Published in final edited form as:

J Biomol Struct Dyn. 2018 September ; 36(12): 3094–3105. doi:10.1080/07391102.2017.1377636.

A comprehensive Analysis of the Computed Tautomer Fractions of the Imidazole Ring of Histidines in *Loligo vulgaris*

Yury N. Vorobjev^{1,2,3}, Harold A. Scheraga³, and Jorge A. Vila⁴

¹Institute of Chemical Biology and Fundamental Medicine of the Siberian Branch of the Russian Academy of Science, Lavrentiev Avenue 8, Novosibirsk 630090, Russia

²Novosibirsk State University, Novosibirsk 630090, Russia

³Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301

⁴IMASL-CONICET, Universidad Nacional de San Luis, Ejército de Los Andes 950, 5700 San Luis, Argentina

Abstract

A recently introduced electrostatic-based method to determine the pKa values of ionizable residues and fractions of ionized and tautomeric forms of histidine (His) and acid residues in proteins, at a given fixed pH, is applied here to the analysis of a His-rich protein, namely *Loligo vulgaris* (pdb id 1E1A), a 314-residue all- β protein. The average tautomeric fractions for the imidazole ring of each of the six histidines in the sequence were computed by using an approach that includes, but is not limited to, molecular dynamic simulations coupled with calculations of the ionization states for all 94 ionizable residues of protein 1E1A in water at pH 6.5 and 300 K. The electrostatic-calculated tautomeric fractions of the imidazole ring of His were compared with predictions obtained from an existent NMR-based methodology. Our results indicate that: (i) the averaged electrostatic-based tautomeric predictions for the imidazole ring of *all* histidines of *Loligo vulgaris* are dominated by the N^{e2}-H rather than the N⁶¹-H form, although such preferences from the NMR-based methodology are not so well defined; (ii) the computed average absolute difference between the electrostatic- and the NMR-based tautomeric predictions among *all* six histidines vary among 0% to 17%; (iii) for the His showing the largest fraction of the neutral form (81%), the absolute difference between the NMR- and electrostatic-based computed tautomeric predictions is only 3%; and (iv) the tautomeric predictions for the imidazole ring of His computed with the NMR-based methodology are stable within a certain, well-defined, range of variations of a tautomer-related parameter.

Keywords

Histidine; Tautomers; Proteins; Molecular Dynamics; NMR

Corresponding Author: jv84@cornell.edu.

Author Contributions

- Y.N.V. and J.A.V. designed and performed research
- Y.N.V., H.A.S. and J.A.V. analyzed data and wrote the paper.

Introduction

Histidine (His) is a unique residue for a number of reasons: *i)* about 50% of all enzymes use His in their active sites (Shimba et al., 2003); *ii)* His, with a pKa of 6.6 at T = 300 K (Demchuk & Wade, 1996), titrates around neutral pH, enabling the deprotonated nitrogen of its imidazole ring to serve as an effective ligand for metal binding; *iii)* the imidazole ring of His (see Figure 1) includes two neutral, chemically-distinct forms, and an ionized form, referred to as N^{δ1}-H and N^{ε2}-H tautomers, and H⁺, respectively; one form is favored over the others by, mainly, the local protein environment and the pH. In particular, it has been suggested that tautomerization of the neutral state and variations of χ^1 of His are crucial parts of the proton-transfer process (Hass et al., 2008).

Development of theoretically accurate methods to account for the fractions of His forms, during the computation of the total free energy, is a long-standing problem in structural biology. The seminal work of Bashford et al (1993) is a standard reference of the pioneer effort to solve the problem of an accurate computation of the pKa of the ionizable residues in proteins by including consideration of the fractions of His forms during the calculations. Although our present work is not intended to be a review of the field, we can say, in general, that there is a large number of theoretical works aimed to compute the pKa's in proteins, accurately (Bashford & Karplus, 1990; Beroza et al., 1991; Yang et al., 1993; Yan & Honig, 1993; Ripoll et al., 1996; Song et al., 2009; Nielsen et al., 2011). In particular, some of these developments were designed to compute the ionization of histidine, and their tautomeric fractions and acid/base residues in proteins, accurately from *only* their 3D structures (Machuqueiro & Baptista, 2009; Witham et al., 2011; Couch & Stuchebrukhov 2011; Anandakrishn, Aguilar & Onufriev, 2012). In this regard, a contribution to solve this problem was introduced recently by Vorobjev, Scheraga & Vila (2017). The method is based on a Molecular Dynamic (MD) simulation, to account for the protein conformational dynamics, together with an effective version of a continuum electrostatic method, for an accurate prediction of the pKa's of protein residues in solution as a function of pH. In particular, the computation of the pKa was validated by using a large set of 297 experimental pKa's from 34 X-ray and NMR-determined 3D structures of proteins.

An accurate theoretical prediction of the forms of His in proteins is certainly essential for a detailed understanding of the protein's functions (Hass et al., 2008). Besides, an accurate experimental determination of the fractions of the His forms is equally important since it will enable us, among other things, to validate such theoretical predictions. However, the latter is an elusive problem in NMR spectroscopy. Actually, there are several reasons as to why this problem is yet unsolved. Among others, we can mention the following. Direct observation of the ¹⁵N chemical shifts of the imidazole ring of His (Pelton et al., 1993) is a common practice, in NMR spectroscopy, to determine the tautomeric fractions of His in proteins. However, there is evidence suggesting that there is a large uncertainty in such predictions (Vila, 2012), depending on the adopted "canonical" limiting values of the ¹⁵N chemical shift of the imidazole ring of His in which each form of His is present to the extent of 100%. A similar concern arises when the one-bond C-H, ¹J_{CH}, spin-spin coupling constants (SSCC) of the imidazole ring of His is intended to be used to determine the tautomeric forms of His (Vila & Scheraga, 2017), i.e., the results are very sensitive to the

adopted “high-pH limiting value” of the $^1J_{\text{CH}}$ SSCC. Finally, an accurate method to calculate the fractions of the tautomeric forms of His as a function of pH, provided that the observed $^{13}\text{C}^\gamma$ and $^{13}\text{C}^{\delta 2}$ chemical shifts and the protein structure, or the fraction of the H^+ form, are known, has recently been developed (Vila et al., 2011). However, in NMR spectroscopy, it is not a common practice to determine the $^{13}\text{C}^\gamma$ chemical shift of the imidazole istidine. In fact, only 109 $^{13}\text{C}^\gamma$, versus 4,813 $^{13}\text{C}^{\delta 2}$, chemical shifts of the imidazole ring of histidine have been deposited in the Biological Magnetic Resonance data Bank (BMRB) (Ulrich et al., 2005). Consequently, the sparse numbers of $^{13}\text{C}^\gamma$ chemical shifts prevent a wide application of this quantum-chemical-based methodology. Despite this, and for the reasons given below, we will use this NMR-based methodology to validate theoretical predictions of the tautomeric fractions of the imidazole ring of His in *Loligo vulgaris* (pdb id 1E1A), a 314-residue all- β protein.

Overall, there is a number of reasons as to why the use of *Loligo vulgaris* is a good test model with which to study the accuracy of the electrostatic-calculated tautomeric fractions of the imidazole ring of each of the histidine residues at pH=6.5 (Vorobjev, Scheraga & Vila, 2017). Among others, these are: (i) this protein, to the best of our knowledge, possesses the largest number of His residues (six) for which the $^{13}\text{C}^\gamma$ and $^{13}\text{C}^{\delta 2}$ chemical shifts have been determined experimentally by NMR spectroscopy and, hence, for which the tautomeric fractions of the imidazole ring of His can be validated by using an NMR-based methodology (Vila et al., 2011); (ii) the *Loligo vulgaris* structure has been accurately determined by X-ray crystallography at 1.8 Å resolution (pdb id 1E1A) [Scharff et al., 2001]; and (iii) this protein possesses a large number (94) of ionizable residues and, hence, our work describes how the *ionized* environment may affect the tautomeric distribution of the imidazole ring of His at a given fixed pH.

Materials and Methods

Molecular dynamics simulation at constant pH: Calculation of the fraction of tautomers and charged form of His

For details to compute the energies of ionization microstates, the free energy of ionization, and an extensive test of the methodology on 34 proteins, see Vorobjev, Scheraga & Vila, (2017).

Protein dynamics in water, at a given pH=6.5, was taken into account by constant-pH MD simulation (Vorobjev, 2012) of protein 1E1A. A preliminary calculation of pKa's of all ionizable residues for energy-optimized protein structure have shown (see Table 1) that pKa's of all six histidine residues of 1E1A are in the range of 6.0–7.5, and all acid residues have pKa < 4.7 (Glu127). Therefore, for further constant-pH simulation it is reasonable to consider *all* acid (Asp, Glu) and base (Lys, Arg) residues in the ionized state, because their respective pK_O's (3.5, 4.0, and 10.5, 12.5, respectively) are shifted by more than 2.5 pK units from the pH (6.5) at which the calculations were carried out. In other words, it is reasonable to maintain *all* acid and base residues as ionized during the calculations. For the same reason, *all* tyrosine residues are considered as neutral, e.g., the minimal computed pKa (9.1) occurs for Tyr144. However, His residue pKa's vary considerably around pH 6.5 at which the calculations are carried out (see Table 1) and hence, consideration of the

ionization of the His forms must be considered explicitly. For this reason, the six histidine residues in the sequence of protein 1E1 A, namely His181, His219, His224, His248, His274 and His287, were considered as having variable ionization states. In order to reduce the computational cost considered for all possible ionization microstates, we determined, first, the distances between all pairs of His. As shown in Table 2, there are *only* three pairs, namely, H219-H224, H224-H248, H224-H287, showing a His-His distance $\approx 13\text{\AA}$. This information, together with the analysis of the pKa of histidine residues, enables us to conclude that *only* three histidine residues are electrostatically coupled in the vicinity of pH=6.5, namely, H219, H224 and H248. Consequently, we will focus on the computation of *only* 27 ionization microstates (namely 3^3 , because there are only three electrostatically coupled histidine residues and three forms for each His).

The ionization partition function for pH=6.5 is calculated as described previously (Vorobjev, Scheraga & Vila, 2017) for a subset of six ionizable His residues, while the ionization states of *all* acid and base (Lys, Arg) residues are taken as ionized. Consequently, the ionization partition function is given by Eq.(1):

$$Z_{inz}(\mathbf{x}, pH) = \sum_{k=1}^{N_z} \exp[-G_{inz}(\mathbf{x}, \mathbf{z}_k, pH)/kT] \quad (1)$$

Where $Z_{inz}(\mathbf{x}, pH)$ is the ionization partition function over all ionized/neutral tautomer states of histidine residues for protein conformation $\mathbf{x}$; $G_{inz}(\mathbf{x}, \mathbf{z}_k, pH)$ is the average free energy of microscopic ionization state $\mathbf{z}_k$ for instant conformation $\mathbf{x}$ at a given pH; $N_z=27$ is the number of ionization microstates. The macroscopic ionization state m of a histidine residue i , is given by Eq.(2)

$$\langle \delta(z_{m,i}) \langle \mathbf{x} \rangle, pH \rangle = \frac{1}{N_{snap}} \sum_{j=1}^{N_{snap}} \langle \delta_{FAMBE}(z_{m,i})(\mathbf{x}_j, pH) \rangle \quad (2)$$

where $\langle \mathbf{x} \rangle$ can be considered as a macroscopic protein structure as an average over a set of instant structures; $m=0, 1$ and 2 are the ionization states of each histidine residue i , viz., ionized and two neutral tautomers; and $\delta_{FAMBE}(\mathbf{z}_{m,i})(\mathbf{x}_j, pH)$ is the occupation number calculated by the Fambe_pH method (Vorobjev & Scheraga, 1997; Vorobjev, Vila & Scheraga, 2008) for the instant protein conformation $\mathbf{x}_j$ at a given pH for a set of snapshots (N_{snap}) along an MD trajectory. The average free energies and occupation numbers are calculated for each of the 27 ionization microstate of the three coupled histidine residues at pH=6.5. In particular, as shown in Table 3, only 5 out of 27 ionization microstates, namely, 001000, 002000, 011000, 012000, 020100, are the most populated ones, i.e., accounting for $\sim 99\%$ of *all* ionization states of the coupled histidine residues. The averaged macroscopically observed ionization states of histidine residues calculated via Eq.(2) are shown in Table 4.

Simulations are carried out using the following protocol: (i) during the MD simulation the conformational dynamics of the protein was taken into account by using the BioPASED program (Popov & Vorobjev, 2010) with the AMBER force field (Cornell et. al., 1995); (ii) the system is heated slowly from 1K to 300K during 250 ps; (iii) a final equilibration at 300K is carried out during ~0.5 ns; (iv) the equilibrium MD trajectories of the protein dynamics, during 10 ns, was collected as snapshots taken every 25ps time-interval; and (v) calculations of the average ionization degrees and tautomer fractions of the imidazole ring of all histidine residues, for each snapshot were carried out by using the well-tested Fambe_pH method (Vorobjev, Vila & Scheraga, 2008), as described in the Material and Methods section of Vorobjev, Scheraga & Vila (2017).

An NMR-based methodology to estimate the tautomeric fraction of His

As far we are aware, there is no experimental method with which to determine, beyond doubts, the protonation *and* the tautomeric fractions of the imidazole ring of His of proteins in solution. For this reason, we will use here a recently proposed NMR-based theoretical approach (Vila et al., 2011). In this approach, the average tautomeric fraction of the N^{e2}-H form of His (f^e) can be determined by using the following equation:

$$f^e = \frac{\Delta^{obs}(1 - \langle f^{H+} \rangle)}{\Delta^e} \quad (3)$$

where $\langle f^{H+} \rangle$ is the average fraction of the ionized form *computed* by using a continuum electrostatic based method (Vorobjev, Scheraga & Vila, 2017); and $\Delta^{obs} = |^{13}C^{\delta 2} - ^{13}C^{\gamma}|$, where $^{13}C^{\delta 2}$ and $^{13}C^{\gamma}$ are the NMR-observed chemical shifts for each His residue in the sequence (see Δ^{obs} values in Table 5); Δ^e is the first-order absolute shielding difference between the $^{13}C^{\delta 2}$ and $^{13}C^{\gamma}$ nuclei for the N^{e2}-H tautomer ($|^{13}C^{\delta 2} - ^{13}C^{\gamma}|^e$), a parameter which must be estimated. As a first approximation, we will use two highly-probable Δ^e values, namely 31.7 ppm, obtained from DFT-based calculations (Vila et al., 2011) and 22.4 ppm, as the NMR-observed value, by Sudmeir *et al.* (2003), at pH 12, and -55 °C in 80% d6-ethanol/20% H₂O, for the N^{e2}-H tautomer. It is worth noting that the small Δ^e value (22.4 ppm), observed by Sudmeier *et al.* (2003), is within ~2 standard deviations (*sd*) from the large mean Δ^e value (31.7 ppm), obtained from the average DFT-computed shielding distribution for each of ~35,000 conformations of histidine in a model tripeptide (Vila et al., 2011); and f^e is the computed, (Eq. 3), fraction of the N^{e2}-H tautomer. Naturally, the f^{δ} fraction, *viz.*, for the N^{δ1}-H tautomer, is obtained straightforwardly as:

$$f^{\delta} = 1 - \langle f^{H+} \rangle - f^e \quad (4)$$

The computed tautomeric fractions with Eqs. (3) and (4) by using the proposed values for Δ^e will be called, from here on, the *reference* values. Implicit in the theoretical approach leading to these *reference* values is that $\Delta^e = \Delta^{obs}$; otherwise, the method does not apply since it will give negative (unphysical) tautomeric fractions. Despite this consideration, it is worth noting that the tautomeric fractions determined by using this NMR-based methodology

depends, critically, on the average fraction of the ionized form of the His residue ($\langle f^{H^+} \rangle$). For self-consistency, we will compute $\langle f^{H^+} \rangle$ by using the new methodology (Vorobjev, Scheraga & Vila, 2017) rather than using the $\langle f^{H^+} \rangle$ values previously reported for protein 1E1A (Vila et al., 2011). The main reason for this choice is that the new methodology, but not the previous one (Vorobjev, Vila & Scheraga, 2008), takes into account explicitly the protein dynamics which is crucial for an accurate estimation of the average fraction of the ionized His form ($\langle f^{H^+} \rangle$) (Vorobjev, Scheraga & Vila, 2017).

Results and discussion

Analysis of the tautomeric distribution of the imidazole ring

The protein 1E1A (Scharff et al., 2001) contains six His residues for which the fractions of the tautomeric forms of each His can be determined by using two methodologies, namely an NMR- and electrostatic-based methodology, respectively.

Calculations of the degree of ionization and the fraction of two neutral histidine tautomers were carried out at pH = 6.5. A representative set of conformations of 1E1A was obtained by carrying out MD simulations of 10ns, for each of the 27 electrostatic-coupled ionization states of the six histidine residues. The fractions of ionized histidine and its tautomers were estimated explicitly by calculating average energies of 27 ionization states of protein 1E1A in water. It is worth noting that all acid and Lys, Arg residues are taken as ionized with Tyr in the neutral state at pH= 6.5, because their pKa's in 1E1A are shifted by more than 2.5 pK units, as shown in Table 1.

As shown in Figure 2 the range of pKa values along the MD trajectory, for each of the six HIS residues of the protein 1E1A, fluctuate within ~ 0.14 – 0.21 pK units around the average values (see Table 5). Moreover, the variation, along the 10 ns MD trajectories, of the fractions of ionized His (f^{H^+}) and tautomeric forms (f^{δ} and f^{ϵ}) for each of the six His residues, at pH = 6.5 and T = 300 K, are shown in Figure 3a-3f. In particular, a considerable variation of the computed fraction occurs for the N^{e2}-H tautomer of His219 (see green-solid line in Figure 3b). The fact that a large variation of the N^{e2}-H tautomeric fraction follows the variation of the fraction of the ionization degree (see black-solid line in Figure 3b) may indicate that the proton binding/release occurs preferentially at the N^{δ1} nucleus. In addition, the large variation of the computed fractions for His219 (see Figure 3b) could be a simple consequence of the fact that this residue is well exposed to the solvent, (see Table 5) and, hence, exhibits a large fluctuation of its surface area. Over all, the distribution of the forms for the imidazole ring varies significantly among different histidines in the protein (see Table 5 and Figure 3), reflecting the importance of the environment of the As shown in Table 5, the NMR-based tautomeric predictions of the *references* values show good qualitative agreement among themselves (80%) except for His219 for which the N^{e2}-H rather than N^{δ1}-H tautomer is preferred for one of the *references*. As to whether this result is related to the computed large variation of the tautomeric fractions mentioned above for His219, remains to be proven. Despite this, the tautomeric predictions of the NMR-based method are stable within a certain, well-defined, range of variations for f^{ϵ} .

The agreement of the electrostatic-based prediction with the *reference* values is variable, e.g., very good agreement is seen for His287 but not for His248 where the opposite occurs, in terms of tautomer preference. Overall, the best agreement between the electrostatic- and the NMR-based prediction, respectively, by using a large ϵ (31.7ppm) or small ϵ (22.4ppm) occurs for His224 and His287 (with averaged absolute differences of 3% and respectively) and for His219 and His274 (with average absolute differences of 11% and 4%, respectively).

There is only one His for which $\delta_{\text{obs}}^{\text{N}^{\text{H}}-\text{H}} (25.9\text{ppm}) > \delta^{\text{N}^{\text{H}}-\text{H}} (22.4\text{ppm})$, namely for His287, leading to an unphysical negative tautomeric fraction. The existence of $\delta_{\text{obs}}^{\text{N}^{\text{H}}-\text{H}} > \delta^{\text{N}^{\text{H}}-\text{H}}$ may perhaps histidines in determining the tautomeric forms. indicate that the NMR-observed $\delta^{\text{N}^{\text{H}}-\text{H}}$ value (22.4ppm) at -50°C , of Sudmeier et al. (2003), may not reflect the conformational chemical-shift dispersion that should occur, for the $\text{N}^{\text{H}}-\text{H}$ tautomer at higher temperature. Not surprising, the best (100%) agreement between the electrostatic-based tautomeric prediction and the *reference* values was obtain for His287 by using a large (31.7ppm) ϵ value.

The $\text{N}^{\text{H}}-\text{H}$ form of His224 is the most populated ($\sim 50\%$) electrostatic-computed tautomer at $\text{pH} = 6.5$, in good agreement with the tautomeric preferences from both *reference* values (see Table 5). Moreover, the fact that His224 shows the smaller molecular surface area is consistent with the smaller fraction of the H^+ form (19%) seen among *all six* His residues. Certainly, the dominant neutral state of His224 (81%) and low ionized fraction (19%) keep the N^{H} nucleus protonated, i.e., in the $\text{N}^{\text{H}}-\text{H}$ form, while the N^{H} nucleus becomes mostly non-protonated (see Figure 3c).

Finally, the electrostatic-base-computed tautomeric preference among all six His are overwhelming dominated (100%) by the $\text{N}^{\text{H}}-\text{H}$ tautomer. However, the predictions from the NMR-based model are not as well defined (see Table 5). Indeed, the $\text{N}^{\text{H}}-\text{H}$ tautomeric preferences are 50% and 60% respectively, for a large and small ϵ value.

Conclusions

Calculations of the fractions of the ionized H^+ and the tautomeric $\text{N}^{\text{H}}-\text{H}$ and $\text{N}^{\text{H}}-\text{H}$ forms of the imidazole ring of His in proteins is a challenging task, because it needs an accurate estimation of the relative free energies of the tautomer forms, within a fraction of 1.0 kcal. In particular, the fractions of the tautomeric forms of the imidazole ring of His in the sequence are accurately predicted within an average absolute difference of 9% to 12%, depending on the adopted values for the ϵ parameter. While this accuracy of the method is encouraging, more improvements are necessary to determine the tautomeric forms of the imidazole ring of His with higher accuracy.

The results obtained by using two highly-probable values for the ϵ parameter, in the NMR-based methodology, enables us to suggest that an averaged ϵ value, namely 27ppm, should be used in future applications of this methodology.

Overall, the only necessary requirement for an accurate prediction of the populations of the His forms in proteins, by using the coupled molecular dynamic and electrostatic-based methods presented here, is the 3D structure. This advantage of the method could be valuable

tool when the determination of the $^{13}\text{C}^\gamma$ and $^{13}\text{C}^{62}$ chemical shifts by NMR spectroscopy is not possible or does not exist.

Acknowledgements

This research was supported by the Russian Foundation for Basic Research [# 15-04- 00387-a], Russian Academy of Sciences [Program 'Molecular and Cell Biology, #6.11], MD modeling and pK calculations of a test set of proteins were supported by the Russian Science Foundation [#16-14-10038] (YNV); U.S. National Institutes of Health [grant number GM-14312], U.S. National Science Foundation [grant number MCB10-19767] (HAS); CONICET-Argentina (PIP-0087), UNSL-Argentina (PROICO 3-2212), ANPCyT-Argentina (PICT-0556, PICT-0767) [JAV].

References

- Anandakrishn R; Aguilar B; and Onufriev AV (2012). H++3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research*, 40, W537–W541. DOI:10.1093/nar/gks375 [PubMed: 22570416]
- Bashford D; Case DA; Dalvit C; Tennant L; Wright PE (1993). Electrostatic calculations of side-chain pKa values in myoglobin and comparison with NMR data for histidines. *Biochemistry*, 32, 8045–8056. DOI: 10.1021/bi00082a027 [PubMed: 8347606]
- Bashford D; Karplus M (1990). pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry*, 29, 10219–10225. DOI: 10.1021/bi00496a010, [PubMed: 2271649]
- Beroza P; Fredkin DR; Okamura MY; Feher G (1991). Protonation of interacting residues in a protein by a Monte Carlo method: Application to lysozyme and the photosynthetic reaction center of *Rhodobacter sphaeroides*. *Proceedings of National Academy of Sciences U.S.A* 88, 5804–5808. DOI:10.1073/pnas.88.13.5804
- Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc* 117:5179–5197. DOI: 10.1021/ja00124a002
- Couch V, Stuchebrukhov A (2011). His in continuum electrostatics protonation state calculations. *Proteins*. 79:3410–3419. DOI: 10.1002/prot.23114 [PubMed: 22072521]
- Demchuk E, & Wade RC (1996). Improving the continuum dielectric approach to calculating pKas of ionizable groups in proteins. *Journal of Physical Chemistry*, 100, 17373–17387. DOI:10.1021/jp960111d
- Hass MAS, Hansen DF, Christensen HEM, Led JJ, Kay LE (2008). Characterization conformational exchange of a histidine side chain: protonation, rotamerization, and tautomerization of His61 plastocyanin from *Anabaena variabilis*. *J Am Chem Soc* 130:8460–8470. DOI: 10.1021/ja801330h [PubMed: 18540585]
- Machuqueiro M, & Baptista AM (2009). Molecular dynamics at constant pH and reduction potential: Application to cytochrome c3. *Journal of the American Chemical Society*, 131, 12586–12594. DOI: 10.1021/ja808463e [PubMed: 19685871]
- Nielsen JE, Gunner MR, & Garcia-Moreno BE (2011). The pKa cooperative: A collaborative effort to advance structure-based calculations of pKa values and electrostatic effects in proteins. *Proteins: Structure, Function, and Bioinformatics*, 79, 3249–3259. DOI 10.1002/prot.23194
- Pelton JG, Torchia DA, Meadow ND, Roseman S (1993) Tautomeric states of the active-site histidines of phosphorylated and unphosphorylated III^{Glc}, a signal-transducing protein from *Escherichia coli*, using two-dimensional heteronuclear NMR techniques. *Prot. Sci* 2:543–558.
- Popov AV, & Vorob'ev YN (2010). GUI-BioPASED: A program for molecular dynamics simulations of biopolymers with a graphical user interface. *Molecular Biology*, 44, 648–654. DOI: 10.1134/S0026893310040217
- Ripoll DR, Vorobjev YN, Liwo A, Vila J, Scheraga HA (1996). Coupling Between Folding and Ionization Equilibria. Effects of pH on the Conformational Preferences of Polypeptides. *Journal of Molecular Biology* 264, 770–783. [PubMed: 8980685]

- Scharff ET., Koepke J, Fritzsche G, Lücke C, Rüterjans H (2001). Crystal structure of diisopropylfluorophosphatase from *Loligo vulgaris*. *Structure*. 9:493–502. PII S0969–2216(01)00601–4 [PubMed: 11435114]
- Shimba N, Serber Z, Ledwidge R, Miller SM, Craik CS, Dötsch V (2003). Quantitative identification of the protonation state of histidine in vitro and in vivo. *Biochemistry*. 42:1021–1034 [PubMed: 12735798]
- Song W, Mao J, & Gunner MR (2009). MCCE2: Improving protein pK calculations with tensile side chain rotamers sampling. *Journal of Computational Chemistry*, 30, 2231–2247. DOI:10.1002/jcc.21222 [PubMed: 19274707]
- Sudmeier JL; Bradshaw EM; Haddad EC; Day RM; Thalhauser CJ; Bullock PA; Bachovchin WW (2003). Identification of histidine tautomers in proteins by 2D $^1\text{H}/^{13}\text{C}^{\delta 2}$ one-bond correlated NMR. *J. Am. Chem. Soc* 125, 8430–8431. [PubMed: 12848537]
- Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao H, Markley JL (2008). BioMagResBank Nucleic Acids Res. 36:D402–D408. DOI:https://doi.org/10.1093/nar/gkm957 [PubMed: 17984079]
- Vila JA (2012). Limiting Values of the ^{15}N chemical shifts of the imidazole ring of histidine at high pH. *J. Phys. Chem. B* 116:6665–6669. [PubMed: 22376024]
- Vila JA, Arnautova YA, Vorobjev YN, & Scheraga HA (2011). Assessing the fractions of tautomeric forms of the imidazole ring of histidine in proteins as a function of pH. *Proceedings of the National Academy of Sciences*, 108, 5602–5607. DOI:10.1073/pnas.1102373108
- Vila JA, Scheraga HA (2017) Limiting values of the one-bond C-H spin-spin coupling constants of the imidazole ring of histidine at high-pH. *J. Mol. Struct* 1134:576–581.
- Vorobjev YN (2012). Potential of mean force of water-proton bath and molecular dynamic simulation of proteins at constant pH. *Journal of Computational Chemistry*, 33, 832–842. DOI:10.1002/jcc.22909 [PubMed: 22278814]
- Vorobjev YN, & Hermans J (1997). SIMS, computation of a smooth invariant molecular surface. *Biophysical Journal*, 73, 722–732. DOI:10.1016/S0006-3495(97)78105-0 [PubMed: 9251789]
- Vorobjev YN, & Scheraga HA (1997). A fast adaptive multigrid boundary element method for macromolecular electrostatic computations in a solvent. *Journal of Computational Chemistry*, 18, 569–583. DOI:10.1002/(SICI)1096-987X(199703)18:4<569::AID-JCC10>3.0.CO;2-B
- Vorobjev YN, Vila JA, & Scheraga HA (2008). FAMBEPH: A fast and accurate method to compute the total solvation free energies of proteins. *The Journal of Physical Chemistry B*, 112:11122–11136. DOI: 10.1021/jp709969n [PubMed: 18683966]
- Vorobjev YN, Scheraga HA, Vila JA. (2017). Coupled molecular dynamic and continuum electrostatic to compute the ionization of protein as function of pH. *J Biomol Struct Dyn*, 35:Feb.24, p.1–14. DOI:10.1080/07391102.2017.1288169 [PubMed: 26727155]
- Witham S, Talley K, Wang L, Zhang Z, Sarkar S, Gao D, Alexov E (2011). Developing hybrid approaches to predict pKa values of ionizable groups. *Proteins: Structure, Function, and Bioinformatics*, 79, 3389–3399. DOI:10.1002/prot.23097
- Yang SA, & Honig B (1993). On the pH dependence of protein stability. *Journal of Molecular Biology*, 231, 459–474. DOI:10.1006/jmbi.1993.1294 [PubMed: 8510157]
- Yang AS, Gunner MR, Sampogna R, Sharp K, Honig B (1993). On the calculation of pKa's in proteins. *Proteins*. 15:252–265. DOI: 10.1002/prot.340150304 [PubMed: 7681210]

Figure 1.

(a) Structure of the $N^{\delta 1}$ -H form of the His residue; (b) same as (a) for the $N^{\epsilon 2}$ -H form; and (c) same as (a) for the ionized H^+ form.

Figure 2.
pKa variation of *all* His residues, of protein 1E1A, along the MD trajectory.

figure 3.

(a) Fractions of the HIS 181 forms along the MD trajectory: *black* line for the H⁺ form; *red* and *green* line for the N^{δ1}-H and N^{ε2}-H tautomer, respectively; (b) same as (a) for HIS219; (c) same as (a) for HIS224; (d) same as (a) for HIS248; (e) same as (a) for HIS274; and (f) same as (a) for HIS287.

Table 1.

pKa of residues of protein 1E1A.^a

Residue		pK _a	pK ₀	Residue		pK _a	pK ₀
NEND	3	7.47	7.50	ASP	168	2.669	4.000
GLU	8	3.250	4.400	ARG	180	12.552	12.000
LYS	12	11.580	10.400	HIS	181	7.152	6.600
GLU	15	3.982	4.400	ASP	184	2.340	4.000
ASP	16	3.677	4.000	ARG	186	12.873	12.000
GLU	21	-1.815	4.400	TYR	188	9.306	9.600
ASP	26	1.970	4.000	GLU	194	2.797	4.400
LYS	27	11.068	10.400	LYS	198	10.854	10.400
ASP	30	3.375	4.000	LYS	199	10.810	10.400
TYR	32	9.428	9.600	TYR	203	9.9039	600
GLU	37	3.647	4.400	ASP	204	3.799	4.000
GLU	39	3.681	4.400	LYS	206	11.360	10.400
LYS	43	11.207	10.400	LYS	210	10.999	10.400
GLU	47	3.403	4.400	GLU	212	3.380	4.400
ARG	50	13.449	12.000	LYS	214	11.684	10.400
ASP	52	3.001	4.000	LYS	215	11.121	10.400
LYS	54	10.893	10.400	HIS	219	6.437	6.600
LYS	57	10.736	10.400	HIS	224	6.085	6.600
LYS	58	10.689	10.400	GLU	225	3.727	4.400
LYS	63	11.297	10.400	ASP	229	0.803	4.000
GLU	65	3.582	4.400	ASP	232	3.185	4.000
TYR	69	9.166	9.600	ASP	234	1.546	4.000
ASP	79	3.412	4.000	GLU	235	4.046	4.400
ARG	80	13.254	12.000	ASP	236	3.483	4.000
ASP	81	1.762	4.000	HIS	248	7.147	6.600
ASP	89	2.545	4.000	GLU	250	2.267	4.400
ARG	91	12.590	12.000	ASP	255	3.295	4.000
ASP	100	3.482	4.000	LYS	260	11.534	10.400
GLU	104	4.113	4.400	ARG	262	13.801	12.000
GLU	105	3.868	4.400	ARG	264	13.837	12.000
LYS	108	11.304	10.400	GLU	268	3.470	4.400
LYS	109	10.400	10.400	LYS	269	11.211	10.400
ASP	110	3.200	54.000	HIS	274	7.358	6.600
GLU	112	4.231	4.400	LYS	276	11.241	10.400
ARG	114	13.021	12.000	LYS	280	11.352	10.400
ARG	115	12.834	12.000	GLU	286	1.492	4.400
ASP	121	2.528	4.000	HIS	287	7.334	6.600
ASP	125	1.853	4.000	GLU	288	4.113	4.400
TYR	126	9.722	9.600	LYS	294	11.171	10.400

Residue		pK_a	pK₀	Residue		pK_a	pK₀
GLU	127	4.732	4.400	GLU	296	4.053	4.400
GLU	138	2.448	4.400	ARG	299	14.152	12.000
ASP	143	3.770	4.000	LYS	302	10.981	10.400
TYR	144	9.151	9.600	LYS	303	11.379	10.400
ARG	146	12.892	12.000	TYR	305	10.136	9.600
GLU	150	2.627	4.400	GLU	307	3.269	4.400
LYS	151	11.756	10.400	LYS	310	11.608	10.400
TYR	156	9.964	9.600	CEND314		3.793	3.800
ASP	161	3.190	4.000				

^aComputed from the energy optimized structure by 250 cycles of a conjugate gradient method; His residues are highlighted in bold face.

Table 2.

Average distances (Å) between pairs of nitrogen atoms ($N^{\delta 1}$, $N^{\epsilon 2}$) for each pair of His residues of IEIA.^a

	H181	H219	H224	H248	H274	H287
H181	—	24.2	26.8	28.3	13.4	24.6
H219	24.2	—	12.8	18.2	24.9	22.5
H224	26.8	12.8	—	8.9	21.1	12.6
H248	28.3	18.2	8.9	—	21.0	16.5
H274	13.4	24.9	21.1	21.0	—	14.1
H287	24.5	22.5	12.6	16.5	14.1	—

^aThose pairs of His residues, for which the distance is lower than 13Å, are shown in bold face.

Table 3.

Average energies of ionization microstates of 1E1A at pH=6.5 obtained by MD simulation.

Ionization State ^a	$\langle E_{\text{mol}} \rangle^b$ (rmsd) ^c	$\langle G_{\text{Solv}} \rangle^d$ (rmsd) ^c	$\langle G_{\text{total}} \rangle^e$ (rmsd) ^c	Occupation ^f
000000	1567.01 (68.81)	-1655.38 (41.24)	-88.37 (54.86)	0.0
001000	1562.71 (66.57)	-1672.40 (37.90)	-103.31 (51.12)	0.268
002000	1588.21 (69.54)	-1681.98 (36.65)	-102.76 (53.15)	0.107
				0.375
010000	1577.38 (61.65)	-1677.97 (20.58)	-93.77 (43.25)	0.0
011000	1588.78 (67.19)	-1695.42 (35.93)	-103.63 (53.61)	0.456
012000	1576.65 (61.92)	-1678.83 (36.98)	-102.18 (51.37)	0.041
				0.871
020000	1568.25 (68.84)	-1670.89 (45.82)	-101.64 (57.35)	0.017
021000	1558.42 (61.71)	-1659.83 (40.33)	-100.38 (52.82)	0.002
022000	1558.42 (61.71)	-1659.94 (40.33)	-99.41 (51.71)	0.0
				0.890
000100	1578.12 (59.81)	-1665.29 (41.76)	-87.17 (54.31)	0.0
001100	1582.46 (61.94)	-1673.88 (36.62)	-91.42 (57.38)	0.0
002100	1587.21 (69.54)	-1679.68 (36.65)	-92.47 (54.75)	0.0
010100	1558.64 (66.21)	-1648.46 (38.15)	-89.82 (57.49)	0.0
011100	1579.28 (61.65)	-1671.95 (20.58)	-92.67 (44.25)	0.0
012100	1584.39 (62.45)	-1677.25 (20.58)	-92.81 (43.71)	0.0
020100	1570.06 (60.59)	-1674.67 (33.05)	-102.75 (46.42)	0.105
021100	1596.77 (67.41)	-1690.67 (33.05)	-93.90 (47.13)	0.0
022100	1573.79 (62.05)	-1666.68 (51.68)	-92.89 (56.86)	0.0
				0.996
000200	1589.16 (60.31)	-1673.62 (42.48)	-84.46 (55.32)	0.0
001200	1585.63 (62.04)	-1674.55 (36.41)	-88.92 (56.78)	0.0
002200	1589.11 (69.54)	-1678.03 (35.95)	-91.18 (54.39)	0.0
010200	1559.75 (65.11)	-1646.47 (37.61)	-86.72 (57.49)	0.0
011200	1574.21 (61.15)	-1664.08 (20.58)	-89.87 (43.31)	0.0
012200	1579.47 (63.38)	-1669.68 (21.13)	-90.21 (42.92)	0.0
020200	1568.26 (60.17)	-1669.14 (33.15)	-100.87 (47.91)	0.005
021200	1592.91 (67.83)	-1685.72 (33.05)	-92.81 (47.16)	0.0
022200	1572.27 (62.37)	-1664.85 (51.68)	-92.38 (56.86)	0.0
				1.00 ^g

^a ionization state **0**, **1**, and **2** designate the histidine ionization state of H⁺, N^{δ1}-H and N^{ε2}-H forms, respectively, for each of 6 histidine residues H181, H219, H224, H248, H274, H287 of protein 1E1A. In bold face the *five* ionization states, that account for ~99% of *all* ionization states of the coupled histidine residues, are highlight.

^b average free-energy of the protein in a given microstate;

c rmsd fluctuations along the MD trajectory;

d energy of atom-atom and protein-solvent interactions (Vorobjev, Scheraga & Vila, 2017)

e total potential energy in water.

f statistical weight of the microstate or occupation number, computed with Eq. (2)

g accumulated occupation fraction of states.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4.

pKa values and tautomeric fractions of the imidazole ring of His computed from five ionization microstates

His	Ionization microstate for MD ^a					Equil. mixture ⁱ
	001000 ^b 0.27 ^c	002000 ^b 0.11 ^c	011000 ^b 0.46 ^c	012000 ^b 0.04 ^c	020100 ^b 0.11 ^c	
His181 ^h						
	6.91 ^d	7.02	7.08	7.07	7.10	7.03
	0.67 ^e	0.74	0.77	0.76	0.07	<u>0.75</u>
	0.10 ^f	0.08	0.08	0.08	0.07	<u>0.08</u>
	0.23 ^g					
His219 ^h		6.64	6.42	6.65	6.66	6.55
	6.67	0.56	0.43	0.56	0.56	<u>0.49</u>
	0.54	0.06	0.07	0.10	0.09	<u>0.07</u>
	0.05	0.38	0.50	0.34	0.36	<u>0.44</u>
His224 ^h		5.87	5.79	5.89	6.17	5.83
	5.75	0.19	0.17	0.19	0.30	<u>0.19</u>
	0.16	0.08	0.39	0.07	0.05	<u>0.29</u>
	0.35	0.73	0.44	0.74	0.65	<u>0.52</u>
	0.49					
His248 ^h		7.23	7.10	7.22	7.13	7.16
	7.22	0.82	0.77	0.82	0.77	<u>0.79</u>
	0.82	0.05	0.05	0.04	0.06	<u>0.05</u>
	0.04	0.13	0.18	0.14	0.17	<u>0.16</u>
	0.14					7.48
His274 ^h		7.52	7.40	7.38	7.58	<u>0.88</u>
	7.56	0.90	0.87	0.86	0.91	<u>0.04</u>
	0.91	0.04	0.03	0.04	0.03	<u>0.08</u>
	0.04	0.06	0.10	0.10	0.06	
	0.05					7.31
His287 ^h		7.47	7.25	7.52	7.56	<u>0.84</u>
	7.56	0.88	0.82	0.89	0.90	<u>0.03</u>
	0.83	0.02	0.03	0.03	0.03	<u>0.13</u>
	0.04	0.10	0.15	0.08	0.07	
	0.13					

^a computed over 10 ns MD trajectory, as explained in Materials and Methods section.^b computed by Eq.(2); only five ionization states accounting for ~99% among *all* 27 ionization states of the coupled histidine, are listed (see main text for details);^c statistical weight of ionization state (see Occupation column in Table 3);

d average pKa of His over the MD trajectory of protein 1E1A for each listed microstate;

e average fraction of the ionized form of histidine (H^+) for each listed microstate;

f average fraction of the $N^{\delta 1}$ -H form for each listed microstate;

g average fraction of the $N^{\epsilon 2}$ -H form for each listed microstate;

h histidine residue number of 1E1A protein;

i average over equilibrium MD trajectories for each ionization microstates; the computed average fraction for the imidazole ring of His, namely for the H^+ , $N^{\delta 1}$ -H and $N^{\epsilon 2}$ -H forms used in Table 5, are underlined.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5.

pKa values and tautomeric fractions of the imidazole ring of His, computed at pH 6.5 and T = 300 K, for protein 1E1 A.

His ^a	pKa computed values ^b	Fractions of the imidazole ring of His forms at pH = 6.5	
	X-ray <pK ^a > _{MD} crystal	Reference Values ^c	Computed values <fraction> ^e RMSD ^f Abs() ^g <MD>
His181 (11.2 Å ²)	6.37.03 (0.14)	<u>0.90</u> ^d 0.75 0.24 (0.24) 0.01 (0.01) N ^{δ1} -H (N ^{δ1} -H)	0.75 0.05 n/a 0.08 0.03 0.17 0.05 N ^{e2} -H 0.16(0.16) ^h
His219 (46.1 Å ²)	7.0 6.55 (0.12)	<u>14.3</u> ^d 0.49 0.28 (0.18) 0.23 (0.33) N ^{δ1} -H (N ^{e2} -H)	0.49 0.09 n/a 0.07 0.02 0.44 0.08 N ^{e2} -H 0.21(0.11)
His224 (5.1 Å ²)	5.9 5.83 (0.21)	0.19 0.32 (0.13) 0.49 (0.68) N ^{e2} -H (N ^{e2} -H)	0.19 0.08 n/a 0.29 0.04 0.52 0.08 N ^{e2} -H 0.03 (0.17)
His248 (20.1 Å ²)	6.9 7.16 (0.13)	<u>3.70</u> ^d 0.79 0.19 (0.18) 0.02 (0.03) N ^{δ1} -H (N ^{δ1} -H)	0.79 0.05 n/a 0.05 0.02 0.16 0.04 N ^{e2} -H 0.14 (0.13)
His274 (17.2 Å ²)	7.2 7.48 (0.16)	<u>22.4</u> ^d 0.88 0.03 (0.00) 0.09 (0.12) N ^{e2} -H (N ^{e2} -H)	0.88 0.05 n/a 0.04 0.02 0.08 0.04 N ^{e2} -H 0.01 (0.04)
His287 (16.7 Å ²)	7.0 7.31 (0.14)	<u>25.9</u> ^d 0.84 0.03 (−0.02) 0.13 (0.18) N ^{e2} -H (n/a)	0.84 0.05 n/a 0.03 0.00 0.13 0.00 N ^{e2} -H 0.00 (n/a)
Averaged Difference	—	—	0.09 (0.12) ⁱ

^a list of His residues in protein 1E1 A; in parenthesis, for each His residue, the average smooth molecular surface area, computed by using the SIMS method (Vorobjev & Hermans, 1997) are listed.

^b pKa values computed by using MC method (Vorobjev, Scheraga & Vila, 2017). In parentheses, we listed the estimated statistical errors of the MD/MC pKa calculations obtained by using multiple MC, for each pH value, with pH step equal to 0.25 (see Vorobjev, Scheraga & Vila, 2017); always the estimated statistical error was computed as the pKa RMSD along the MD 10.5 ns trajectory, and the maximal RMSD obtained was 0.21 pH units for His224, Figure 2.

^c for each of the three rows of every His in the first column we listed the corresponding *reference* values for the His forms, namely for the H⁺, N^{δ1}-H and N^{e2}-H form, respectively; with the average fraction of the H⁺ computed as described below in footnote (*e*) and the tautomeric form values computed by using Eq.(3) and Eq.(4). There are 2 reference values listed, the first one was computed by using $\epsilon = 31.5$ ppm and the second one (in parenthesis) by using 22.4 ppm (see main text for details). The last row, for each His residue, indicate the NMR-based tautomeric preference for each of the two references.

^d for every His we listed, underlined, the ϵ_{obs} value from Table SI of Vila et al. (2011).

^e in the first three rows the averaged fractions (see underlined values of Table 4) computed over 10 ns MD trajectory, by using Vorobjev et al (2017) methodology, are listed, namely for the H⁺, N^{δ1}-H and N^{ε2}-H form, respectively; the last row, for each His residue, list the electrostatic-computed tautomeric preference.

^f RMSD of fluctuations along the MD trajectory for fractions of ionized and tautomer forms of HIS residues.

^g in the first three rows the absolute difference () for each His form, i.e., between the *reference* and the electrostatically-calculated tautomeric fraction, respectively, are listed.

^h average absolute difference among *all* 's of each His, i.e., computed as: $1/3 \sum_{j=1,3} f_j$, are listed in the last row.

ⁱ averaged difference value, computed as: $1/n \sum_{i=1, n} < >_i$, with n the number of His residues and $< >$ defined in item (g) above, are listed in the last row.