Supporting Information

Vila et al. 10.1073/pnas.1102373108

SI Text

Assessing the Accuracy of the Calculations. Among the possible sources of error in our predictive method, e.g., those that originate in the computation of the shielding values, we focus *only* on those directly affecting the computation of the relative amounts of the different forms of the imidazole ring of histidine.

On one hand, the Fast Adaptive Multigrid Boundary Element (FAMBE-pH) method (1) provides results for pK predictions within $\sim \pm 0.3$ pK units, and, hence, it affects the accuracy of the computed average degree of charge, $\langle\rho\rangle$, which represents the fraction of the protonated form. The errors in the computed $\langle\rho\rangle$ for a given residue would reach a maximum at pH $\sim$ pK*, although in most of the applications to proteins, the disagreements originating in structural differences between the protein determined in crystals and in solution are expected to dominate the uncertainty in the computation of an accurate average degree of charge $\langle\rho\rangle$.

On the other hand, from Fig. 1 (main text), it is clear that $\Delta^{\delta} \sim 0$, for the $N^{\delta 1}$-H tautomer, but not exactly zero. In fact, $\Delta^{\delta} = 2.0 \pm 1.9$ ppm. To derive Eqs. **1** and **2**, in the main text, we have assumed that the first-order difference for the $N^{\delta 1}$-H tautomer is $\Delta^{\delta} = 0$ ppm, rather than $\Delta^{\delta} \sim 2.0$ ppm. Let us illustrate the error introduced by this assumption. If no charged form exists and $\Delta^{\text{obs}} = \sim 32$ ppm, then the average contribution of the $N^{\delta 1}$-H tautomer ($\Delta^{\delta} \sim 2.0$ ppm) would be less than $\sim$10% of the Δ^{obs} value, with the remaining contribution, i.e., more than 90%, coming from the $N^{\epsilon 2}$-H tautomer ($\Delta^{\epsilon} \sim 32$ ppm, as shown in Fig. 1). On the other hand, if both a charged, H^+, and the $N^{\delta 1}$-H tautomer coexist, then the same conclusion holds, i.e., about $\sim$90% of the Δ^{obs} value will be provided by the H^+ form because the Δ^{obs} value would be greater than $\sim$20 ppm, as for Residues H276 and H289 of protein 1E1A, which are fully protonated (Table S1).

Overall, from here on, we assume that the calculated fraction of each of the three forms of the imidazole ring of His, at a given pH, is affected by an estimated error of $\sim \pm 10\%$.

Determine an Applicability Limit. Residue H289 of protein 1E1A has the largest observed difference ($\Delta^{\text{obs}} \sim 26$ ppm) between $^{13}C^{\delta 2}$ and $^{13}C^{\gamma}$ chemical shifts among all the eight proteins (Table S1). This value is in good agreement with the computed difference in the shielding between $^{13}C^{\delta 2}$ and $^{13}C^{\gamma}$ nuclei in the protonated, H^+, form (Fig. 1, main text); i.e., the shielding difference computed for this tautomer can be fit by a Gaussian distribution with a mean value Δ^+ of 22.9 ppm and a standard deviation of $\pm$4.7 ppm. This agreement is not surprising because the observed chemical shifts for H289 were obtained at pH 6.5, and, at this pH, the computed average degree of charge of H289 is $\langle\rho\rangle = 0.98$. In other words, 98% of the imidazole rings of H289 are in the H^+ form. It should be noted that the average degree of charge of H289 is significantly higher than that of a free His in solution at the same pH ($\langle\rho\rangle = 0.56$), because there is a favorable electrostatic interaction of the imidazole ring of H289 with the highly deprotonated ($\langle\rho\rangle = 0.97$) Glu39 side chain (Fig. S5). A detailed analysis of all five remaining His residues in protein 1E1A are discussed in *Computation of the Tautomeric Distribution of the Imidazole Ring of His in Proteins*, in the main text.

At this point, it is worth noting that $\langle\rho\rangle \sim 1.0$ does not always imply $\Delta^{\text{obs}} \sim 26$ ppm, e.g., as observed for H64 and H66 from protein 1HOE, with $\langle\rho\rangle = 1.0$ and $\Delta^{\text{obs}} = |^{13}C^{\delta 2}\text{-}^{13}C^{\gamma}| \sim 10$ ppm. Because of the nature of the Gaussian statistics of the differences between $^{13}C^{\delta 2}$ and $^{13}C^{\gamma}$ nuclei, shielding values within $\sim$3 standard deviations ($\sim \pm 14$ ppm) from the mean value, $\Delta^+ = \sim 23$ ppm, can be considered as likely to occur. However, Δ^{obs} values far from these likely ones, i.e., outliers, may be observed; see, for example, H183 of protein 1E1A, with $\langle\rho\rangle \sim 0.9$ and $\Delta^{\text{obs}} \sim 1$ ppm (Table S1). Conceivably, the origin of this inconsistency could be found in significant differences between the structure in solution, for which the chemical shifts are observed, and the one in the crystal, for which the averaged degree of charge is computed. Consequently, from here on, if $\langle\rho\rangle \sim 1.0$ and $\Delta^{\text{obs}} < \Delta^{\text{cutoff}} = \sim 9$ ppm [as for the lowest limit, 3 standard deviations (14 ppm) of the mean value (23 ppm) of the Gaussian distribution], this combination of data will be considered incompatible and, hence, the method does not apply.

Overall, validation of the computed shielding ranges compared to the observed ^{13}C chemical-shift values from single proteins enables us to define a cutoff for the Δ^{obs} value, namely $\Delta^{\text{cutoff}} = \sim 9$ ppm, with which to decide if the method can or cannot be applied.

Cross-Validation Test. A study of the mechanism of proton conduction in influenza M2 proton channels from solid-state NMR by Hu et al. (2) reveals important findings. In particular, by using two-dimensional ^{13}C-^{13}C and ^{15}N-^{13}C correlation spectra of labeled His37, the authors show that only the neutral imidazole ring of His37 exists at pH 8.5. From a graphic analysis of their figure 1, Hu et al. (2) conclude that the $N^{\epsilon 2}$-H tautomer and $N^{\delta 1}$-H tautomer exist in an $\sim$3:1 ratio. Using their reported (2) $^{13}C^{\delta 2}$ (112.9 ppm) and $^{13}C^{\gamma}$ (135.8 ppm) chemical shifts for the $N^{\epsilon 2}$-H tautomer (see their figure 1*E*), we can compute the tautomeric fractions of His37. Thus, using the value of $\Delta^{\text{obs}} = |^{13}C^{\Delta 2}\text{-}^{13}C^{\gamma}| \sim 23$ ppm in Eqs. **1** and **2** in the main text, we obtain $f^{\epsilon} \sim 0.72$ and $f^{\delta} \sim 0.28$, respectively. This result is in fair agreement with their calculated $\sim$3:1 ratio.

1. Vorobjev YA, Vila JA, Scheraga HA (2008) FAMBE-pH: A fast and accurate method to compute the total solvation free energies of proteins. *J Phys Chem B* 112:11122–11136.
2. Hu F, Wenbin L, Hong M (2010) Mechanism of proton conduction and gating in influenza M2 proton channels from solid-state NMR. *Science* 330:505–508.

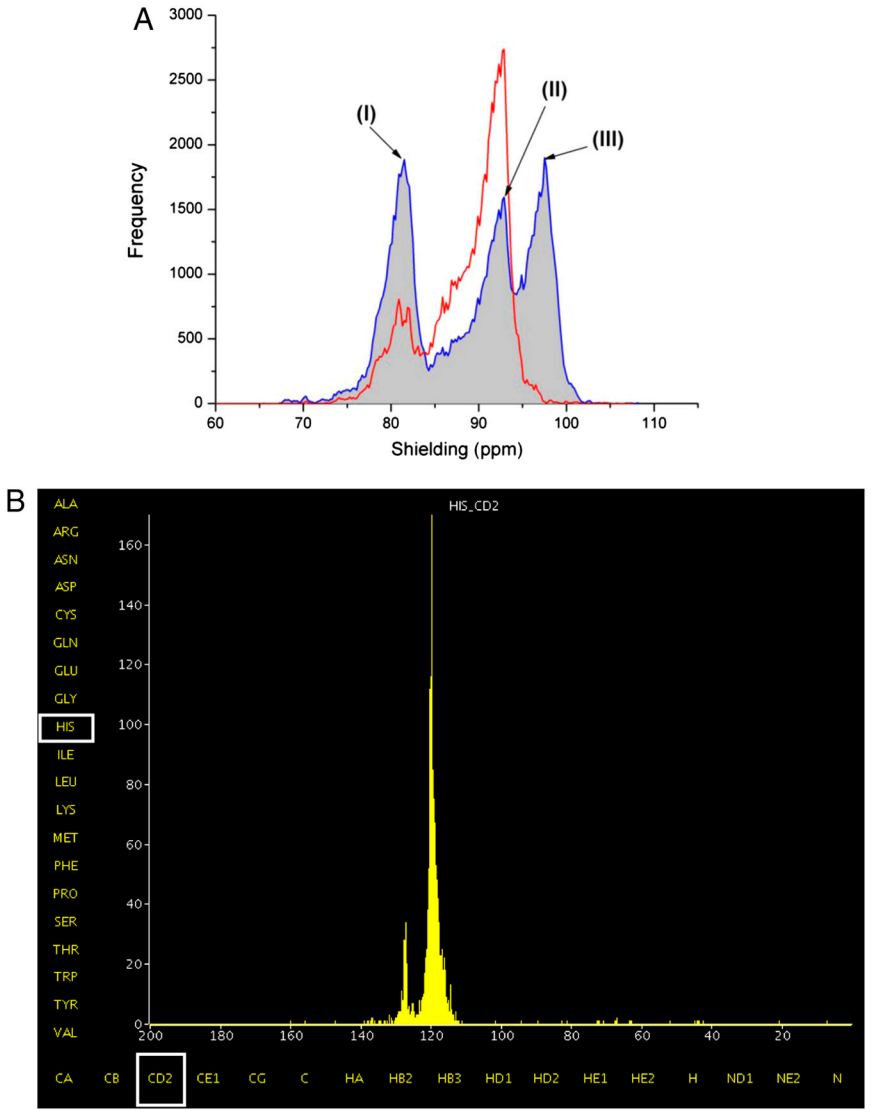

Fig. S1. (*A*) The blue-line profile represents the histogram of the shielding distribution of the $^{13}C^{\delta 2}$ nucleus for all forms of the imidazole ring of His without distinction among them, i.e., with the surface under the blue line representing a collection of more than ~110,000 shieldings from the model tripeptide Ac-G**H**G-NMe. The heights of the peaks reflect the different frequencies of the shielding distributions. The red-line profile is the histogram of the assumed shielding distribution (see main text) with computed arbitrarily selected contributions of each form as follows: 70% from H^+, and 15% from each of the tautomers $N^{\delta 1}$-H and o$N^{\epsilon 2}$-H, respectively. (*B*) Histogram of the frequency of the chemical-shifts distribution of the 2,267 observed values of $^{13}C^{\delta 2}$, in yellow, as obtained from the Biological Magnetic Resonance Data Bank (BMRB) (as of October 2010). The white frame on the *y* axis shows the chosen amino acids, and on the *x* axis the selected nucleus, among all nuclei for which chemical-shift values are deposited [courtesy of Kent R. Wenger (BMRB, Madison, WI)].

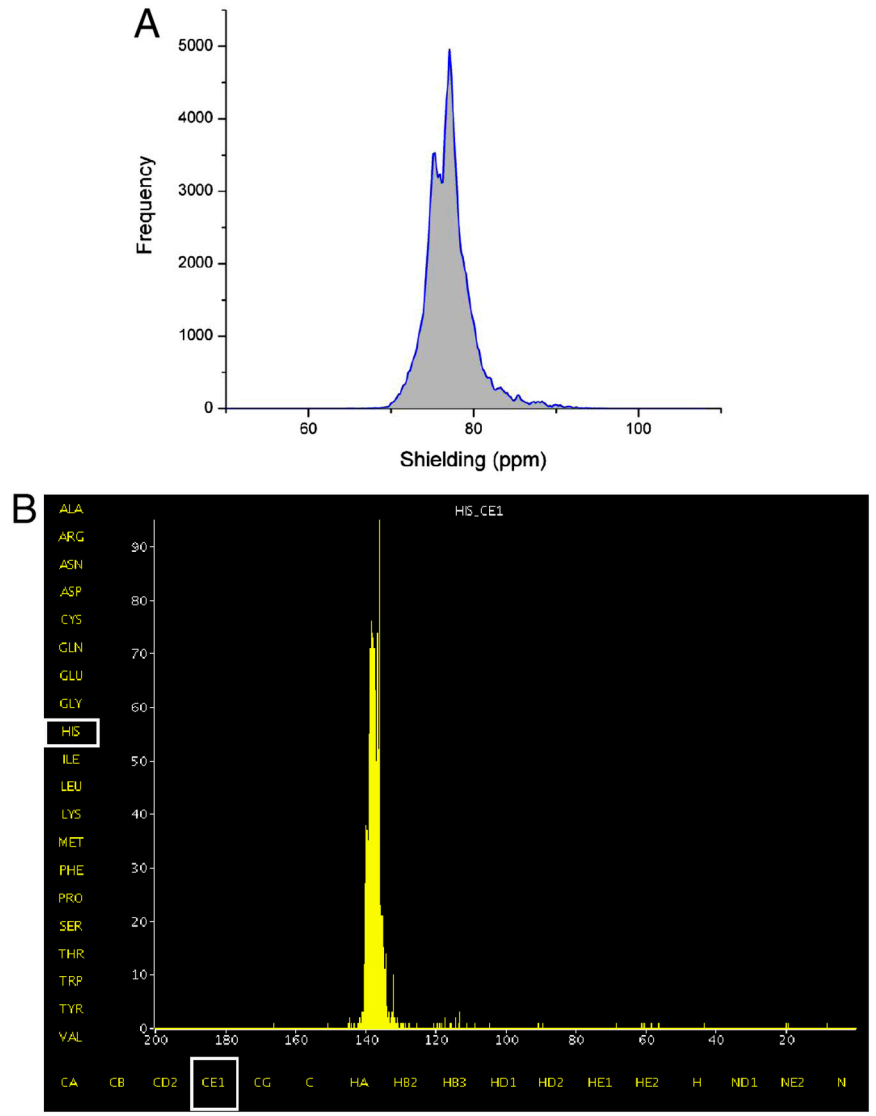

Fig. S2. (*A*) The blue-line profile represents the frequency of the shielding distribution of the $^{13}C^{\epsilon 1}$ nucleus for all forms of the imidazole ring of His; i.e., with the surface under the blue-line representing a collection of more than ~110,000 shieldings with the same contribution among all three forms, as in the model tripeptide Ac-G**H**G-NMe. (*B*) Frequency of the chemical-shift distribution of the 1,754 observed values of $^{13}C^{\epsilon 1}$, in yellow, as obtained from the BMRB database (as of October 2010). The white frame on the *y* axis shows the chosen amino acids, and on the *x* axis the selected nucleus (courtesy of Kent R. Wenger of the BMRB)

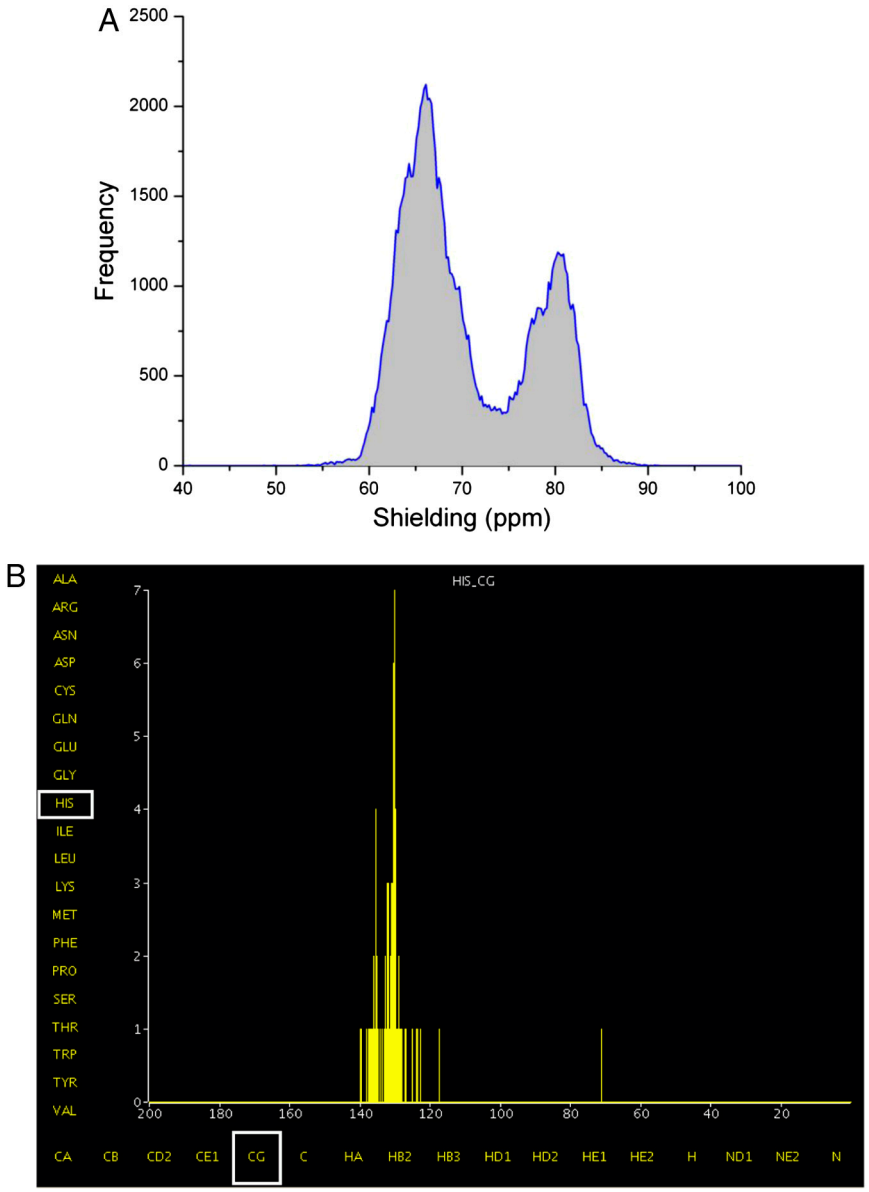

Fig. S3. (*A*) The blue-line profile represents the frequency of the shielding distribution of the $^{13}C^{\gamma}$ nucleus for all forms of the imidazole His ring, i.e., with the surface under the blue-line representing a collection of more than ~110,000 shieldings with the same contribution among all three forms. (*B*) Frequency of the observed 84 $^{13}C^{\gamma}$ chemical-shift distribution, in yellow, as obtained from the BMRB database (as of October 2010). The white frame on the *y* axis shows the chosen amino acids, and on the *x* axis the selected nucleus (courtesy of Kent R. Wenger of the BMRB).

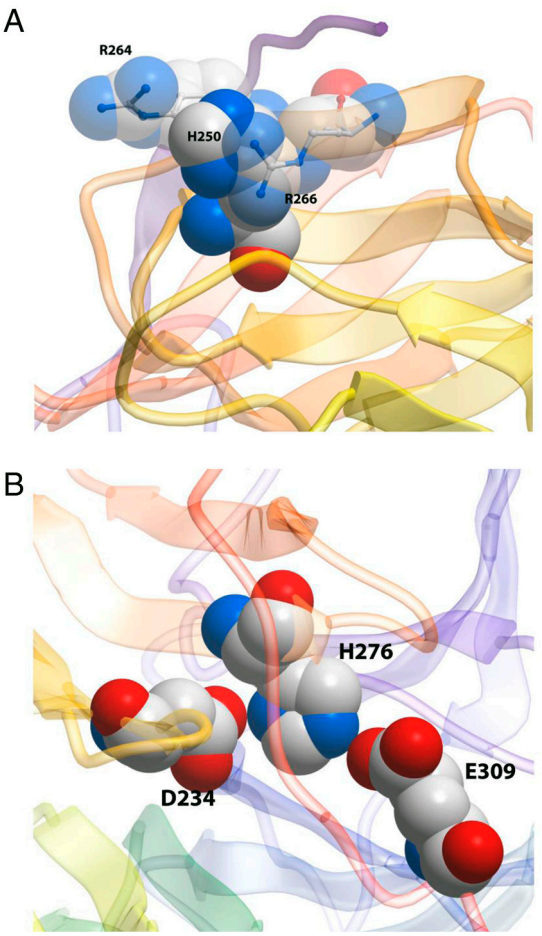

Fig. S4. (*A*) Space-filling-sphere representation for H250, R264, and R266. The gray, blue, and red colors designate carbon, nitrogen, and oxygen atoms, respectively. The background shows a ribbon diagram of part of the protein 1E1A; (*B*) same as *A* for H276, D234, and E309 in protein 1E1A.

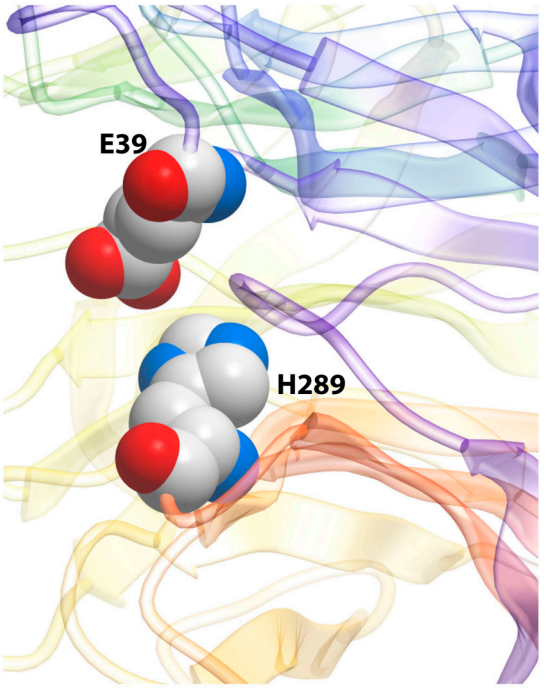

Fig. S5. Space-filling-sphere representation for H289 and E39 in protein 1E1A. The gray, blue, and red colors designate carbon, nitrogen, and oxygen atoms, respectively. The background shows a ribbon diagram of part of protein 1E1A.

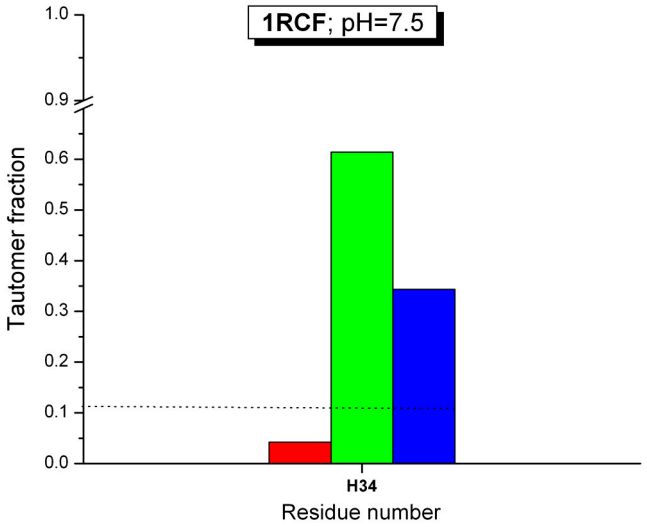

Fig. S6. Fraction of tautomeric distribution for H34 of protein 1RCF, for which the chemical shifts were determined in solution at pH 7.5. Red and blue bars represent the fraction of the $N^{\epsilon 2}$-H and $N^{\delta 1}$-H tautomer, respectively, and the green bar represents the fraction of the protonated form, H^+. The dotted horizontal line indicates the fraction of the H^+ tautomer that a free His residue would have in solution at pH 7.5.

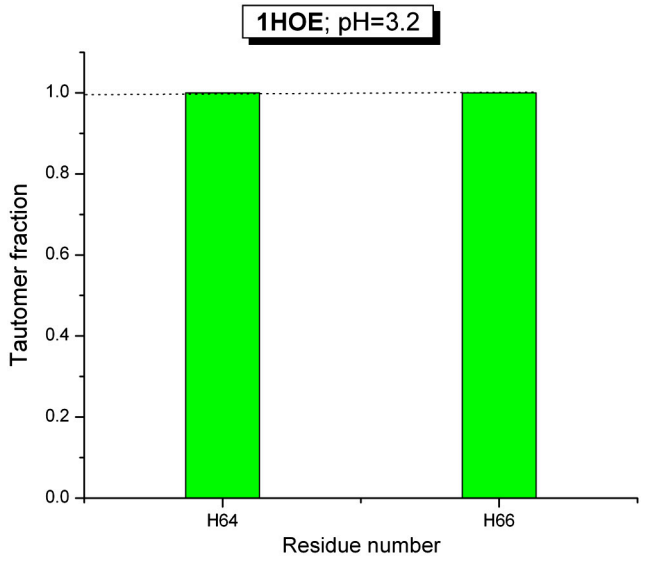

Fig. S7. Fraction of tautomeric distribution for each of the two His residues of protein 1HOE, for which the chemical shifts were determined in solution at pH 3.2. Green bars represent the fraction of the H^+ form for each His. The dotted horizontal line indicates the fraction of the H^+ form that a free His residue would have in solution at pH 3.2.

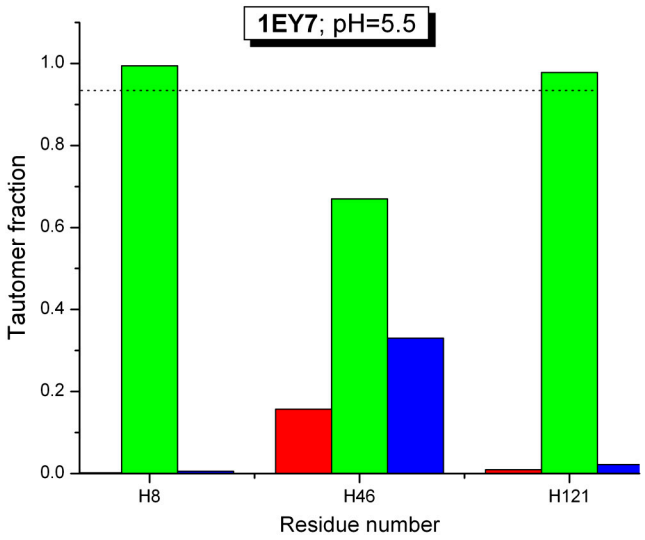

Fig. S8. Fraction of tautomeric distribution for each of the three His residues of protein 1EY7, for which the chemical shifts were determined in solution at pH 5.5. Red and blue bars represent the fraction of the $N^{\epsilon 2}$-H and $N^{\delta 1}$-H tautomer, respectively, and the green bar represents the fraction of the protonated form, H^+. The dotted horizontal line indicates the fraction of the H^+ tautomer that a free His residue would have in solution at pH 5.5.

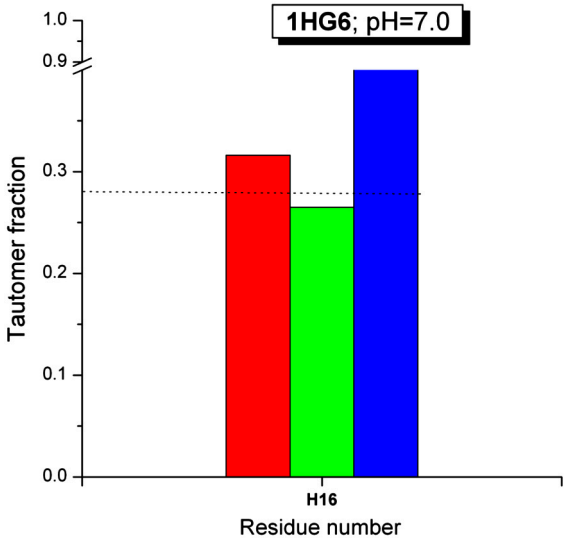

Fig. S9. Fraction of tautomeric distribution for H16 of protein 1HG6, for which the chemical shifts were determined in solution at pH 7.0. Red and blue bars represent the fraction of the $N^{\epsilon 2}$-H and $N^{\delta 1}$-H tautomer, respectively, and the green bar represents the fraction of the protonated form, H^+. The dotted horizontal line indicates the fraction of the H^+ form that a free His residue would have in solution at pH 7.0.

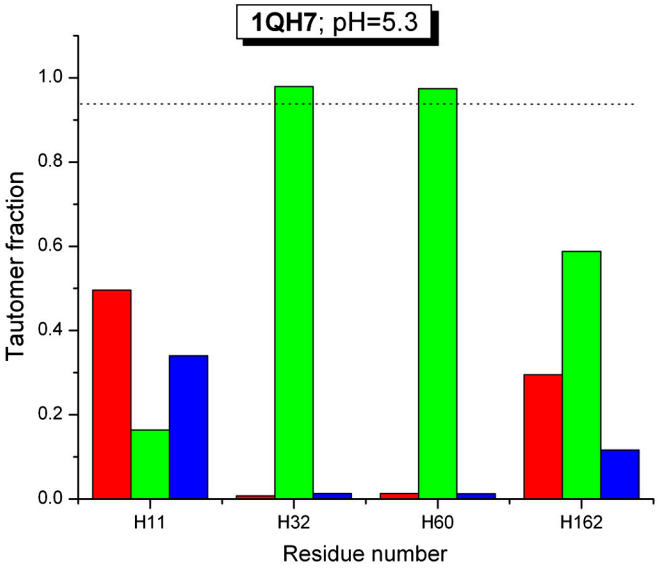

Fig. S10. Fraction of tautomeric distribution for each of the four His residues of the protein 1QH7, for which the chemical shifts were determined in solution at pH 5.3. Red and blue bars represent the fraction of the $N^{\epsilon 2}$-H and $N^{\delta 1}$-H tautomer, respectively, and the green bar represents the fraction of the protonated form, H^+. The dotted horizontal line indicates the fraction of the H^+ form that a free His residue would have in solution at pH 5.3.

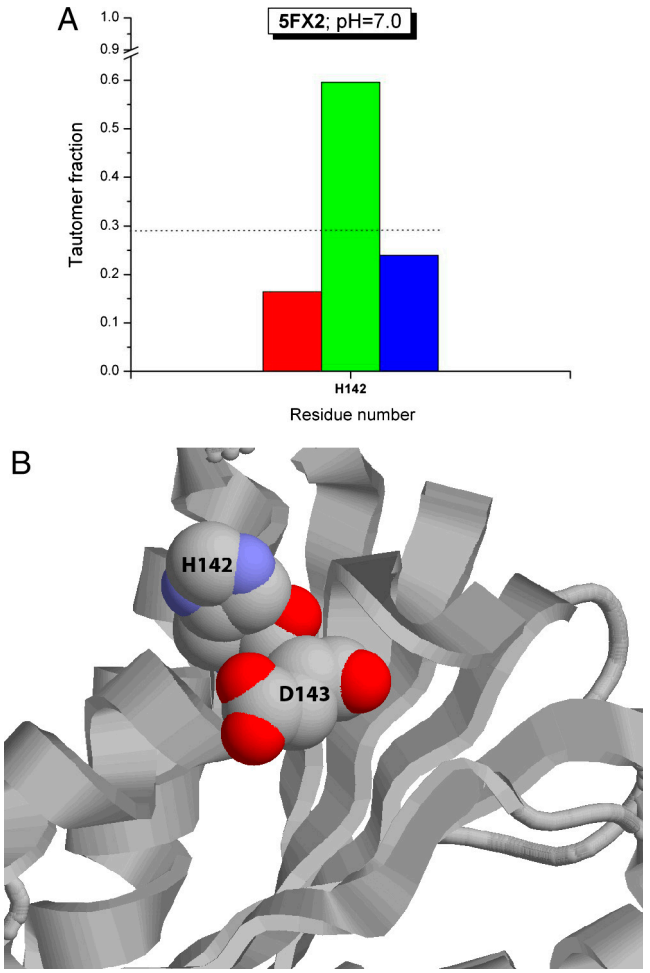

Fig. S11. (*A*) Fraction of tautomeric distribution for the His of protein 5FX2, for which the chemical shifts were determined in solution at pH 7.0. Red and blue bars represent the fraction of the $N^{\epsilon 2}$-H and $N^{\delta 1}$-H tautomer, respectively, and the green bar represents the fraction of protonated form, H^+. The dotted horizontal line indicates the fraction of the H^+ form that a free His residue would have in solution at pH 7.0. (*B*) Space-filling-sphere representation for H142 and D143, in protein 5FX2. The gray, blue, and red colors designate carbon, nitrogen, and oxygen atoms, respectively. The background shows a ribbon diagram of part of protein 5FX2.

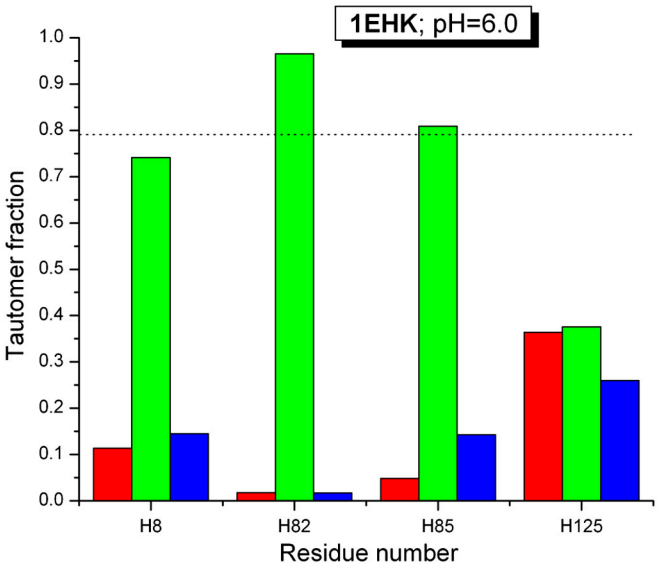

Fig. S12. Fraction of tautomeric distribution for each of the four His residues of the protein 1EHK, for which the chemical shifts were determined in solution at pH 6.0. Red and blue bars represent the fraction of the $N^{\epsilon 2}$-H and $N^{\delta 1}$-H tautomer, respectively, and the green bar represents the fraction of the protonated form, H^+. The dotted horizontal line indicates the fraction of the H^+ form that a free His residue would have in solution at pH 6.0.

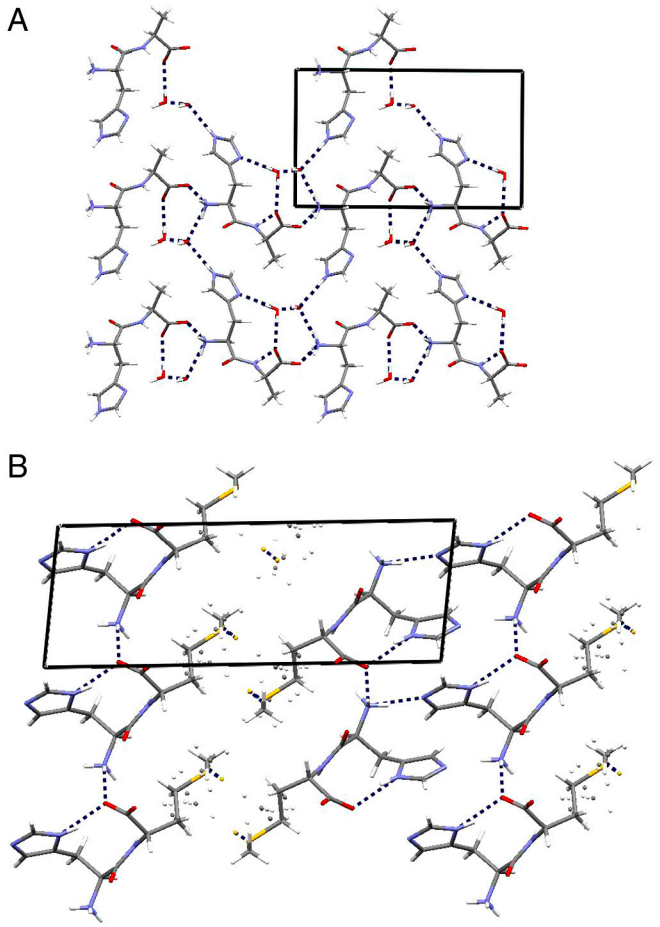

Fig. S13. The crystal-packing scheme and pattern of conventional hydrogen bonds (dark blue dashed lines) for (*A*) His-Ala, compound 10 (viewed along the *a* axis) and (*B*) His-Met, compound 6 (viewed along the *b* axis). The white, red, and light blue colors designate hydrogen, oxygen, and nitrogen atoms, respectively.

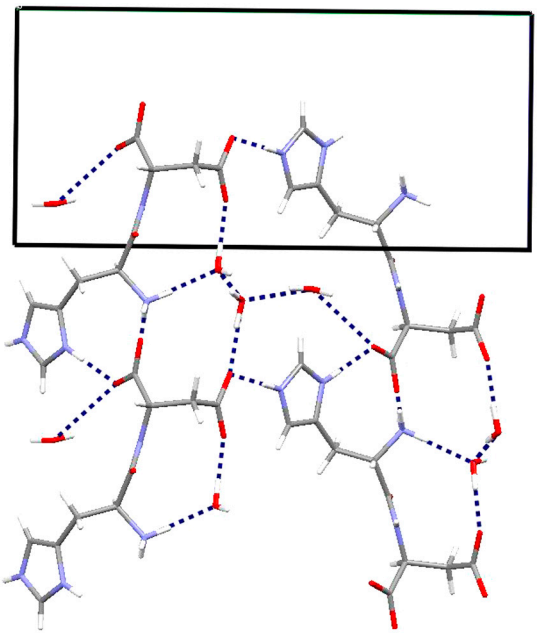

Fig. S14. Crystal-packing scheme, based on data of ref. 1, illustrating the pattern of conventional hydrogen bonds (dark blue dashed lines) formed by the imidazole ring of His in His-Asp (compound 9 in Table S2, viewed along the *a* axis). The white, red, and light blue colors designate hydrogen, oxygen, and nitrogen atoms, respectively.

1. Cheng F, Sun H, Zhang Y, Mukkamala D, Oldfield E (2005) A solid state 13C NMR, crystallographic, and quantum chemical investigation of chemical shifts and hydrogen bonding in histidine dipeptides *J Am Chem Soc* 127:12544–12554.

Table S1. Data used to compute the His tautomeric distribution in proteins

Protein*	BMRB†	His number‡	Chemical shifts§ $^{13}C^{\gamma}$	$^{13}C^{\delta 2}$	Δ^{obs}	Charge¶ $\langle\rho\rangle$
1RCF	1580 (7.5)	34	126.8	123.5	3.3	0.61
1HOE	1642 (3.2)	64	129.6	117.6	12	1.00
		66	128.9	118.1	10.8	1.00
1EY7	1704 (5.5)	8	129.7	118.7	11	0.99
		46	132.1	118.6	13.5	0.67
		121	130.8	117.8	13	0.98
1HG6‖	5160 (7.0)	16	131.2	118.3	12.9	0.27
1QH7	5352 (5.3)	11	135.4	117.6	17.8	0.16
		32	131.0	120.3	10.7	0.98
		60	135.4	120.3	15.1	0.97
		162	137.5	116.0	21.5	0.59
5XF2	5571 (7.0)	142	131.8	119.6	12.2	0.60
1E1A**	5618 (6.5)	183	*128.3*	*127.4*	*0.9*	0.85
		221	134.4	120.1	14.3	0.69
		226	138.1	119.2	18.9	0.41
		250	130.0	126.3	3.7	0.38
		276	139.7	117.3	22.4	1.00
		289	139.8	113.9	25.9	0.98
1EHK††	5819 (6.0)	8	133.0	119.8	13.2	0.74
		82	135.5	120.2	15.3	0.97
		85	129.8	122.2	7.6	0.81
		125	136.5	119.0	17.5	0.38

*PDB (1) ID of the protein used in this work. The structures of all these proteins, except 1HG6, have been determined by X-ray crystallography and are represented by a single conformation.

†BMRB (2) accession number from which the NMR-determined chemical shifts were deposited. In parentheses the pH at which the chemical shifts were determined.

‡List number of His in the BMRB protein sequence.

§Values for the observed $^{13}C^{\gamma}$ and $^{13}C^{\delta 2}$ chemical shifts, in ppm, from the accession number listed in column 2. The listed chemical-shift values are used to obtain $\Delta^{obs} = |^{13}C^{\delta 2}\text{-}^{13}C^{\gamma}|$ and, hence, to compute f^{ϵ} in Eq. **1**. Values in italics were not used to compute the tautomer distribution for the imidazole His ring because $\langle\rho\rangle \sim 1.0$ and the Δ^{obs} value (~1 ppm) is much lower than the Δ^{cutoff} value (~9 ppm).

¶Value of the average degree of charge of His computed with FAMBE-pH (3) at the pH value listed in column 2.

‖Protein (1HG6) is obsolete in the PDB; *only* model 1 of the NMR-determined structure was used to compute the average degree of charge because it has not been replaced. Hence, caution should be exercised in interpreting the results derived from this protein.

**Computation of the fraction of tautomers for residue H183 was avoided for the reason explained in the main text.

††Chemical shifts observed for only four out of six His residues in the sequence of 1EHK.

1. Berman HM, et al. (2000) The protein data bank. *Nucleic Acids Res* 28:235–242.
2. Ulrich EL, et al. (2008) BioMagResBank. *Nucleic Acids Res* 36:D402–D408.
3. Vorobjev YA, Vila JA, Scheraga HA (2008) FAMBE-pH: A fast and accurate method to compute the total solvation free energies of proteins. *J Phys Chem B* 112:11122–11136.

Table S2. Tautomeric distribution of His-containing dipeptides

Compound*	Form†	Observed Δ^{obs} (ppm)	Computed‡ f^{ϵ} (%)	f^{δ} (%)
His-Leu (5)	$N^{\delta 1}$-H	3.6	11	89
His-Met (6)	$N^{\delta 1}$-H	0.0	0	100
Gly-His (7)	$N^{\delta 1}$-H	3.8	12	88
Leu-His (8)	$N^{\delta 1}$-H	9.5	30	70
His-Ala (10)	$N^{\epsilon 2}$-H	14.3	45	55
His-Glu (11)	$N^{\epsilon 2}$-H	15.0	47	53
Ala-His (12)	$N^{\epsilon 2}$-H	22.9	72	28
His-Asp (9)	H^{+}	8.3	unknown	unknown

*Sequence of the unblocked dipeptides, crystallized by Cheng et al. (1), from which the Δ^{obs} value was computed (see column 3), namely as $\Delta^{obs} = |^{13}C^{\delta 2}\text{-}^{13}C^{\gamma}|$, with $^{13}C^{\delta 2}$ and $^{13}C^{\gamma}$ being the observed chemical shifts (1), from ^{13}C solid-state NMR. The number in parentheses designates the compound number notation used by Cheng et al. (1) in their table 2.

†Experimental form (1) for the His of each of the 8 crystallized dipeptides.

‡Tautomeric fractions of the imidazole ring of His computed here with Eqs. **1** and **2**. For each of the first seven compounds, namely 5–8 and 10–12, we assume, as the experimental evidences indicates (1), that *all* His in the crystal are in the neutral form; hence $\langle\rho\rangle = 0$ in Eqs. **1** and **2**. Concerning the result "unknown" obtained for the protonated form, compound 9, see *Computation of the Tautomeric Distribution of the His-Containing Dipeptides* in the main text.

1. Cheng F, Sun H, Zhang Y, Mukkamala D, Oldfield E (2005) A solid state 13C NMR, crystallographic, and quantum chemical investigation of chemical shifts and hydrogen bonding in histidine dipeptides *J Am Chem Soc* 127:12544–12554.