1.6028	0.7722	0.0000

AA	J_1^V	J_2^V	J_3^V	J_4^V	J_5^V
Ala	0.9000	0.6644	0.2467	0.0000	0.0000
Arg	0.6500	0.6301	0.1793	0.0629	0.0318
Asn	0.8500	0.7611	0.1932	0.0632	0.0266

Asp	0.9875	0.7611	0.2078	0.0676	0.0156
Cys	0.7500	0.5537	0.2363	0.0644	0.0000
Gln	0.7556	0.6401	0.2052	0.0572	0.0229
Glu	0.8778	0.7012	0.2188	0.0648	0.0327
Gly	1.1250	0.7340	0.1236	0.0000	0.0000
His	0.8650	0.7658	0.1863	0.0470	0.0317
Hyp	0.9750	0.5448	0.1444	0.0575	0.0078
Ile	0.6250	0.4431	0.2003	0.1101	0.0298
Leu	0.6875	0.4153	0.1780	0.0755	0.0596
Lys	0.5667	0.3750	0.1573	0.0537	0.0296
Met	0.5625	0.4153	0.1773	0.0619	0.0388
Phe	0.4773	0.8778	0.1578	0.0513	0.0320
Pro	0.8000	0.5290	0.1623	0.0746	0.0000
Ser	1.0333	0.5713	0.2146	0.0185	0.0000
Thr	0.8857	0.5683	0.2103	0.0710	0.0000
Trp	0.4964	0.9359	0.1794	0.0603	0.0303
Tyr	0.6000	0.8787	0.1551	0.0607	0.0367
Val	0.7143	0.4746	0.2290	0.1103	0.0000

The calculated and experimental isoelectric points pI for the 21 amino acids are listed in Table 2.

Table 2. Calculated and experimental values of pH at isoelectric point pI for 21 amino acids (AA).

AA	pI (Eq. 10)	pI (Eq. 14)	Experiment
Ala	5.80	5.76	6.00
Arg	10.31	10.33	10.76
Asn	5.80	5.66	5.41
Asp	2.79	2.65	2.77
Cys	5.80	5.79	5.07
Gln	5.80	5.85	5.65
Glu	2.79	2.86	3.22
Gly	5.80	5.90	5.97
His	8.80	8.38	7.59
Hyp	5.80	5.81	5.80
Ile	5.80	5.96	6.02
Leu	5.80	5.99	5.98

Lys	8.80	9.28	9.74
Met	5.80	5.98	5.74
Phe	5.80	5.86	5.48
Pro	5.80	6.07	6.30
Ser	5.80	5.56	5.68
Thr	5.80	5.65	5.60
Trp	5.80	5.57	5.89
Tyr	5.80	5.58	5.66
Val	5.80	5.81	5.96

For the $\{G_k,J_k\}$ chosen databasis the following best linear model turns out to be:

$$\mathrm{p}I = 12.0 - 12.7J_1 - 22.5J_4 \quad n = 21 \quad r = 0.478 \quad s = 1.598 \quad F = 2.7 \tag{6}$$

$$\mathrm{MAPE} = 17.73\% \qquad \mathrm{AEV} = 0.7718$$

where the mean absolute percentage error (MAPE) is 17.73% and the approximation error variance (AEV) is 0.7718. The inclusion of N improves the correlation

$$\mathrm{p}I = 7.13 + 0.751N - 7.99J_1 - 15.7J_3 - 81.7J_4 \tag{7}$$

$$n = 21 \quad r = 0.629 \quad s = 1.499 \quad F = 2.6 \quad \mathrm{MAPE} = 16.95\% \quad \mathrm{AEV} = 0.6065$$

and AEV decreases by 21%. However, the model is limited to small N because N increases with both n_A and n_B, resulting inadequate for polypeptides and proteins.

The databasis $\{G_k,J_k,G_k^V,J_k^V\}$ improves the model:

$$\mathrm{p}I = 16.9 + 0.421G_2 - 22.2J_4 - 2.62J_1^V - 9.53J_2^V - 13.9J_3^V - 23.1J_4^V \tag{8}$$

$$n = 21 \quad r = 0.699 \quad s = 1.475 \quad F = 2.2 \quad \mathrm{MAPE} = 14.73\% \quad \mathrm{AEV} = 0.5465$$

and AEV decreases by 29%. The inclusion of N improves the correlation

$$\mathrm{p}I = -11.1 + 3.35N + 5.56G_3 - 18.2G_5 + 3.39J_2 - 66.6J_4 - 3.31G_2^V + 0.955G_4^V - 8.12J_1^V + 24.7J_2^V - 22.3J_3^V - 50.7J_4^V - 14.0J_5^V \tag{9}$$

$$n = 21 \quad r = 0.958 \quad s = 0.781 \quad F = 7.5 \quad \mathrm{MAPE} = 8.25\% \quad \mathrm{AEV} = 0.1754$$

and AEV decreases by 77%. However, the model is inadequate for proteins because N, G_3, G_5, G_2^V and G_4^V increase with n_A and n_B. The use of $(1+\Delta n/n_T) = 0.5$ for Arg, 4/3 for Asp and Glu, 2/3 for His and Lys, as well as one for all others improves the fit:

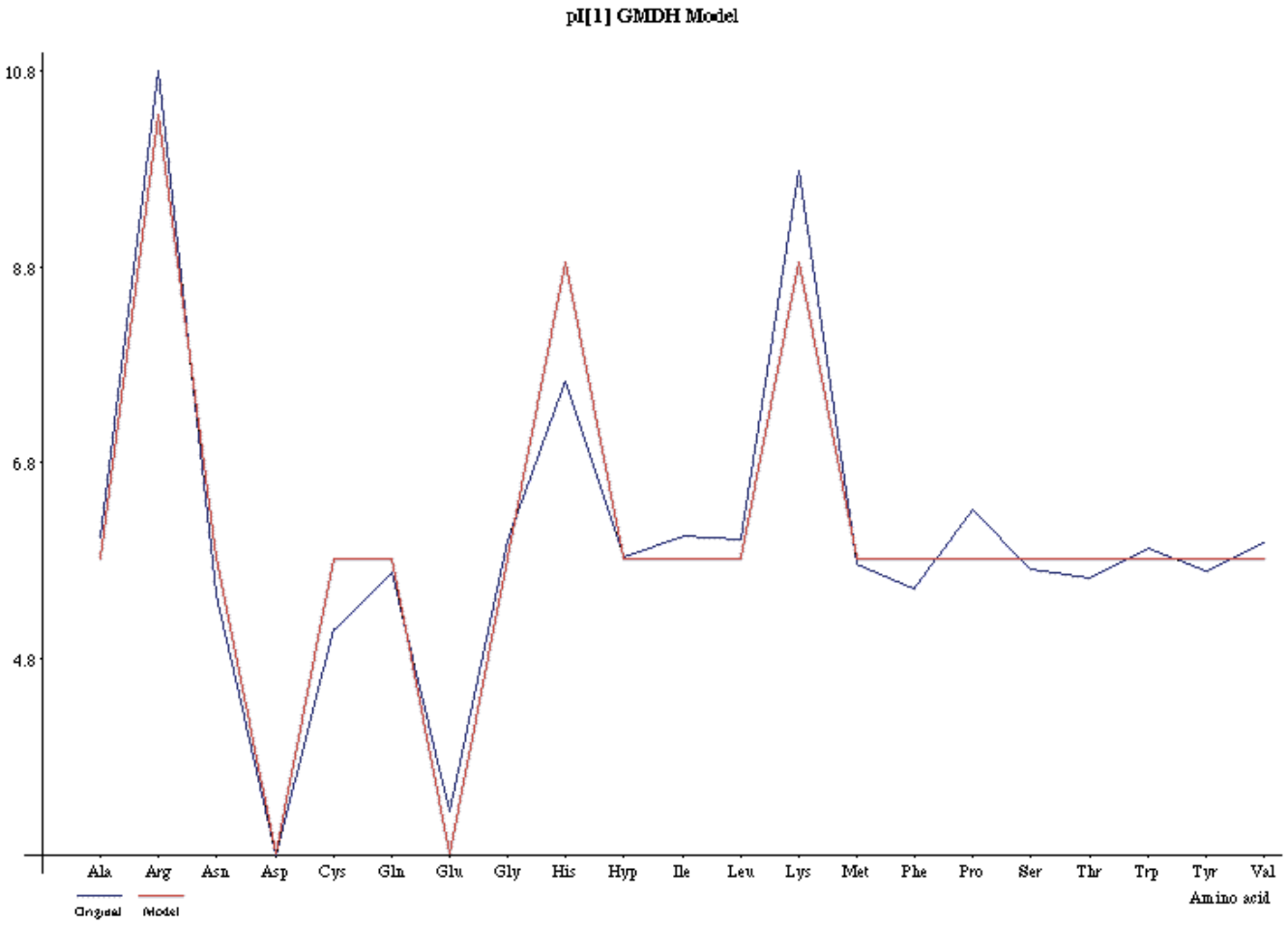
$$\mathrm{p}I = 14.8 - 9.01(1 + \Delta n/n_T) \quad n = 21 \quad r = 0.965 \quad s = 0.462 \quad F = 259.8 \tag{10}$$

$$\mathrm{MAPE} = 5.29\% \qquad \mathrm{AEV} = 0.0682$$

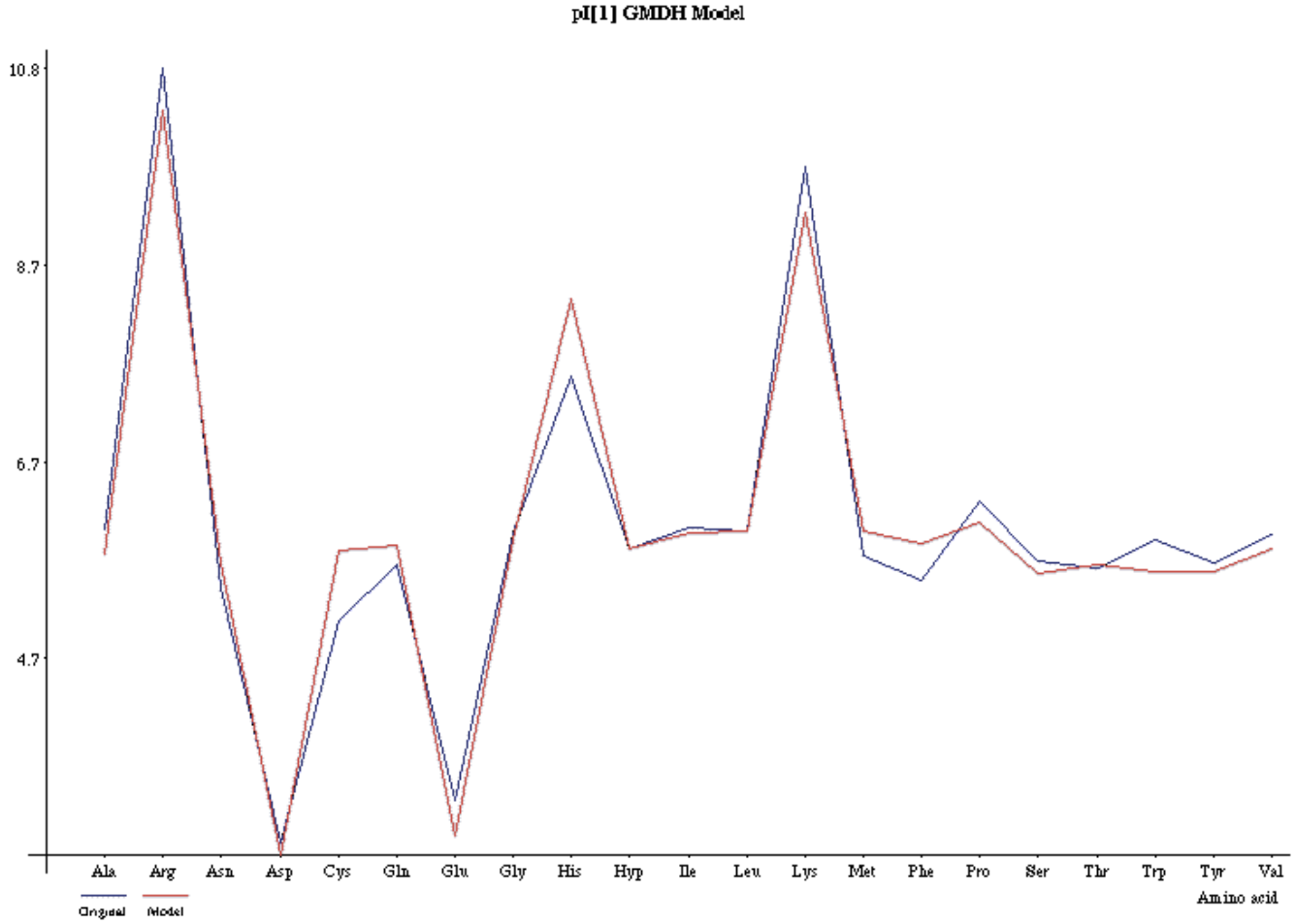
and AEV decreases by 91%. The correlation coefficient represents the 96.8% of that of the correlation of the means ($n = 4$, $r = 0.997$).

The pI isoelectric points (calculated with Equation 10) for the 21 amino acids are also included in Table 2. For Equation (10) the absolute relative errors results 5%. The pI isoelectric points

(calculated with Equation 10 and experimental) for the 21 amino acids are shown in Figure 1a. For Equation (10) the two amino acids farthest from the experimental value are His and Lys, with an absolute error of *ca*. 1.1 units.



(a)



(b)

Figure 1. Isoelectric points pI for the 21 amino acids: Equations (10) (a) and (14) (b).

The variation of the pI isoelectric point as a function of $(1+\Delta n/n_T)$ for the 21 amino acids (*cf.* Figure 2) shows that some amino acids appear superposed. The fitting line corresponds to the 21 amino acids; both amino acids that are the farthest are His and Lys ($n_B = 2$).

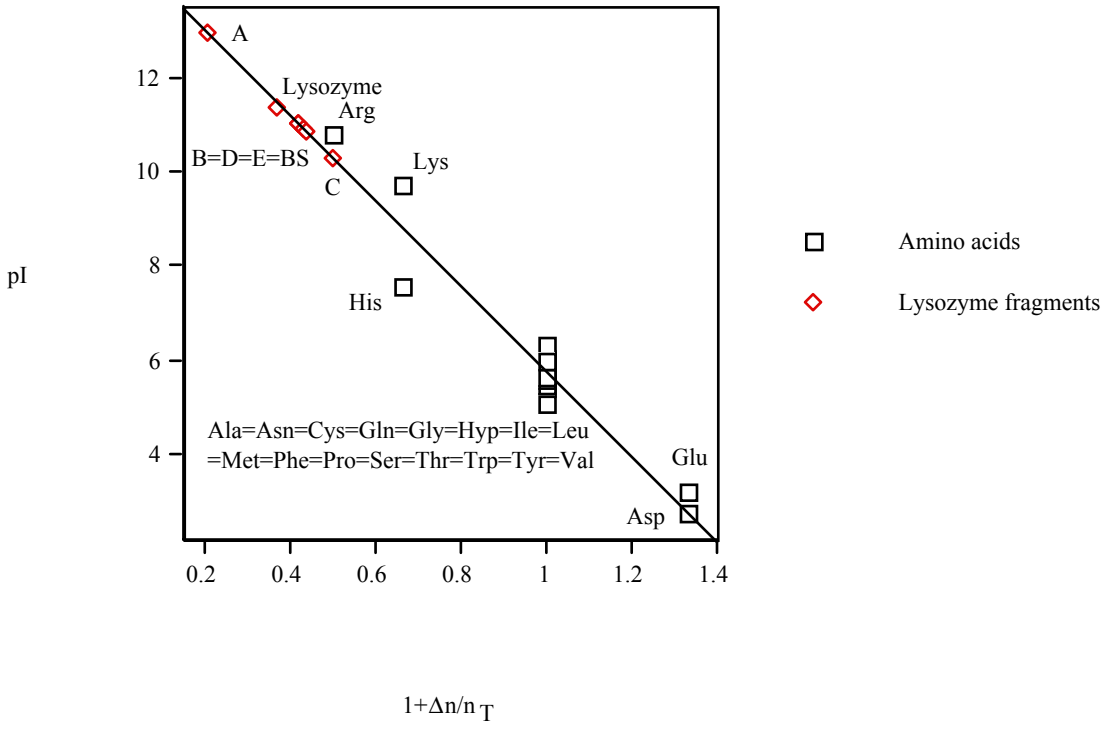


Figure 2. Variation of pI *vs*. $(1+\Delta n/n_T)$ for 21 amino acids, lysozyme and its fragments. The fitting line corresponds to the amino acids.

The inclusion of $\{G_k,J_k\}$ improves the fit:

$$\mathrm{p}I = 15.3 - 8.99(1+\Delta n/n_T) - 1.00J_2 \quad n = 21 \;\; r = 0.971 \qquad s = 0.435 \qquad F = 147.9 \qquad (11)$$

$$\mathrm{MAPE} = 5.10\% \qquad \mathrm{AEV} = 0.0573$$

and AEV decreases by 93%. The inclusion of $\{J_k^V\}$ improves the fit:

$$\mathrm{p}I = 16.8 - 8.59(1+\Delta n/n_T) - 0.958J_2 - 8.98J_3 - 1.16J_1^V \qquad (12)$$

$$n = 21 \;\; r = 0.977 \qquad s = 0.407 \qquad F = 85.7 \qquad \mathrm{MAPE} = 4.70\% \qquad \mathrm{AEV} = 0.0450$$

AEV decreases by 94% and allows studying polypeptides, proteins and protein fragments. The inclusion of $\{G_k^V\}$ improves the fit:

$$\mathrm{p}I = 16.0 - 8.94(1+\Delta n/n_T) - 0.828J_2 - 9.77J_3 + 0.619G_4^V \qquad (13)$$

$$n = 21 \;\; r = 0.981 \qquad s = 0.378 \qquad F = 100.1 \qquad \mathrm{MAPE} = 4.12\% \qquad \mathrm{AEV} = 0.0425$$

and AEV decreases by 94%. However, the model is inadequate for proteins because G_4^V increases with n_A and n_B. No superposition of the corresponding G_k–J_k or G_k^V–J_k^V pairs is observed in Equations (6–8,10–13), which decreases the risk of collinearity in the fits, given the close relationship between each pair G_k–J_k in Equation (4) [23,24].

The simultaneous inclusion of $\{G_k^V,J_k^V\}$ improves the fit:

$$\mathrm{p}I = 16.6 - 8.71(1+\Delta n/n_T) - 0.787J_2 - 9.52J_3 + 0.485G_4^V - 0.801J_1^V - 2.73J_4^V \qquad (14)$$

$$n = 21 \;\; r = 0.989 \qquad s = 0.306 \qquad F = 103.9 \qquad \mathrm{MAPE} = 4.20\% \qquad \mathrm{AEV} = 0.0380$$

and AEV decreases by 95%. However, the model is inadequate for proteins because G_4^V increases with n_A and n_B.

The pI isoelectric points (calculated with Equation 14) for the 21 amino acids are also included in Table 2. For Equation (14) the absolute relative error decreases to 4%. The pI isoelectric points (calculated with Equation 14 and experimental) for the 21 amino acids are displayed in Figure 1b. For Equation (14) the error is reduced for most amino acids; in particular for His and Lys the error decreases to 0.6 units.

The molecular CT indices are collected in Table 3 for lysozyme, five fragments of its tertiary structure and its binding site. In general, the CT indices do not distinguish α-helices, 3.0_{10}-helix, β-sheet and binding site. In particular both J_k and J_k^V indices for the whole molecule are similar to those for the α-helices and, specially, for α-helix D.

Table 3. Values of G_k and J_k charge-transfer indices up to fifth order for lysozyme and its fragments.

Fragment	N	G_1	G_2	G_3	G_4	G_5
α-Helix A	87	28.5000	39.3889	10.9097	6.4083	4.5069
α-Helix B	83	26.2500	45.5000	11.5417	6.5189	4.7258
3.0_{10}-Helix C	39	13.2500	14.2222	5.2500	3.2222	2.1181
α-Helix D	61	20.2500	24.6667	8.3750	5.1022	3.6111
β-Sheet E	104	39.0000	55.5556	13.8750	8.4000	6.4167
Binding site BS	105	30.2500	60.1111	13.5833	5.0556	2.5906
Lysozyme	1001	349.2500	541.7222	137.0833	85.9639	64.8828

Fragment	J_1	J_2	J_3	J_4	J_5
α-Helix A	0.3314	0.4580	0.1269	0.0745	0.0524
α-Helix B	0.3201	0.5549	0.1408	0.0795	0.0576
3.0_{10}-Helix C	0.3487	0.3743	0.1382	0.0848	0.0557
α-Helix D	0.3375	0.4111	0.1396	0.0850	0.0602
β-Sheet E	0.3786	0.5394	0.1347	0.0816	0.0623
Binding site BS	0.2909	0.5780	0.1306	0.0486	0.0249
Lysozyme	0.3493	0.5417	0.1371	0.0860	0.0649

Fragment	G_1^V	G_2^V	G_3^V	G_4^V	G_5^V
α-Helix A	56.5000	50.9778	19.8069	11.3040	7.0642

α-Helix B	51.3500	56.1472	19.0750	10.6837	6.8304
3.0_{10}-Helix C	27.9500	19.9000	9.2583	5.2535	3.2994
α-Helix D	41.5500	34.1083	14.6944	8.9512	5.1353
β-Sheet E	79.6000	74.1861	22.9611	13.2918	8.6063
Binding site BS	59.5500	70.9806	19.1278	7.0843	3.1818
Lysozyme	683.4500	691.8222	230.6111	139.8265	91.6138

Fragment	J_1^V	J_2^V	J_3^V	J_4^V	J_5^V
α-Helix A	0.6570	0.5928	0.2303	0.1314	0.0821
α-Helix B	0.6262	0.6847	0.2326	0.1303	0.0833
3.0_{10}-Helix C	0.7355	0.5237	0.2436	0.1382	0.0868
α-Helix D	0.6925	0.5685	0.2449	0.1492	0.0856
β-Sheet E	0.7728	0.7203	0.2229	0.1290	0.0836
Binding site BS	0.5726	0.6825	0.1839	0.0681	0.0306
Lysozyme	0.6835	0.6918	0.2306	0.1398	0.0916

The pI isoelectric points for lysozyme and its fragments not included in the fit are calculated by a modification of Equation (10):

$$\mathrm{p}I = 14.8 - 9.01(M + \Delta n)/n_T \qquad (15)$$

where M is the number of amino-acid residues in the protein or fragment. The choice seems sensible as pI values are strongly dependent on the type of side-chain functional groups.

The pI isoelectric points (calculated and experimental) for lysozyme and its fragments not included in the fit are reported in Table 4. The calculation result for α-helix A (M = 11 residues) is an estimate for that of the whole lysozyme (M = 129 residues) with a relative error of 13%. Furthermore, the inclusion of the other two α-helices (A+B+D, M = 31 residues) reduces the error to 1%.

Table 4. Values of the pH at the isoelectric point, pI for lysozyme fragments not included in the fit.

Fragment	Residues	pI	Experiment
α-Helix A	5–15	12.95	–
α-Helix B	24–34	10.89	–
3.0_{10}-Helix C	80–85	10.31	–
α-Helix D	88–96	11.02	–

Total α-helix	5–15,24–34,88–96	11.62	–
Total helix	5–15,24–34,80–85,88–96	11.29	–
β-Sheet E	41–54	10.87	–
Total helix+sheet	5–15,24–34,41–54,80–85,88–96	11.21	–
Binding site BS	34,35,37,44,57,59,62,63,101,107,114	11.00	–
Total helix+sheet+BS	5–15,24–34,35,37,41–54,57,59,62,63,80–85,88–96,101, 107,114	11.17	–
Lysozyme	1–129	11.49	11.35

The variation of the pI isoelectric point for lysozyme (experiment) and its fragments (calculation) as a function of $(N+\Delta n)/n_T$ (Figure 2) shows that some fragments appear superposed. Both lysozyme and its fragments lie in the fitting line obtained for the amino acids.

Experimental Procedures

The most important matrices that delineate the labelled chemical graph are the *adjacency* (**A**) [18] and *distance* (**D**) *matrices*, wherein $D_{ij} = \square_{ij}$ if $i = j$, "0" otherwise; $\square_{ij}$ is the shortest edge count between vertices i and j [19]. In **A**, $A_{ij} = 1$ if vertices i and j are adjacent, "0" otherwise. The $\mathbf{D}^{[-2]}$ matrix is that whose elements are the squares of the reciprocal distances D_{ij}^{-2}. The intermediate matrix **M** is defined as the matrix product of **A** by $\mathbf{D}^{[-2]}$:

$$\mathbf{M} = \mathbf{A}\mathbf{D}^{[-2]}$$

The *CT matrix* **C** is defined as $\mathbf{C} = \mathbf{M} - \mathbf{M}^{\mathrm{T}}$ where $\mathbf{M}^{\mathrm{T}}$ is the transpose of **M** [20]. By agreement $C_{ii} = M_{ii}$. For $i \neq j$, the C_{ij} terms represent a measure of the intramolecular *net charge* transferred from atom j to i. The *topological CT indices* G_k are described as the sum of absolute values of the C_{ij} terms defined for the vertices i,j placed at a topological distance D_{ij} equal to k:

$$G_k = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left| C_{ij} \right| \delta\left(k, D_{ij}\right) \tag{3}$$

where N is the number of vertices in the graph, D_{ij} are the entries of the **D** matrix, as well as δ is the Kronecker δ function being $\delta = 1$ for $i = j$ and $\delta = 0$ for $i \neq j$. The G_k represent the sum of all the C_{ij} terms, for every pair of vertices i and j at topological distance k. Other topological CT index, J_k, is defined as:

$$J_k = \frac{G_k}{N-1} \tag{4}$$

The index represents the mean value of CT for each edge, since the number of edges for acyclic compounds is $N - 1$.

When heteroatoms are present, some way of discriminating atoms of different kinds needs to be considered [21]. In valence CT-index terms, the presence of each heteroatom is taken into account by

introducing its electronegativity in the corresponding entry of the main diagonal of the adjacency matrix **A**. For each heteroatom X its entry A_{ii} is redefined as:

$$A_{ii}^{V} = 2.2(\chi_X - \chi_C) \quad (5)$$

to give the *valence adjacency* $\mathbf{A}^V$ *matrix*, where χ_X and χ_C are the electronegativities of heteroatom X and carbon, respectively, in Pauling units. The subtractive term keeps $A_{ii}^{V} = 0$ for the C atom, and the factor gives $A_{ii}^{V} = 2.2$ for O, which was taken as standard. From $\mathbf{A}^V$ instead of **A**, $\mathbf{M}^V$, $\mathbf{C}^V$, G_k^V and J_k^V are calculated following the former procedure. The C_{ii}^V, G_k^V and J_k^V are graph invariants.

The enzyme protein lysozyme (129 amino-acid residues, molecular weight 14307g·mol^{-1}) has been taken from the Protein Data Bank code 2LYM. The charge on lysozyme is +12.0e at pH 4.0, +8.0e at pH 7.0, +4.0e at pH 10.0 and decreases rapidly as the isoelectronic point at pH 11.35 is approached [22].

From the present results and discussion the following conclusions can be drawn.

1. The inclusion of heteroatoms in the π-electron system was beneficial for the description of the isoelecric point, owing to either the role of the additional p orbitals provided by the heteroatom or the role of steric factors in the π-electron conjugation.

2. The use of only charge-transfer and valence charge-transfer indices $\{G_k,J_k,G_k^V,J_k^V\}$ gave limited results for modelling the isoelectric point of amino acids. Furthermore, the inclusion of $(1+\Delta n/n_T)$ improved all the models. The effect is especially noticeable for those amino acids with more than two functional groups, *viz*. Arg, Asp, Glu, and, specially, His, and Lys. Moreover, the fractional index casts some light on the importance of the side-chain functional groups in the pI simulations of functional-rich molecules. The satisfactory modelling of the pI of 21 amino acids by the aid of a fractional index, based mainly on the Δn index, shows how to bypass the problem to derive and work with an extended set of charge-transfer indices (here, $m = 20$) as, in this case, a good description can be obtained with only one index.

3. The fitting line obtained for the 21 amino acids can be used to estimate the isoelectric point of lysozyme and its fragments, by only replacing $(1+\Delta n/n_T)$ with $(M+\Delta n)/n_T$.

4. For lysozyme, the results of smaller fragments can estimate that of the whole protein with 1–13% errors. An extension of the present study to other enzymes and proteins would give an insight into a possible generality of these conclusions, because most globular, water-soluble proteins are ionic, *e.g*., lysozyme (charge +8.0e) and bovine serum albumin (anionic) at pH 7.0. The present study may be also of interest in charge-migration peptide studies.

Work is in progress on the further elucidation of the value of Δn in the fractional indices for a better definition of indices, which are highly dependent on side-chain functional groups.

Acknowledgements

The authors acknowledge financial support from the Spanish MEC DGCyT (Project No. CTQ2004-07768-C02-01/BQU), EU (Program FEDER) and Generalitat Valenciana (DGEUI INF01-051, INFRA03-047 and OCYT GRUPOS03-173).

References

[1] Pogliani, L. Molecular Connectivity Model for Determination of Isoelectric Point of Amino Acids. *J. Pharm. Sci*. **1992**, *81*, 334-336.

[2] Pogliani, L. Molecular Connectivity: Treatment of the Electronic Structure of Amino Acids. *J. Pharm. Sci*. **1992**, *81*, 967-969.

[3] Pogliani, L. Modeling with Special Descriptors Derived from a Medium-Sized Set of Connectivity Indices. *J. Phys. Chem*. **1996**, *100*, 18065-18077.

[4] Pogliani, L. Modeling Biochemicals with Leading Molecular Connectivity Terms. *Med. Chem. Res*. **1997**, *7*, 380-394.

[5] Pogliani, L. Modeling Properties with Higher-Level Molecular Connectivity Descriptors. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 104-111.

[6] Pogliani, L. From Molecular Connectivity Indices to Semiempirical Connectivity Terms: Recent Trends in Graph Theoretical Descriptors. *Chem. Rev*. **2000**, *100*, 3827-3858.

[7] Stadtman, E. R. Oxidation of Free Amino Acids and Amino Acid Residues in Proteins by Radiolysis and by Metal-Catalyzed Reactions. *Annu. Rev. Biochem*. **1993**, *62*, 797-821.

[8] Berlett, B. S.; Stadtman, E. R. Protein Oxidation in Aging, Disease, and Oxidative Stress. *J. Biol. Chem*. **1997**, *272*, 20313-20316.

[9] Chu, I. K.; Rodriquez, C. F.; Lau, T.-C., Hopkinson, A. C.; Siu, K. W. M. Molecular Radical Cations of Oligopeptides. *J. Phys. Chem. B* **2000**, *104*, 3393-3397.

[10] Simon, S.; Gil, A.; Sodupe, M.; Bertrán, J. Structure and Fragmentation of Glycine, Alanine, Serine and Cysteine Radical Cations. A Theoretical Study. *J. Mol. Struct. (Theochem)* **2005**, *727*, 191-197.

[11] Gil, A.; Bertran, J.; Sodupe, M. Effects of Ionization on *N*-Glycylglycine Peptide: Influence of Intramolecular Hydrogen Bonds. *J. Chem. Phys*. **2006**, *124*, 154306–1-10.

[12] Torrens, F. A New Topological Index to Elucidate Apolar Hydrocarbons. *J. Comput.-Aided Mol. Design* **2001**, *15*, 709-719.

[13] Torrens, F. Valence Topological Charge-Transfer Indices for Dipole Moments. *J. Mol. Struct. (Theochem)* **2003**, *621*, 37-42.

[14] Torrens, F. Valence Topological Charge-Transfer Indices for Dipole Moments. *Mol. Diversity* **2004**, *8*, 365-370.

[15] Torrens, F. Valence Topological Charge-Transfer Indices for Reflecting Polarity: Correction for Heteromolecules. *Molecules* **2005**, *10*, 334-345.

[16] Torrens, F. Valence Topological Charge-Transfer Indices for Dipole Moments. *Molecules* **2003**, *8*, 169-185.

[17] Torrens, F. Valence Topological Charge-Transfer Indices for Dipole Moments: Percutaneous Enhancers. *Molecules* **2004**, *9*, 1222-1235.

[18] Randi•, M. The Analysis and Selection of Variables in Linear Regression. *J. Am. Chem. Soc*. **1975**, *97*, 6609-6615.

[19] Hosoya, H. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn*. **1971**, *44*, 2332-2339.

[20] Gálvez, J.; García, R.; Salabert, M. T.; Soler, R. Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci*. **1984**, *34*, 520-525.

[21] Kier, L. B.; Hall, L. H. Molecular Connectivity VII: Specific Treatment to Heteroatoms. *J. Pharm. Sci*. **1976**, *65*, 1806-1809.

[22] Bergers, J. J.; Vingerhoeds, M. H.; van Bloois, L.; Herron, J. N.; Janssen, L. H. M.; Fischer, M. J. E.; Crommelin, D. J. A. The Role of Protein Charge in Protein–Lipid Interactions. PH-Dependent Changes of the Electrophoretic Mobility of Liposomes through Adsorption of Water-Soluble, Globular Proteins. *Biochemistry* **1993**, *32*, 4641-4649.

[23] Box, G. E. P.; Hunter, W. G.; MacGregor, J. F.; Erjavec, J. Some Problems Associated with the Analysis of Multiresponse Data. *Technometrics* **1973**, *15*, 33-51.

[24] Hocking, R. R. The Analysis and Selection of Variables in Linear Regression. *Biometrics* **1976**, *32*, 1-49.