ELSEVIER

Biophysical Chemistry 115 (2005) 263 – 266

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

Empirical parametrization of pK values for carboxylic acids in proteins using a genetic algorithm

Raquel Godoy-Ruiz[a,b], Raul Perez-Jimenez[a,b], Maria M. Garcia-Mira[a,b], Isabel M. Plaza del Pino[a,b], Jose M. Sanchez-Ruiz[a,b,]*

[a]*Facultad de Ciencias, Departamento de Quimica Fisica, 18071-Granada, Spain*
[b]*Institute for Biocomputation and Physics of Complex Systems, E50009-Zaragoza, Spain*

Received 21 June 2004; received in revised form 29 October 2004; accepted 10 December 2004
Available online 23 December 2004

Abstract

Considerable effort has been devoted to the development of theoretical electrostatic methods to predict the pK values of ionizable residues in proteins. However, predictions appear often to be still at the qualitative or semi-quantitative level. We believe that, with the increasing number experimentally available pK values for proteins of known structure, an alternative approach becomes feasible: the empirical parametrization of the experimental protein pK database. Of course, in the long term, this empirical approach is no substitute for rigorous electrostatic analysis but, in the short term, it may prove to have useful predictive power and it may help to pinpoint the main structural determinants of pK values in proteins. Here we demonstrate the feasibility of the parametrization approach by fitting (using a genetic algorithm as fitting tool) the database for carboxylic acid pK values in proteins on the basis of an empirical equation that takes into account the two following kinds of effects: (1) long-range charge–charge interactions; (2) interactions of the given carboxylic acid group with its environment in the protein, which are described in terms of contributions from the different kind of atoms present in the protein (atomic contributions).

Keywords: Protein ionization; Carboxylic acids; Genetic algorithm

1. Introduction

Electrostatic interactions are likely to play essential roles in molecular processes involving proteins, including ligand-binding, protein–protein interactions, and protein folding–unfolding (see Ref. [1], and references quoted therein). Electrostatic interactions should be most clearly evident in the pK values for ionizable residues in proteins and, consequently, considerable effort has been devoted to the development of theoretical electrostatic methods to predict such pK values. However, predictions appear often to be still at the qualitative or semi-quantitative level (see, for instance, Refs. [2,3]).

We believe that, with the increasing number of experimentally available pK values for proteins of known structure [4], an alternative approach becomes feasible: the empirical parametrization of the protein pK database. In this kind of approach an empirical expression for the pK value is posed including parameters that describe plausible structural effects and fitting parameters whose values are determined from the fitting of the empirical expression to the set of experimental pK values. Certainly, in the long term, this empirical approach is no substitute for rigorous electrostatic analysis but, in the short term, it may prove to have useful predictive power and it may help to determine the main structural determinants of pK values in proteins.

Here, we explore the feasibility of the empirical parametrization approach. To do so, we fit (using a genetic

* Corresponding author. Facultad de Ciencias, Departamento de Quimica Fisica, Campus Fuentenueva s/n, 18071-Granada, Spain. Tel.: +34 958243189; fax: +34 958272879.
E-mail address: sanchezr@ugr.es (J.M. Sanchez-Ruiz).

0301-4622/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2004.12.028

algorithm as fitting tool) the recently compiled database for carboxylic acid pK values in proteins [4] on the basis of an empirical equation that takes into account the two following kinds of interactions: (1) charge–charge interactions; these are modeled using the Tanford–Kirkwood model, as we have previously described [5]; (2) interactions of the given carboxylic acid group with its environment in the protein, which is described in terms of atomic terms; that is, the number of atoms of a given type (for instance, aliphatic carbons or hydroxyl oxygens) within a certain distance of the COO^- group.

2. Theory, results, and discussion

2.1. The pK database

We use the data base of 212 pK values for caboxylic acid groups in proteins compiled by Andrew Robertson and coworkers [4]. These values correspond to 24 proteins of known structure and had been determined by NMR. From this set, we selected the 121 pK values corresponding to the 16 proteins of structure known by X-ray crystallography (that is, the pK values of the 8 proteins of structure known only by NMR were not included). Out of these 121 pK values, 63 correspond to aspartate residues and 58 to glutamate residues.

2.2. Parametrization of pK values in terms of atomic contributions

The following equation was used to describe the pK values in terms of atomic contributions:

$$\mathrm{p}K = pK_{TK} + a_0 + a_1 \cdot p_1 + a_2 \cdot p_2 + a_3 \cdot p_3 + a_4 \cdot p_4 + \ldots \quad (1)$$

The meaning of the symbols in Eq. (1) is as follows:

- pK_{TK} is the pK value calculated taking only into account the effect of charge–charge interactions. pK_{TK} values are obtained, from the structure of the protein, using the Tanford–Kirkwood model with the Gurd correction, as we have previously described in detail [5]. Note that pK_{TK} values are not fitting parameters.
- p_1, p_2, p_3,... are atomic parameters that are meant to describe the environment of each given carboxylic acid group in the protein. Each of these parameters is simply the number of atoms of a given kind within a sphere of radius r centered at the carbon of the COO^- moiety. Actually, we considered 14 different types of atoms: (1) aliphatic carbons (i.e., carbons from Ala, Val, Leu, and Pro); (2) oxygen atoms of –OH groups (Ser, Thr, and Tyr); (3) nitrogen atoms of $-NH_2$ groups in Asn and Gln; (4) sulphur atoms of Cys and Met; (5) oxygen atoms of $-COO^-$ groups in Glu and Asp; (6) nitrogen NZ of Lys; (7) nitrogen atoms in –NH groups of Arg; (8) nitrogen atoms of His sidechains; (9) carbon atoms of aromatic sidechains (Phe, Tyr, Trp); (10) backbone carbons; (11) backbone oxygens; (12) backbone nitrogens; (13) carbonylic carbons of Asn and Gln sidechains; (14) carbonylic oxygens of Asn and Gln. For each of these 14 types of atoms we considered 6 spheres of radii 4, 5, 6, 7, 8, and 9 Å. This makes a total of 84 p_i parameters, the values of which being calculated, for each carboxylic acid residue, from the protein structure.
- a_0, a_1, a_2, a_3,... are the fitting parameters whose values are to be determined according to a least-squares criterion (see below).

2.3. A genetic algorithm as a fitting tool

The above approach leads to a total of 85 fitting parameters. Clearly, our purpose is not to fit a database of 121 pK values with an equation including 85 parameters, but to achieve an acceptable description of that database on the basis of a small subset of the fitting parameters. In order to find such subset, we use a genetic algorithm similar to that we have previously employed in the design of stabilizing surface charges in proteins [6]. Briefly, each set of values of the fitting parameters, $\{a_0,\ a_1,\ a_2,\ldots\}$, is referred to as a chromosome and, for any given chromosome, the standard deviation (σ) associated to the difference between the experimental and predicted (Eq. (1)) pK values can be easily calculated. We start with a population of several chromosomes randomly chosen, which are ranked according to the following score:

$$Z = \sigma + \delta \left[\frac{(N_{\mathrm{N}} - N_{\mathrm{C}})}{5} \right]^4 \quad (2)$$

where N_{N} is the number of a_i parameters different from zero in the chromosome and N_{C} is the desired (or target) value for that number. The best chromosome (lower value of Z) is passed unmodified to the next generation which is completed with “children” chromosomes generated by crossover from couples of “parent” chromosomes randomly selected from the original population. The new population is subjected to low-probability mutation (random changes in the values of the parameters a_i), Z scores are calculated and a new cycle starts with the new population. This process is continued until the best chromosome has remained unchanged for a specified number of cycles.

It must be noted that the fourth power term in Eq. (2) guarantees that, for reasonable values of δ, chromosomes with N_{N} values clearly different than the target N_{C} value are strongly penalized. Therefore, using different values for the desired number of a_i parameters different from zero (N_{C}) allows the genetic algorithm to target subsets of different size of the full fitting parameters set.

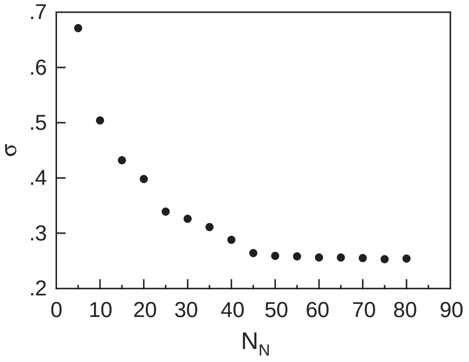

Fig. 1. Plot of standard deviation associated to the difference between the experimental and predicted (Eq. (1)) pK values versus the number of non-zero a_i coefficients in Eq. (1). The values shown correspond to the best chromosomes obtained using a genetic algorithm as a fitting tool (see text for details).

2.4. Results of the fitting process

Fig. 1 shows a plot of σ (the standard deviation associated to the difference between the experimental and predicted pK values) versus N_N (the number of non-zero a_i coefficients) for the best chromosomes derived from the application of the genetic algorithm to the fitting of Eq. (1) to experimental pK database. Clearly, for high values of N_N, the σ value is very low, indicating a very good fit; however, good fits achieved with high numbers of fitting parameters are likely to have little physical meaning. On the other hand, low values of N_N lead to poor fittings (high values of σ). The results shown in Figs. 2 and 3 correspond to the best chromosome obtained with N_N=15 (i.e., 15 non-zero fitting parameters), which appears to be a reasonable compromise between the high-N_N and low-N_N situations. The fitting obtained with N_N=15 appears to be a significant improvement over the predictions of the Tanford–Kirkwood

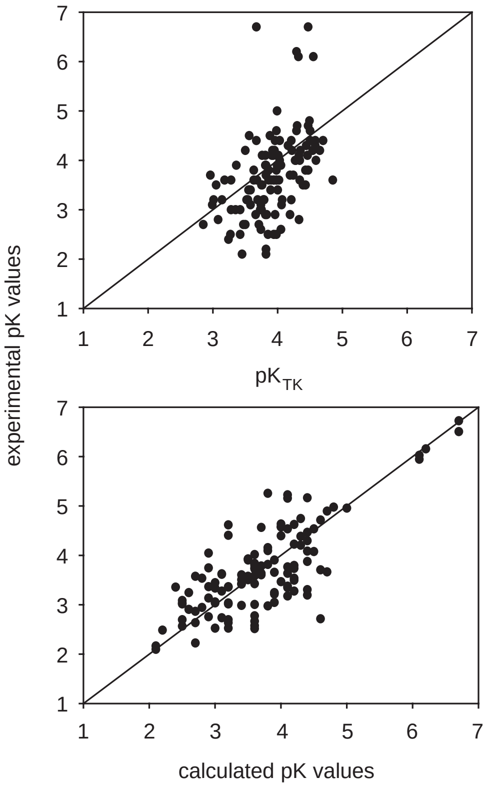

Fig. 2. Upper panel: plot of experimental pK values for carboxylic acids in several proteins versus the values calculated using the Tanford–Kirkwood model with the correction of Gurd. Lower panel: plot of experimental pK values for carboxylic acids in several proteins versus the values calculated from Eq. (1) using the a_i values corresponding to the best chromosome with 15 non-zero a_i coefficients.

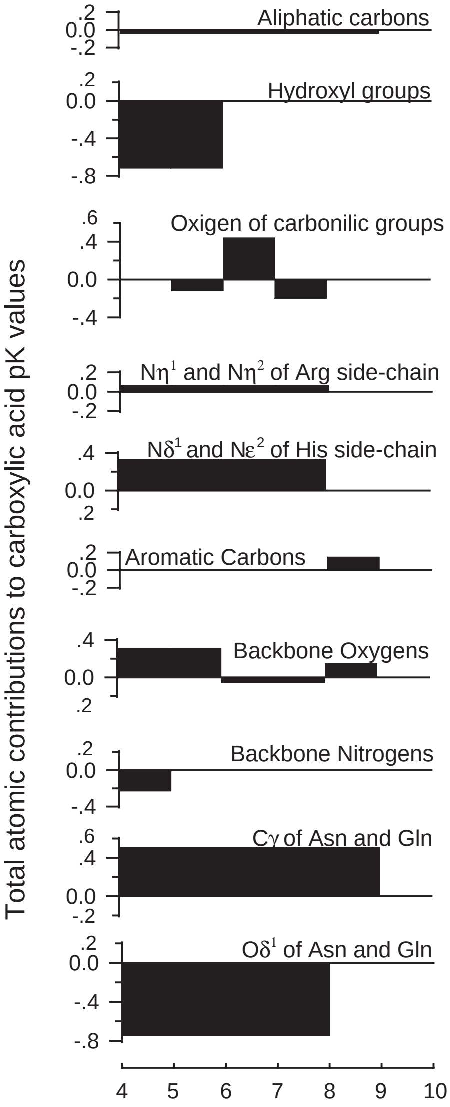

Fig. 3. Total atomic contributions to carboxylic acid pK values as derived from the a_i values corresponding to the best chromosome with 15 non-zero a_i coefficients. For a given atom type (see text for details) at a distance d from the carbon of the COO^- group, the total atomic contribution is given by $\sum_{R>d} a_{i(R)}$, where $a_i(R)$ is the a_i coefficient for the atom type and a sphere of radius R (see the text for details) and the sum over all the spheres of radius larger than d.

model (Fig. 2) and the values obtained for the fitting parameters (Fig. 3) seem to provide a comparatively simple picture of the environment effect on pK values, although we must recognize that the values found for the contributions of aliphatic carbons and backbone oxygens appear puzzling.

Overall, the results we report here support the feasibility of an empirical parametrization of carboxylic acid pK values in terms of atomic contributions. A definitive conclusion about the usefulness of the approach will have to wait, however, until a larger database of experimental pK values is available.

Acknowledgements

This work was supported by Grant BIO2003-02229 from the Spanish Ministry of Science and Technology and by FEDER funds. Raul Perez-Jimenez is a recipient of a predoctoral fellowship from the Spanish Ministry of Science and Technology.

References

[1] R. Perez-Jimenez, R. Godoy-Ruiz, B. Ibarra-Molero, J.M. Sanchez-Ruiz, The efficiency of salts to screen charge-interactions in proteins: a Hofmeister effect? Biophys. J. 86 (2004) 2414–2429.

[2] M. Sundd, N. Iverson, B. Ibarra-Molero, J.M. Sanchez-Ruiz, A.D. Robertson, Electrostatic interactions in ubiquitin: stabilization of carboxylates by lysine amino groups, Biochemsitry 41 (2002) 7586–7596.

[3] M.F. Garcia-Mayoral, J.M. Perez-Cañadillas, J. Santero, B. Ibarra-Molero, J.M. Sanchez-Ruiz, J. Lacadena, A. Martinez del Pozo, J.G. Gavilanes, M. Rico, M. Bruix, Dissecting structural and electrostatic interactions of charged groups in α-sarcin. An NMR study of some mutants involving the catalytic residues, Biochemistry 42 (2003) 13122–13133.

[4] W.R. Forsyth, J.M. Antosiewicz, A.D. Robertson, Empirical relationships between protein structure and carboxyl pK_a values in proteins, Proteins: Struct., Funct., Genet. 48 (2002) 388–403.

[5] B. Ibarra-Molero, V.V. Loladze, G.I. Makhatadze, J.M. Sanchez-Ruiz, Thermal versus guanidine-induced unfolding of ubiquitin. Analysis in terms of the contributions from charge–charge interactions to protein stability, Biochemistry 38 (1999) 8138–8149.

[6] B. Ibarra-Molero, J.M. Sanchez-Ruiz, Genetic algorithm to design stabilizing surface-charge distributions in proteins, J. Phys. Chem., B 106 (2002) 6609–6613.