Can the protonation state of histidine residues be determined from molecular dynamics simulations?

Uranga, Jon; Mikulskis, Paulius; Genheden, Samuel; Ryde, Ulf

Published in:
Computational and Theoretical Chemistry

DOI:
10.1016/j.comptc.2012.09.025

2012

Can the protonation state of histidine residues be determined from molecular dynamics simulations?

Jon Uranga, Paulius Mikulskis, Samuel Genheden, Ulf Ryde*

Department of Theoretical Chemistry, Lund University, Chemical Centre, P. O. Box 124, SE-221 00 Lund, Sweden

Correspondence to Ulf Ryde, E-mail: Ulf.Ryde@teokem.lu.se, Tel: +46 – 46 2224502, Fax: +46 – 46 2228648

2012-09-10

Histidine (His) residues in proteins can attain three different protonation states at normal pH. This constitutes a prominent problem when adding protons to a protein crystal structure, e.g. in order to perform molecular simulations. Typically, the His protonation is deduced from the hydrogen-bond pattern in crystal structures. Here, we study whether it is possible to detect erroneous His protonation state by analysing short molecular dynamics (MD) trajectories. We systematically vary the His protonation state and measure the root-mean-squared deviation (RMSD) of the His residues and nearby residues relative to the starting structure, as well as the distribution of the dihedral angle that determines the rotation of the His side chain. We study three proteins, hisactophilin with 31 solvent-exposed His residues, galectin-3, for which an experimental assignment is available for two of the His residues, and trypsin, for which the hydrogen-bond analysis is quite conclusive. The results show that improper protonation states have larger RMSD values and larger widths of the dihedral distribution, compared to the correct protonation states. Unfortunately, the variation among different His residues in the same and different proteins is so large that it is hard to define unambiguous thresholds between proper and improper protonation states. Therefore, simulations of all three protonation states are needed for conclusive results. For trypsin, we could obtain a conclusive assignment for all three His residues, which was better than the simple hydrogen-bond analysis. For galectin-3, the MD trajectories confirmed the results of hydrogen-bond analysis and experiments. They also gave additional, more uncertain information for some of the residues. However, for the solvent-exposed His residues in hisactophilin, no unambiguous conclusions regarding the protonation states could be reached. On the other hand, this indicates that protein structures are quite insensitive to the protonation state of the His residues, besides those that involve direct hydrogen bonds to the His side chain.

Key Words: Histidine protonation, molecular dynamics simulations, side-chain rotation, dihedral distribution.

Introduction

Molecular dynamics (MD) simulations are a versatile tool to study the structure and function of proteins, because they give a picture in atomistic detail of the time behaviour of the system. They are based on the numerical solution of Newton's equations, typically describing the protein with a molecular-mechanics potential.

Most protein MD simulations are started from a crystal structure. Unfortunately, no coordinates are available for hydrogen atoms in the great majority of protein structures. Therefore, the positions of the hydrogen atoms have to be generated from the coordinates of the heavy atoms. For many groups in proteins, this is either simple (e.g. for the backbone, aromatic, CH_2, and CH groups) or of minor energetic importance (e.g. methyl groups). However, for water and hydroxyl groups, the addition of protons is much harder and the selection may strongly affect the energies, because these groups form hydrogen bonds.

Even worse, for histidine (His) residues, it is not even known how many protons should be added or to which atom they should bind. The reason for this is that the imidazole side chain of His contains two nitrogen atoms that can be protonated, ND1 and NE2. Moreover, the pK_a of the side chain is close to 7 [1], so that at neutral pH, it is almost equally probable that one or both of the nitrogen atoms are protonated. Therefore, there are three protonation states of His that all are possible around pH 7. They are shown in Figure 1 and they will be called HID, HIE, and HIP in the following, depending on whether the ND1, the NE2, or both nitrogen atoms are protonated. Most MD modelling softwares include tools to protonate protein structures, but many of them assume that the user will manually assign the protonation state of His residues (the three protonation states of His have different residue names).

However, there are several softwares to help such an assignment. First, there are many different methods to estimate the pK_a value of protein residues [1,2]. Unfortunately, most of these methods are quite time-consuming, so the only method that has gained wide use in the setup of proteins is the simple and fast PROPKA approach [3]. Second, there are several automatic methods to select which nitrogen atom to protonate in His, e.g. What_Check [4] and REDUCE [5], which analyse the hydrogen-bond pattern and possible steric clashes that might arise between the His hydrogen atoms and surrounding residues. They also consider the possibility that the C and N atoms in the His side chain are mixed up, which often happens because they are not clearly discernible in crystal structures [4]. In addition, Signorini et al. have tried to determine the His protonation state by analysing energies obtained for the three protonation states after energy minimisation [6], but such an approach is questionable because the various protonation states have different topologies, making the energies not comparable.

All these methods try to estimate the His protonation state before the MD simulations, which of course would be very valuable. However, once the MD simulations are run, much additional information is available. In this study, we investigate whether erroneous choices of the His protonation can be detected from the MD trajectories. We do this by systematically simulating three proteins with all possible protonation states of each His residue and then analysing the fluctuations of the His residue and the surrounding residues, as well as the rotation around the CB–CG bond in the His side chain.

Methods

Preparation of the proteins

The simulations of trypsin were based on the 1HJ9 crystal structure of trypsin in complex aniline[7], but the aniline ligand was changed to 2,4-difluoroaniline, ligand 3 in the Sample3 challenge [8]. The simulations of galectin-3 were based on the 2XG3 crystal structures in complex with a benzamido-N-acetyllactoseamine inhibitor [9]. The hisactophilin calculations were based on the 1HCD NMR structure [10]. Protons were added to all three structures using the tleap software in the Amber suite of programs, assuming typical protonation states at pH

7.0 (i.e. Asp and Glu residues were assumed to be negatively charged, whereas Arg and Lys residues were assumed to be positively charged). The protonation of the His residues were systematically varied, as will be discussed below. The proteins were modelled by the Amber99SB force field [11] and the ligands were described with the general Amber force field (GAFF) [12], using charges calculated with the RESP method [13], based on quantum mechanical calculations of the electrostatic potential at the Hartree–Fock level with the 6-31G* basis set and points sampled with the Merz–Kollman scheme [14]. The setup of the ligands has been described before [8,15].

Trypsin contains a Ca^{2+} ion, which was described with a non-bonded potential, using the formal +2 charge and Lennard-Jones parameters of 1.60 Å and 0.42 kJ/mol (from the Amber parm91.dat file). This gave Ca–O distances of ~2.25 Å in the MD simulations. Each protein–ligand complex was immersed in a periodic truncated octahedral box of TIP3P water molecules, [16] which extended at least 10 Å outside the protein.

Simulation protocol

All MD simulations were run by either the sander or pmemd modules in Amber 11 [17]. The temperature was kept at 300 K using Langevin dynamics [18] with a collision frequency of 2.0 ps^{-1}. The pressure was kept at 1 atm using a weak-coupling approach [19] with isotropic position rescaling and a relaxation time of 1 ps. The long-range electrostatics were treated by particle-mesh Ewald summation [20] with a fourth-order B-spline interpolation and a tolerance of 10^{-5}. The non-bonded cutoff was 8 Å and the non-bonded pair list was updated every 50 fs. The MD time step was 2 fs and the SHAKE algorithm [21] was used to constrain bonds involving hydrogen atoms.

The proteins were simulated the following way: First, the protein was minimised with all non-hydrogen atoms except water oxygen atoms restrained towards the starting structure with a force constant of 418 kJ/mol/$Å^2$, followed by a 20 ps restrained MD simulation in the *NPT* ensemble, and a 1 ns unrestrained equilibration. After this equilibration, ten independent simulations were initiated by assigning different starting velocities. Each of the simulations was further equilibrated for 100 ps in the *NPT* ensemble, before a 200 ps production run was performed, in which snapshots were saved every 5 ps. For trypsin, we also tested production runs of 100 and 1000 ps.

Analysis

We analysed the root-mean-squared deviation (RMSD) from the starting structure of the atoms in each histidine residue, as well as of all residues within 3.5 Å of the protonated side chain of each His residue (RRD). These RMSD values were obtained for each snapshot from the MD simulations and they are presented as the average over the ten independent simulations. The uncertainties are the standard errors of the mean values, i.e. the standard deviation over the ten independent trajectories divided by $\sqrt{10}$.

Moreover, we studied the CA–CB–CG–ND1 dihedral angle (called *φ* in the following), to measure wiggling and rotation of the imidazole ring. The distribution of φ was modelled with a Gaussian mixture model (GMM) [22,23] to determine the number of minima, their locations, and widths. This approach models the total distribution as a sum of univariate Gaussian distributions. Each of these distributions will be called a state. The probability that a data point (a particular value of *φ*, denoted *y* in the following formulas) comes from state *k* is denoted π_k, and the distribution of each state is

$$p(y \mid \text{from class } k, \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{(-y-\mu_k)^2}{2\sigma_k^2}\right) \quad (1)$$

where μ_k and σ_k^2 are the mean and variance of state *k*. The total distribution is

$$p(y|\vec{\pi},\vec{\mu},\vec{\sigma}) = \sum_k \pi_k p(y|\text{from class } k,\mu_k,\sigma_k) \quad (2)$$

To determine which state each data point belongs to and the values of the parameters π_k, μ_k, and σ_k^2, we use an expectation-maximisation algorithm [24]. Initially, we assume that there are two states located at φ = –120° and 120°, both with a variance of 10 and a probability of 0.5. The parameters are then iteratively updated until convergence. If the probability of a state in any iteration falls below 0.001, that state is discarded. The distribution of φ was obtained by pooling the data from the ten independent simulation, i.e. 400 snapshots were used to create the GMM. Standard deviations of the model parameters were obtained by bootstrapping [25].

Result and Discussion

We have studied whether erroneous protonation states of His residues can be detected from short MD simulations. As test cases, we use three different proteins, hisactophilin, galectin-3, and trypsin. The first was chosen because it contains many His residues, which are solvent exposed. Therefore, these calculations may show the typical behaviour of surface His residues. On the other hand, there is no information of the preferred protonation state of any of these residues. Galectin-3 was chosen because there are experimental information about the protonation state for two of the His residues. Finally, trypsin was chosen because the hydrogen-bond analysis gave clear information about the protonation state of all three residues.

For each protein, we started the investigation by analysing the hydrogen-bond pattern and solvent-accessibility of all His residues. By such an analysis, we get some tentative information about the protonation states of the His residues. All His residues were also examined by the PROPKA [3], What_Check [4], and REDUCE [5] softwares to get further information about the protonation state. Then, we performed MD simulations, systematically varying the protonation states of the His residues, and analysed the RMSD of the His residues and their surroundings, as well as the variation of the CA–CB–CG–ND1 dihedral angle (φ). The hypothesis was that if we use an unfavourable protonation state of a His residue, it will be incompatible with the experimental structure and therefore it will show larger fluctuations and possibly rotations of the His side chain to resolve these problems. Therefore, an unusually large RMSD and a large variation in the φ dihedral angle may indicate an erroneous protonation state of the His residue. In the following, we will discuss the results for each protein individually.

Trypsin

Trypsin contains three His residues, His40, 57 and 91. The analysis of the hydrogen-bond pattern in the crystal structure is presented in Table 1 and Figure 2. It showed that the NE2 atom of His40 forms a hydrogen bond with the backbone carbonyl O atom of Gly193 (H–O distance of 2.0 Å). The latter atom can only accept hydrogen bonds, showing that NE2 must be protonated. Therefore, His40 is unlikely to be in the HID state. The ND1 atom forms a hydrogen bond with the side-chain OH group Ser32, which can be both an acceptor and donor of hydrogen bonds. Consequently, we cannot decide from the hydrogen-bond analysis whether His40 is in the HIE or HIP state.

For His57, which is a part of the catalytic triad in the enzyme, the NE2 atom forms

hydrogen bond with the side-chain hydroxyl group of catalytic Ser195, which can be both an acceptor and donor of hydrogen bonds. The ND1 atom makes a hydrogen bond with OD2 of Asp102 (1.7 Å distance), which is a hydrogen-bond acceptor. Therefore, ND1 must be protonated and the HIE state is unlikely.

Finally, we found that the ND1 atom of His91 forms a hydrogen bond with the backbone N atom of Ser93 (2.0 Å distance). This atom is a hydrogen-bond donor, so ND1 cannot be protonated. This settles the protonation state of this residue to HIE. The NE2 atom forms a hydrogen bond to a water molecule, which can be both a donor and acceptor of hydrogen bonds, consistent a HIE state of this residue.

Calculations with the PROPKA approach [3], suggested that His40 and 91 are neutral (HID or HIE) at pH 7, because the calculated pK_a values were below 7. However, His57 was predicted to be in the doubly protonated HIP state with a pK_a of 7.4. The What_Check software [4] gave ambiguous predictions for His40 and 57 (the program uses two different criteria, which for these two residues gave different results). However, for His91, What_Check predicted a HIE state, in agreement with the hydrogen-bond analysis. Finally, the REDUCE software [5] predicted that His40 and 91 are in the HIE state, whereas His57 is in the HIP state.

The latter assignment is in agreement also with the predictions of the other three methods, so we will consider this as the preferred protonation state, and it was used in the simulation setup set 1 (SS1). Then, we set up six additional setup sets (SSs) by changing the protonation state of one of the residues in turn, as is described in the first part of Table 2. For each of the seven SSs, ten independent simulations were run. During the 200 ps production simulation (after 1.0 + 0.1 ns equilibration), snapshots were collected every 5 ps and the RMSD and φ dihedral variation was analysed for each snapshot. All reported results are averages over the ten independent simulations and standard errors (SE) are the standard deviation over these ten simulations divided by $\sqrt{10}$. The results of the RMSD analysis for the trypsin simulations are collected in Table 2. The distribution of the φ dihedral angle was analysed with a Gaussian mixture model (GMM) by pooling the data from the ten independent simulations. Statistics of the models are presented in Table 3.

Starting with His40, the RMSD is 0.66–0.73 Å with SEs of 0.01–0.04 Å for the five SSs with this residue in the HIE state (SS1 and SS4–SS7). However, SS3 with His40 in the HIP state gave the lowest RMSD, 0.53±0.02 Å, indicating that this is actually the proper protonation state of this residue. On the other hand, the HID protonation state in SS2 gave a much larger RMSD than all the other SSs, 1.03±0.04 Å, indicating that this is an unfavourable state. This is in agreement with the hydrogen-bond analysis.

It is conceivable that an unfavourable interaction between an erroneous protonation state of a His residue and its surroundings may be relived by the movement of the surrounding residues, rather than of the His residue itself. Therefore, we have also calculated the RMSD values (relative to the starting crystal structure) of all residues within 3.5 Å of the His side chain. We will call this the residue RMSD, RRD, in the following. The RRD in each SS are presented in the third part of Table 2. It can be seen that SS2 (with His40 in the HID state) still gives the largest deviation, 0.98±0.03 Å, in agreement with the RMSD and hydrogen-bond analysis, indicating the HID state is unfavourable for His40. However, SS1 and SS7 gave only slightly smaller RRD (0.97±0.03 Å). Moreover, SS3 only gave the second lowest RRD value. Thus, the RRD results are less conclusive than the RMSD results.

The centre of the GMM distributions of the φ dihedral angle (μ in Table 3) for His40 was 72–84°, whereas the angle in the crystal structure is 84°. This shows that in most SSs, the φ angle is slightly less than in the crystal structure. The standard error of μ was small, less than 1° for all simulations. The only exception is SS5, for which there was a 1% probability that φ becomes –7° rather than 83°, caused by a single flip of this residue (to –70°) in one of the ten simulations. This also influences the uncertainty of μ for the second state. The width of the distributions (σ in Table 3) is 10–14°, except for the minor distribution of SS5. The narrowest

distribution was that of SS3, in which His40 is in the HIP state, again indicating that this state may be preferable over the other states. This is also the only state for which μ was the same as the dihedral angle in the crystal structure. Therefore, we conclude that HIP is the most probable protonation state for His40. Moreover, the RMSD and RRD analyses pointed out the HID state to be unfavourable, in agreement with the hydrogen-bond analysis.

Next, we considered the His57 residue. For the five SSs with His57 in the HIP state, the RMSD was 0.53–0.68 Å with SEs of 0.01–0.03 Å, and the RRD was 0.86–0.99 Å with SEs of 0.03–0.04 Å. On the other hand, SS5 with His57 in the HIE state gave much higher RMSD and RRD of 1.81±0.08 Å and 1.12±0.04 Å, respectively, confirming that this is a highly unfavourable state. SS4, with His57 in the HID state, gave an intermediate RMSD of 0.84±0.01 Å, indicating that this state is also less favourable than the HIP state. The RRD is also slightly higher for this set, 1.01±0.03 Å, but in this case, the difference is not statistically significant.

The centre of the GMM distributions of the φ angle for His57 was –84° to –93°. This is slightly less than the angle of –98° observed in the crystal structure. The only exception was SS5 (with His57 in the HIE state), in which the side chain had flipped so that the distribution was centred on 106°. The width of the distribution for SS5 was also significantly wider than for the other sets, 16°, compared to 10–12° (±0.3–0.5°) . This clearly shows that HIE is unfavourable for His57, in agreement with the hydrogen-bond analysis. On the other hand, SS4, with His57 in the HID state gave results ($\mu = -92°$ and $\sigma = 11°$) that are similar to those of the other SSs with His57 in the HIP state. Still, considering the results of the RMSD analysis, the HIP state seems to be most likely for His57.

Finally, for His91, the five SSs with His91 in the HIE state gave RMSDs of 0.64–0.74 Å with SEs 0.02 Å and RRDs of 0.73–0.76 Å with SEs of 0.02 Å. On the other hand, SS6 and SS7, with this residue in the other two protonation states, gave significantly higher RMSDs, 1.01–1.11 Å with SEs of 0.03–0.04 Å, as well as significantly higher RRDs 1.03±0.04 Å. Hence, it seems clear that the HIP and HID states are unfavourable for His91, in accordance with the hydrogen-bond analysis.

Similarly, the centre of the GMM distribution ranged from –85° to –89° for SS1–SS5, whereas SS6 and SS7 has a significantly different centre of –97°. The crystal structure has a φ dihedral angle of –85°. It is also clear that the widths of the distribution for SS6 and SS7, 13–14°, are significantly larger than those of the other sets (10–12° with SEs of 0.3–0.5°). Thus, for this residue, all three criteria indicated that this residue should be in the HIE state.

In conclusion, for trypsin, the MD simulations gave conclusive protonation states for all three His residues, viz. HIP for His40 and His57, and HIE for His91. Thus, the simulations extend the results of the hydrogen-bond analysis by providing conclusive assignments also for His40 and 57. The assignments of His57 and 91 are in agreement with that of PROPKA and REDUCE, and at least one of the suggestions of What_Check. However, the assignment for His40 is in disagreement with all these three softwares, but in agreement with the hydrogen-bond analysis.

An interesting question is how long the MD simulations need to be to give converged predictions regarding the protonation states of the His residues. Therefore, we have also performed the same analysis for five times longer trajectories (i.e.1 ns simulations after the equilibration). The results of the RMSD and RRD analyses are presented in Table S1 in the (supplementary data. It can be seen that the RMSD results differ by 0.04 Å on average (in absolute terms), which is dominated by a difference of 0.17 Å for SS5 of His57, whereas the second largest difference is 0.08 Å. In fact, these differences follow quite closely the estimated standard errors and in only two cases (SS3 for His40 and SS1 for His57) the differences are statistically significant at the 95% level. This is only one more than expected (at 95% confidence level, 1.05 out of 21 samples should give a significant difference by chance), so there is no clear indication that the results obtained after 1 ns sampling is different from those obtained after 200 ps. This conclusion is supported by the RRDs, for which the

differences are even smaller, 0.02 Å on average (absolute), with a maximum of 0.08 Å and none of the differences is statistically significant at the 95% level. In particular, all conclusions obtained by the 200 ps simulations are still valid after 1 ns sampling.

The GMM results in Table S2 (supplementary data) are somewhat more different. In particular, the second model observed for His40 in SS5 has almost disappeared (π has dropped to below 0.005). On the other hand, two new models are observed for His91 in SS6 and SS7, with $\pi = 0.05$ and 0.10. Moreover, His57 in SS5 shows a skewed distribution, probably indicating two overlapping Gaussians with rather close μ values. These differences reflect modified sampling of rare events, which of course is strongly affected by the length of the simulation. For the other distributions, the differences between the two sampling times are modest, e.g. 0.9° mean absolute difference for μ (maximum 2.4°) and 0.5° average difference for σ (maximum 1.4°). Three of the differences for μ but none of the (difference for σ are statistically significant at the 95% level. However, all conclusions based on the GMM models for the φ dihedral angle in trypsin from the 200 ps simulations are still valid after 1 ns simulation. In fact, the extra models for SS6 and SS7 of His91 and the skewed distribution of His57 in SS5 only further emphasize that these protonation states are unfavourable.

Finally, we also tried to reduce the simulation time to 100 ps. The RMSD and RRD results from these simulations are also included in Table S1. It can be seen that the mean absolute differences to the 1 ns results increases to 0.05 and 0.04 Å for the RMSD and RRD, respectively. In particular, the differences now are statistically significant for six of the 21 RMSD values. This indicates that 100 ps simulation time is too short to obtain fully reliable results. Consequently, we will only discuss results for 200 ps simulation times for the other two proteins.

Galectin-3

Galectin-3 is a member of a lectin family that controls the intra- and extracellular trafficking and localisation of certain glycoproteins in human cells, which in turn affect cell signalling, adhesion, and differentiation [26]. We have studied the carbohydrate-binding domain of this protein, with a potent inhibitor, benzamido-N-acetyllactoseamine. This protein contains four His residues, His158, 208, 217, and 223. We started by performing an analysis of the solvent accessibility and hydrogen-bond pattern of these residues in the crystal structure [9]. The results are shown in Table 4 and Figure 3.

It can be seen that HD1 of His158 makes a hydrogen bond to Asp148 (1.7 Å distance). Hence, ND1 must be protonated, making the HIE protonation state unlikely. HE2 forms a hydrogen bond to an OH group of the ligand, which can both accept and donate hydrogen bonds, depending on the direction of the proton. Therefore, we cannot conclude whether His158 is protonated or not on the NE2 atom, based on the hydrogen-bond pattern.

The ND1 atom of His208 makes a hydrogen bond to a water molecule (1.6 Å distance). Water can also both donate and accept hydrogen bonds, so the protonation state of the ND1 cannot be decided. The NE2 atom also forms a hydrogen bond to a water molecule (2.4 Å). However, the disordered residue Glu205 is close to this atom and in one of the two reported conformations in the crystal structure, there is a hydrogen bond between HE2 and OE2 of 1.7 Å. In that conformation, His208 must be protonated on the NE2 atom. However, in the MD simulations, only a single conformation of Glu205 can be simulated, and all simulations were based on the other conformation, in which Glu205 is not close enough to form a hydrogen bond to His208.

For His217, both side-chain N atoms form hydrogen bonds to water molecules (1.9 and 2.0 Å distances). Therefore, the protonation state of this residue cannot be decided from the hydrogen-bond pattern.

Finally, ND1 of His223 also forms a hydrogen bond to a water molecule. However, HE2 forms a hydrogen bond to the backbone O atom of Glu205. Therefore, we can conclude that

the NE2 atom should be protonated, making the HID state unlikely.

The PROPKA calculations did not give much information on the protonation states, apart from that all histidines most likely are deprotonated (all predicted pKa values were less than 7.0). What_Check assigned the HID state to His158 and the HIE state to the other three residues, in accordance with the hydrogen-bond analysis. Finally, REDUCE assigned a HID state to His158 and 217, and a HIE state to His208 and 223.

For galectin-3, NMR data are available for the protonation of some of the His residues (Weininger, U.; Akke, M., personal communication). These data are also included in Table 4 and it can be seen that it completely supports the assignments made by the hydrogen-bond analysis and the three softwares, showing that His158 is in the HID state, whereas His223 is in the HIE state. For the other two His residues, the experimental data were not conclusive.

Based on these results, we set up nine sets of MD simulations that differed in the protonation state of the four His residues, as is shown in the first section of Table 5. In SS1, we assumed that His158 is in the HID state, whereas the other three are in the HIE state. The other eight SSs were obtained by systematically changing the protonation state of each of the four His residues to the other possible states. For each SS, ten independent MD simulations were started. The analyses are presented in Tables 5 and 6.

For His158, the nine SSs gave RMSDs of 0.48–0.61 Å with SEs of 0.01–0.03 Å. The largest RMSD (0.61 Å) was obtained for SS3, in which His158 is in the HIE state, in agreement with the hydrogen-bond analysis. However, SS9 (with His158 in the standard HID state) gave the same RMSD, so the results are not conclusive. Interestingly, the HIP state of His158 in SS2 gave the smallest RMSD of all nine setups (0.48±0.01 Å), but this value is close to the second lowest one (0.49±0.01 Å), so it may be a coincidence.

All SSs gave RRDs of 0.70–0.81 Å with SEs of 0.02–0.04 Å, except for the two SSs in which His158 is not in the HID state. SS3 gave the highest RRD of 1.16 Å, indicating again that the HIE state of His158 is unfavourable. In particular, Asp148, which forms a hydrogen bond to His158 in the crystal structure, showed an increased RMSD, 0.94 Å, compared to 0.42–0.54 Å in the other eight setups, in agreement with a broken hydrogen bond when His158 is in the HIE state. Interestingly, SS2 gave also a large RRD of 0.98±0.03 Å, indicating that also the HIP state of His158 is unfavourable, in agreement with the NMR experiments. The increased RRD is caused mainly by Arg143, which swings out into solution in all the simulations.

The centres of the φ GMM distributions were 76–86° with SEs of up to 1° (Table 6). The angle in the crystal structure is in between these values, 79°. The widths of the distributions were 11–15°, except for SS3, for which the width is 19°, indicating again that the HIE state is unfavourable for His158. Thus, all three measures point out the HIE state as unfavourable, whereas they are not fully conclusive concerning the HIP state.

His208 shows an appreciably larger RMSD in all simulations than His158, 0.89–1.18 Å with SEs of 0.02–0.05 Å. This reflects that this residue is more solvent exposed than His158. The deviating protonation states in SS4 and SS5 gave RMSDs in the middle of this range.

His208 also gave larger RRD values than His158, 1.32–1.77 Å, with SEs of 0.03–0.08 Å. The largest RRD (1.90±0.08 Å) was found for SS5. A detailed examination of the latter results shows that it caused mainly by the movement of Glu205, which shows two conformations in the crystal structure. Thus, the simulations confirm that this residue is flexible and that the HID protonation state of His208 may stabilise an alternative conformation of this residue. Moreover, SS4 gave the smallest RRD (1.32±0.05 Å), which may indicate that HIP is a more proper protonation state of this residue, but the difference is not fully significant.

The centres of the GMM distributions for His208 ranged from –106° to –115°. The φ dihedral angle in the crystal structure is –100°, i.e. slightly smaller than in the simulations. For SS5 and SS8, two additional distributions were found, around –48° and –30°, but the probability of these distributions were less than 1%. However, for SS9, there is a 9%

probability that the His208 side chain has flipped with a φ dihedral of 70° (observed in one of the ten simulations). This also affects the spread of the distribution, which (is significantly wider than for the other simulations. The narrowest distribution is found for SS4, in which His208 is in the HIP state, but the difference to SS2 and SS7 is not significant. Thus, there some indications that the HIP state may be favourable for His208, but the results are far from conclusive.

For His217, the RMSDs in the nine SSs were intermediate between the two other His residues, 0.70–0.85 Å with SEs of 0.02–0.04 Å. SS6 with His217 in the HID state gave the lowest RMSD, and SS7 with His217 in the HIP state gave the third lowest RMSD. This might indicate that the HIE state is not the most favourable state for this residue, but the differences are not fully significant.

His217 gave RRDs of 1.28–1.43 Å with SEs of 0.04–0.07 Å, without any statistically significant differences between any of the setups. In particular, SS6 and SS7 with the deviating protonation states were in the middle of the observed RRDs.

The centres of the GMM distributions for the φ dihedral of His217 were 55–65° and the widths were 16–21° for the seven SSs with His217 in the HIE state. However SS6 and SS7 with His217 in the HID and HIP states, respectively, gave significantly larger φ dihedral angles of 69° and 75°, respectively, and a smaller width of 13°. This might indicate that HID, which gave the φ dihedral angle closest to the crystal structure (69°) is the most favourable state for this residue. Thus, the three measures gave somewhat diverging results, but there is an indication that the HID state might be more favourable than the HIE state in the simulations.

Finally, for His223, the RMSDs were 0.56–0.71 Å with SEs of 0.01–0.03 Å, except for SS8, which stands out with the largest RMSD of 0.82±0.05 Å. This indicates that the HID state is unfavourable for His223, confirming the hydrogen-bond analysis and the experimental assignment. SS9 (with His223 in the HIP state) had the fourth largest RMSD.

The RRDs were 1.09–1.35 Å with SEs of 0.03–0.06 Å. Quite unexpectedly, SS8 gave only the third largest RRD and Glu205 did not show any particularly high RMSD for this setup (1.34 Å, compared to 1.27–2.05 Å for the other setups). This shows that removal of the His223–Glu205 hydrogen bond in the HID state is compensated only by a rotation of the His residue in this case, not by a significant reorganisation of the surroundings. The large RRD for SS5 (1.35±0.05 Å) is caused by the unfavourable HID state of His208, because His208 and 223 are close in space and share three residues in the RRD.

The centres of the φ dihedral GMMs of His223 ranged from –34° to –37° for the seven SSs with His223 in the HIE state. This is less than the dihedral angle observed in the crystal structure, –54°. For SS8 and SS9, the centres of the distribution were –51° and –47°, respectively. The width of the distribution for SS8 (32±1°) was much larger than for the other sets (12–15°). This clearly shows that the HID state is unfavourable for His223. Thus, the MD simulations point out the HID state of His223 as highly favourable, in accordance with the hydrogen-bond analysis, but they give no indication that HIP state should be unfavourable.

Consequently, we can conclude that the simulation data clearly pointed out that the HIE state is unfavourable for His158 and the HID state is unfavourable for His223, in agreement with the hydrogen-bond analysis and experiments. There are also some indications that the HIP state may be unfavourable for His158, and that the HIP and HID states may be most favourable for His208 and 217, respectively.

Hisactophilin

Hisactophilin is a small protein with 118 amino acids, out of which 31 are histidines. They are known to have similar pK_a values and the protein acts as a pH sensor, binding actin at low pH [10]. All His residues are on the surface of the protein and only a few of them form unambiguous hydrogen bonds with other residues. In fact, as can be seen from Table 7, there

is only one putative hydrogen bond that is shorter than 2 Å (for His100 ND1) and only six additional hydrogen bonds are shorter than 2.5 Å (for the other two proteins in Tables 1 and 4, the longest hydrogen bonds to protein residues is 2.04 Å; note that the calculations are based on a NMR structure with no reported water molecules[10]), so there are very little information regarding the protonation state of these residues. PROPKA suggests that all residues are deprotonated at pH 7, except His100 (pK_a = 7.1). What_Check predicts that all His residues are in the HIE state, although a HID state is also possible for the His28, 31, 79, 90, 100, and 109 residues. Reduce suggests that all residues are in the HIE state, except residues His28, 31, 91, 100, and 107, which are suggested to be in the HID state. Considering these uncertain and partly conflicting results, we simply run three sets of simulations: One with all His residues in HID state, one with all of them in the HIE state, and the last with all of them in the HIP state. These simulations provide information of the behaviour of His residues on the surface of the protein.

The RMSD of all 31 His residues are presented in Table 8. It can be seen that all His residues show large RMSDs from the starting structure, with values varying from 1.4–7.4 Å (3.3. Å on average), i.e. (much) larger than the RMSD for any His residue and protonation state in the other two proteins, with a single exception (His57 for SS5 of trypsin). The standard errors are 0.04–0.4 Å (0.2 Å on average). It can also be seen that the RMSD depends more on what residue is considered than on the protonation state; for example, His35 always has a low RMSD (1.4–1.9 Å), whereas that of His28 is high (5.9–6.8 Å). However, there are also many large differences between the various protonation states. In fact, 53 of the possible 93 pairwise comparisons of two different protonation states gave statistically significant differences at the 95% level. For eleven of the His residues, one protonation state gave significantly smaller RMSD than the other two protonation state (marked in bold face in Table 8) and can therefore be considered to be the most favourable state. These are the HID state for His27, 30, 35, 79, and 106, the HIE state for His48 and 78, and the HIP state for His 71, 91, 98, and 100.

Most of the His residues in hisactophilin are close to each other on the surface of the protein. Therefore, no RRD was calculated because there would be too many residues shared two or more His residue, making the analysis ambiguous.

We also analysed the φ dihedral angles of the His residues in hisactophilin with the GMM and the results are shown in Table 9. It can directly be seen that they also show a much larger variation than for the other two proteins. In fact, all 31 His need more than one Gaussian to describe the distribution in at least one of the three protonation states, indicating that the side chain has flipped once during the simulation. Moreover, the widths of the distributions are in general large, 9–99°, with averages of 25±11°, 27±13°, and 25±10° for the HID, HIE, and HIP simulations, respectively (the great majority of the distributions for the other two proteins had widths of 10–15°). This reflects that the His residues are on the surface of the protein, without any pronounced hydrogen bonding. The SEs of the widths are in general 1–2°. Only 12 of the central values of the distributions are within 10° of the dihedral angle observed in the starting structure [10]. However, considering that we started from an NMR structure, this is not too unexpected. Still, for 12 of the His residues, one of the three protonation states had a significantly narrower distribution than the other two protonation states, indicating that this is the more favourable protonation state. These are the HID state for His27, 31, 39, 58, 78, and 106, the HIE state for His100, and the HIP state for His33, 48, 75, 97, and 107. They are marked in bold face in Table 9.

Unfortunately, the RMSD and φ dihedral angle analysis give the same prediction of the most stable residue for only two residues, His27 and 106. Moreover, all suggestions disagree with at least one of the PROPKA, What_Check, or REDUCE analyses, except for His31, for which all methods (including the analysis of the hydrogen-bond pattern) indicate that the HID state is more stable. The RMSD analysis also gives this results, although the difference to the HIE state is not statistically significant (owing to a large uncertainty of the HIE state).

Therefore, we have to conclude that the MD simulations of hisactophilin do not provide much information about the protonation state of the His residues. This is probably a typical result for His residues on the protein surface.

Conclusions

We have studied how the protonation of histidine residues affect the protein structure and dynamics during MD simulations, with the aim of detecting erroneous protonation state from the results of MD simulations. Previous studies [4,5,6] have focused on predicting the protonation state before the MD simulations are started (which of course is preferable), but the decision is in many cases ambiguous and different software give conflicting results, as has been seen in Tables 1, 4, and 7. We test whether additional information from by MD simulations can help the decision and pinpoint erroneous selections. Three proteins were studied, hisactophilin, galectin-3, and trypsin.

We have analysed three different measures: the width of the distribution of the His CA–CB–CG–ND1 dihedral angle using GMMs, the RMSD of the His residue compared to the starting structure, and the RMSD of all residues within 3.5 Å of the His residue during the MD simulations (RRD). The GMM was found to be more discriminatory than other more simple measures of the spread of the φ dihedral, e.g. the range. The results also showed that in all simulations, the φ dihedral distribution could be well described by 1–3 Gaussians. Our results indicate that with a combination of these three measures, erroneous protonation states can be identified by unusually wide dihedral distributions, or unusually large His RMSD and RRD values. Each measure alone cannot point out all erroneous protonation states, but using all three measures together, an unambiguous decision could often be reached.

However, the decisions in this paper are based on a comparison of simulations with all three possible protonation states of each His residue. Unfortunately, there does not seem to be any specific thresholds for the three measures that can be used to accept or reject the results. For example, for trypsin, acceptable widths of the φ dihedral distributions seem to be 10–12°, whereas erroneous protonation states were characterised by widths of 12–16°. For galectin-3, the φ widths seem to be somewhat larger: 10–15° and 16–32°, respectively. However, for the solvent-exposed His residues in hisactophilin, even the minimum φ widths among the three possible protonation states were 13–30°, strongly overlapping with unfavourable ranges for the other two proteins. It is possible that some discriminatory power can be obtained by considering buried and solvent-exposed His residues separately, but for reliable results, it is most likely necessary to run all three protonation states.

For trypsin, the MD simulations allowed assignment of a single protonation state for all three residues, although the hydrogen-bond analysis was ambiguous for two of them. For galectin-3, our results confirm the hydrogen-bond analysis regarding the protonation state of the His residues, but they also provide some further, although less certain, information about the protonation state of three of the His residues. It might be somewhat disappointing that the analysis was not conclusive, but it should be remembered that the experiments were also ambiguous for two of the His residues, indicating they actually may be in a mixture of several protonation states.

For the solvent-exposed His residues in hisactophilin, essentially no unambiguous information about the protonation states could be obtained. Of course, this is disappointing considering the high cost of MD simulations. This indicates that the MD simulations are quite insensitive to the protonation state of the His residues, besides when there are direct hydrogen bonds to the His residues. On the other hand, this is a very important conclusion, showing that the problematic and often quite arbitrary choice of the protonation state of the His residues has a minor influence of the structure of the protein when the hydrogen-bond analysis does not give any conclusive results. This is also in accordance with the observation that ligand-binding free energies are insensitive to the protonation of His residues in the protein, except

when they are in direct contact with the ligand [27]. Consequently, that the analysis of the His protonation may be concentrated around a possible site of central interest.

Acknowledgements

This investigation has been supported by grants from the Swedish research council (project 2010-5025) and from the FLÄK research school in pharmaceutical science. The simulations were performed on computer resources provided by the Swedish National Infrastructure for Computing (SNIC) at Lunarc, Lund University. We thank Ulrich Weininger and Mikael Akke for providing us the NMR assignment of the His protonation state in galectin-3 and for fruitful discussions.

References

[1] G.M. Ullmann, E.W. Knapp, Electrostatic models for computing protonation and redox equilibria in proteins, Eur. Biophys. J. (28 (1999) 533-551.

[2] E. Alexov, E.L. Mehler, N. Baker, A.M. Baptista, Y. Huang, F. Milletti, J.E. Nielsen, D. Farrell, T. Carstensen, M.H.M. Olsson, J.K. Shen, J. Warwicker, S. Williams, J.M. Word, Progress in the prediction of pKa values in proteins, Proteins, 79 (2011) 3260-3275.

[3] M.H.M. Olsson, C.R. Sondergaard, M. Rostkowski, J.H. Jensen, PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions, J. Chem. Theory Comput, 7 (2011) 525-537.

[4] R.W.W. Hooft, G. Vriend, C. Sander, E.E. Abola, Errors in protein structures, Nature 381 (1996) 272-272.

[5] J.M. Word, S.C. Lovell, J.S. Richardson, D.C. Richardson, Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation, J. Mol. Biol., 285 (1999) 1735-1747.

[6] G.F. Signorini, R. Chelli, P. Procacci, V. Schettino, Energetic Fitness of Histidine Protonation States in PDB Structures, J. Phys. Chem. B, 108 (2004) 12252-12257.

[7] H.-K.S. Leiros, S.M. Mcsweeney, A.O. Smalås, Atomic resolution structures of trypsin provide insight into structural radiation damage, Acta Crystallogr., Sect.D 57 (2001) 488

[8] P. Mikulskis, S. Genheden, P. Rydberg, L. Sandberg, L. Olsen, U. Ryde, Binding affinities of the SAMPL3 trypsin and host-guest blind tests estimated with the MM/PBSA and LIE methods, J. Comput.-Aided Mol. Design, 26 (2012) 527-541.

[9] C. Diehl, O. Engström, T. Delaine, M. Håkansson, S. Genheden, K. Modig, H. Leffler, U. Ryde, U. Nilsson, M. Akke, (Protein flexibility and conformational entropy in ligand design targeting the carbohydrate recognition domain of galectin-3, J. Am. Chem. Soc. 132 (2010) 14577-14589.

[10] J. Habazettl, D. Gondol, R. Wiltscheck, J. Otlewski, M. Schleicher, T.A. Holak, Structure of hisactophilin is similar to interleukin-1b and fibroblast growth factor, Nature, 359 (1992) 855-858

[11] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, (Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters, Proteins (65 (2006) 712-725.

[12] J.M. Wang, R.M. Wolf, K.W. Caldwell, P.A. Kollman, D.A. Case, Development and Testing of a General Amber Force Field, J. Comput. Chem., 25 (2004) 1157-1174

[13] C.I. Bayly, P. Cieplak, W.D. Cornell, P.A. Kollman, (J. Phys. Chem., A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model, 97 (1993) 10269-10280

[14] B.H. Besler, K.M. Merz, P.A. Kollman, (Atomic charges derived from semiempirical methods, J. Comput. Chem. 11 (1990) 431-439

[15] S. Genheden, U. Ryde, (How to obtain statistically converged MM/GBSA results, J. Comput. Chem., 31 (2010) 837-846

[16] W.L. Jorgensen, J. Chandrasekhar, (J.D. Madura, (R.W. Impley, M.L. Klein, (Comparison of simple potential functions for simulating liquid water, J. Chem. Phys. 79 (1983) 926-935

[17] D.A. Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, M. Crowley, R.C. Walker, W. Zhang, K.M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvary, K.F. Wong, F. Paesani, J. Vanicek, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak G. Cui, D.H. Mathews, M.G. Seetin, C. Sagui, V. Babin, P.A. Kollman, Amber 10, University of California, San Francisco, 2008.

[18] X. Wu, B.R. Brooks, (Self-guided Langevin dynamics simulation method, Chem. Phys.

Lett. 381 (2003) 512-518
[19] H.J.C. Berendsen, J.P.M. Postma, W.F. Van Gunsteren, A. Dinola, J.R. Haak, Molecular dynamics with coupling to an external bath, J. Chem. Phys. 81 (1984) 3684–3690
[20] T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N -log(N) method for Ewald sums in large systems, J. Chem. Phys. 98 (1993) 10089-10092
[21] J.P. Ryckaert, G. Ciccotti, H.J.C. Berendsen, Numerical integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of *n*-Alkanes, J. Comput. Phys. 23 (1977) 327-341
[22] J. Bowers, B. Devolder, L. Yin, T. Kwan, A maximum likelihood method for linking particle-in-cell and Monte-Carlo transport simulations, Comput. Phys. Commun. 164 (2004) 311-317
[23] S. Genheden, C. Diehl, M. Akke, U. Ryde, Starting-condition dependence of order parameters derived from molecular dynamics simulations, J. Chem. Theory Comput., 6 (2010) 2176-2190
[24] A.P. Dempster, N.M. Laird, M.D. Rubin, Maximum likelihood from incomplete data via EM algorithm, J. R. Stat. Soc. B 39 (1977) 1-38
[25] B.Efron, Nonparametric estimates of standar error- the jackknife, the bootstrap and other methods, Biometrika 68 (1981) 589-599.
[26] F.-T. Liu, G.A. Rabinovich Galectins: regulators of acute and chronic inflammation, Ann. N. Y. Acad. Sci. 1183 (2010) 158-182.
[27] S. Genheden, U. Ryde, A comparison of different initialisation protocols to obtain statistically independent molecular dynamics simulations, J. Comput. Chem., 32 (2011) 187-195

Table 1. Results of an analysis of the surroundings of the His residues in trypsin [7]. The solvent accessibility (SA, in %, based on the number of heavy atoms around the residue, divided by 115 [3]) and possible hydrogen-bond partners to the HD1 and HE2 atoms, with distances in Å, are listed for each residue. The results of calculations with the PROPKA (presented as estimated pK_a values) [3], What_Check [4], and REDUCE [5] softwares are also listed.

Residue	SA	HD1	HE2	Conclusion	PROPKA	What_Check	REDUCE
40	14	1.71 OG Ser32	1.99 O Gly193	not HID	5.1	HID or HIE	HIE
57	26	1.72 OD2 Asp102	1.97 OG Ser195	not HIE	7.4 (HIP)	HIE or HIP	HIP
91	33	2.04 N Ser93	1.83 Wat	HIE	4.9	HIE	HIE

Table 2. Results from the trypsin simulations. The first part of the Table shows the protonation states of the three His residues in the seven simulation setup sets (SS1–SS7). The second part shows the RMSD of the three His residues with the respect to the starting crystal structure [7], whereas the third part shows the average RMSD of all residues within 3.5 Å of each His residue (RRD). All results are presented as averages and standard errors over 10 independent simulations for each protonation setup. Particularly high values are marked in bold face and low values in bold and Italics.

	His states			RMSD			RRD		
	40	57	91	40	57	91	40	57	91
SS1	HIE	HIP	HIE	0.72±0.04	0.68±0.01	0.64±0.02	0.97±0.04	0.94±0.04	0.73±0.02
SS2	**HID**	HIP	HIE	**1.03**±0.04	0.56±0.02	0.69±0.02	**0.98**±0.03	0.87±0.04	0.73±0.02
SS3	**HIP**	HIP	HIE	***0.53***±0.02	0.62±0.02	0.71±0.02	0.78±0.04	0.86±0.04	0.76±0.02
SS4	HIE	**HID**	HIE	0.71±0.02	**0.84**±0.01	0.68±0.02	0.86±0.04	**1.01**±0.03	0.75±0.02
SS5	HIE	**HIE**	HIE	0.72±0.02	**1.81**±0.08	0.74±0.02	0.78±0.02	**1.12**±0.04	0.75±0.02
SS6	HIE	HIP	**HID**	0.66±0.01	0.59±0.03	**1.01**±0.03	0.75±0.02	0.99±0.04	**1.03**±0.03
SS7	HIE	HIP	**HIP**	0.73±0.03	0.53±0.01	**1.11**±0.04	0.97±0.04	0.92±0.03	**1.03**±0.04

Table 3. Gaussian mixture models for the φ dihedral angle in trypsin. Particularly poor values are marked in bold face and good values in bold and Italics.

	His40		His57		His91	
	μ	σ	μ	σ	μ	σ
SS1	72.1±0.7	14.0±0.6	-83.7±0.5	10.6±0.4	-85.2±0.5	10.4±0.4
SS2	75.6±0.6	13.1±0.5	-89.6±0.6	12.3±0.4	-85.0±0.5	9.9±0.4
SS3	***84.2***±0.5	***10.1***±0.3	-91.1±0.6	11.0±0.4	-87.7±0.5	10.5±0.3
SS4	81.9±0.7	13.0±0.5	-91.9±0.5	10.5±0.4	-88.5±0.6	11.8±0.5
SS5[a]	83.0±0.7	13.7±0.5	**106.3**±0.8	**16.2**±0.5	-88.7±0.5	10.8±0.4
	-7.1±4.3	44.8±2.4				
SS6	81.2±0.6	12.4±0.5	-89.3±0.5	10.0±0.3	**-97.4**±0.7	**13.8**±0.5
SS7	80.5±0.6	12.6±0.5	-92.5±0.5	9.9±0.4	**-97.4**±0.6	**13.2**±0.5
Crystal	84.2		-97.8		-84.6	

[a] The first and second models for His40 of SS5 have π = 0.99 and 0.01, respectively.

Table 4. Results of an analysis of the surroundings of the His residues in galectin-3 [9]. The solvent accessibility (SA; cf. Table 1) and possible hydrogen-bond partners to the HD1 and HE2 atoms, with distances in Å, are listed for each residue. In addition, experimental assignments of the protonation state, based on NMR measurements are indicated (Weininger, U.; Akke, M., personal communication), as well as the results of calculations with the PROPKA (presented as estimated pK_a values) [3], What_Check [4], and REDUCE [5] softwares.

Residue	SA	HD1	HE2	PROPKA	What_Check	REDUCE	Exp.
158	15	1.70 OD1 Asp148	1.87 O19 Ligand	6.0	HID	HID	HID
208	39	1.64 Wat	2.38 Wat 1.72 OE2 Glu205[a]	6.4	HIE	HIE	?
217	44	1.88 Wat	1.98 Wat	5.8	HIE	HID	?
223	6	2.12 Wat	1.95 O Glu205	5.6	HIE	HIE	HIE

[a] Only for one of two reported conformations of Glu205 in the crystal structure.

Table 5. Results from the galectin-3 simulations. The first part of the Table shows the protonation states of the four His residues in the nine setup sets (SS1–SS9). The second part shows the RMSD of the four His residues with the respect to the starting crystal structure, whereas the third part shows the average RMSD of all residues within 3.5 Å of each His residue (RRD). All results are averages and standard errors over ten independent simulations for each protonation setup. Particularly large values are marked in bold face and particularly small values in bold and Italics face.

	His state				RMSD				RRD			
	158	208	217	223	158	208	217	223	158	208	217	223
SS1	HID	HIE	HIE	HIE	0.50±0.01	1.05±0.03	0.79±0.03	0.56±0.02	0.76±0.03	1.61±0.05	1.43±0.06	1.19±0.04
SS2	**HIP**	HIE	HIE	HIE	***0.48***±0.01	0.97±0.03	0.78±0.03	0.62±0.02	**0.98**±0.03	1.64±0.03	1.32±0.04	1.17±0.03
SS3	**HIE**	HIE	HIE	HIE	**0.61**±0.02	0.97±0.04	0.74±0.02	0.64±0.02	**1.16**±0.06	1.42±0.07	1.33±0.06	1.14±0.05
SS4	HID	**HIP**	HIE	HIE	0.52±0.03	0.98±0.04	0.77±0.04	0.68±0.01	0.81±0.04	***1.32***±0.05	1.36±0.05	1.09±0.05
SS5	HID	**HID**	HIE	HIE	0.54±0.01	1.04±0.05	0.83±0.03	0.65±0.02	0.77±0.02	**1.77**±0.07	1.39±0.07	1.35±0.05
SS6	HID	HIE	**HID**	HIE	0.49±0.01	1.00±0.02	***0.70***±0.02	0.71±0.02	0.70±0.02	1.61±0.05	1.38±0.05	1.22±0.03
SS7	HID	HIE	**HIP**	HIE	0.58±0.02	1.18±0.05	0.76±0.02	0.66±0.03	0.76±0.03	1.62±0.06	1.32±0.06	1.30±0.05
SS8	HID	HIE	HIE	**HID**	0.57±0.01	0.89±0.02	0.85±0.03	**0.82**±0.05	0.75±0.02	1.40±0.07	1.28±0.06	1.25±0.05
SS9	HID	HIE	HIE	**HIP**	0.61±0.02	0.97±0.10	0.76±0.03	0.68±0.03	0.81±0.04	1.49±0.08	1.39±0.05	1.16±0.06

Table 6. Gaussian mixture models for the φ dihedral-angle distributions in galectin-3.

	His158		His208		His217		His223	
	μ	σ	μ	σ	μ	σ	μ	σ
SS1	82.1±0.6	13.0±0.4	-113.0±0.6	12.1±0.4	65.0±1.0	20.8±0.9	-34.3±0.7	14.7±0.6
SS2	85.6±0.6	12.3±0.4	-108.6±0.5	11.1±0.4	60.7±0.8	16.1±0.6	-34.3±0.7	14.7±0.6
SS3	84.9±1.0	**18.8**±0.6	-110.1±0.7	13.2±0.6	65.0±1.0	20.8±0.9	-36.5±0.6	13.2±0.5
SS4	80.5±0.7	14.0±0.6	-113.2±0.5	***10.2***±0.4	59.2±0.9	19.8±0.9	-34.0±0.6	12.0±0.4
SS5	85.3±0.6	12.2±0.4	-107.3±0.6	11.5±0.4	59.2±0.9	19.8±0.9	-33.9±0.7	14.3±0.5
SS6	75.6±0.7	14.8±0.6	-113.2±0.6	12.4±0.6	***69.1***±0.7	***13.3***±0.6	-33.9±0.7	14.3±0.5
SS7	82.1±0.7	12.6±0.5	-111.7±0.5	11.0±0.4	75.2±0.6	***12.7***±0.4	-36.5±0.7	14.0±0.5
SS8	85.6±0.5	10.6±0.3	-114.6±0.8	15.0±0.8	54.9±0.8	16.0±0.7	-51.2±1.5	**32.2**±1.3
SS9[a]	82.8±0.6	12.7±0.5	-106.3±2.7	20.4±3.6	54.9±0.8	16.0±0.7	-47.0±0.7	13.7±0.5
			70.2±2.7	35.5±3.6				
Crystal	79.2		-100.2		68.9		-53.7	

[a] The first and second models for His208 of SS9 have π = 0.91 and 0.09, respectively.

Table 7. Result of an analysis of the surroundings of the His residues in hisactophilin [10]. The solvent accessibility (SA; cf. Table 1) and possible hydrogen-bond partners to the HD1 and HE2 atoms, with distances in Å, are listed for each residue. In addition, results are presented for calculations with the PROPKA (presented as estimated pK_a values) [3], What_Check [4], and REDUCE [5] softwares.

Residue	SA	HD1	HE2	PROPKA	What_Check	REDUCE
9	81	2.0 N His10		6.3	HIE	HIE
10	83			6.1	HIE	HIE
12	39	2.6 O Gly11	2.9 O His97	5.7	HIE	HIE
24	20	3.0 O Phe13		6.0	HIE	HIE
25	72			5.8	HIE	HIE
27	90			6.3	HIE	HIE
28	57			6.8	HIE or HID	HID
30	94			6.3	HIE	HIE
31	55	3.0 O Asp29	2.3 NH1 Arg4	5.4	HIE or HID	HID
33	66			6.0	HIE	HIE
35	66			5.9	HIE	HIE
39	43	2.7 NE2 His68		5.9	HIE	HIE
48	88			6.2	HIE	HIE
58	93	2.2 NZ Lys59		6.4	HIE	HIE
65	80			6.0	HIE	HIE
66	82			6.3	HIE	HIE
68	79			6.2	HIE	HIE
71	49		2.2 O Ile55	6.0	HIE	HIE
75	65			6.0	HIE	HIE
78	62			6.7	HIE[a]	HIE
79	60	2.9 N Gly80		6.1	HIE or HID	HIE
88	78			6.2	HIE	HIE
89	76			6.0	HIE	HIE
90	66			6.2	HIE or HID	HIE
91	26	2.5 O His88		4.7	HIE	HID
97	66			6.1	HIE	HIE
98	94			6.4	HIE	HIE
100	64	1.7 O Gly99	2.5 O Lys59 & Val101	7.1 (HIP)	HIE or HID	HID
106	86			6.4	HIE	HIE
107	60		2.4 OG Ser84	5.7	HIE	HID
109	86	2.7 OD2 Asp108		6.6	HIE or HID	HIE

[a] What_Check predicts that this residue should be flipped.

Table 8. RMSDs of the 31 His residues in hisactophilin with the respect to the starting structure. The results are averages and standard errors over ten independent simulations for each protonation setup. RMSDs that are significantly lower for one protonation state than for the other two at the 95% level are marked in bold face.

His	HID	HIE	HIP
9	3.1±0.1	3.0±0.1	3.1±0.1
10	3.3±0.2	3.1±0.2	4.3±0.2
12	2.0±0.1	2.3±0.1	2.2±0.1
24	2.2±0.0	2.0±0.1	2.3±0.1
25	3.8±0.2	3.2±0.0	3.4±0.1
27	**3.1**±0.1	4.6±0.2	3.7±0.1
28	6.8±0.3	6.3±0.3	5.9±0.2
30	**2.9**±0.1	4.3±0.2	3.4±0.1
31	3.3±0.1	3.9±0.3	4.7±0.2
33	2.1±0.1	2.5±0.2	2.7±0.1
35	**1.4**±0.0	1.7±0.1	1.9±0.0
39	2.7±0.1	3.2±0.3	2.5±0.4
48	3.4±0.2	**2.8**±0.2	4.3±0.3
58	2.7±0.3	3.0±0.2	4.1±0.2
65	3.2±0.1	3.3±0.1	4.2±0.1
66	3.3±0.1	2.3±0.2	2.7±0.2
68	2.9±0.1	2.6±0.2	2.3±0.2
71	3.3±0.1	4.3±0.1	**2.9**±0.1
75	2.2±0.2	2.2±0.1	2.3±0.1
78	3.5±0.3	**2.7**±0.2	3.5±0.2
79	**2.6**±0.2	3.9±0.3	3.3±0.2
88	2.4±0.2	2.8±0.3	2.8±0.2
89	3.0±0.2	3.3±0.4	4.3±0.2
90	2.6±0.1	2.5±0.1	2.2±0.1
91	2.7±0.1	2.5±0.1	**2.1**±0.1
97	6.8±0.4	4.1±0.2	3.6±0.3
98	7.4±0.2	7.0±0.3	**2.9**±0.2
100	4.0±0.1	3.4±0.1	**2.8**±0.1
106	**2.6**±0.1	4.1±0.2	3.2±0.1
107	2.5±0.1	2.4±0.1	3.1±0.1
109	2.5±0.2	3.8±0.1	2.1±0.2

Table 9. Gaussian mixture models for the φ dihedral angles in hisactophilin. Distributions with significantly smaller widths than for the two other protonation states are marked in bold face.

	HID			HIE			HIP			Exp.
	π	μ	σ	π	μ	σ	π	μ	σ	
His9	0.94	-93	34	1.00	43	20	1.00	-69	22	-58
	0.07	151	14							
His10	0.50	-105	34	0.90	-75	17	1.00	-58	27	19
	0.51	86	29	0.10	23	22				
His12	1.00	124	32	0.60	-118	31	0.72	-70	26	-100
				0.40	26	20	0.28	134	27	
His24	1.00	84	19	1.00	123	19	0.90	-16	24	-10
							0.10	131	18	
His25	0.80	-76	26	1.00	57	21	1.00	78	18	104
	0.20	121	24							
His27	1.00	63	**16**	0.33	-98	26	0.52	-110	29	115
				0.67	115	30	0.48	84	38	
His28	0.40	-83	22	1.00	2	99	0.78	-68	29	76
	0.60	78	23				0.22	73	25	
His30	1.00	-29	39	1.00	30	22	0.53	-106	21	100
							0.24	23	43	
							0.22	120	23	
His31	1.00	137	**20**	0.84	-70	28	0.32	-53	38	-60
				0.16	134	22	0.68	132	19	
His33	0.70	-115	18	1.00	-94	31	1.00	-89	**13**	-156
	0.30	102	30							
His35	1.00	38	16	0.30	-101	22	1.00	90	17	-78
				0.70	46	24				
His39	1.00	78	**14**	0.31	-57	49	0.89	-79	18	-100
				0.69	88	22	0.11	58	31	
His48	0.40	-67	39	0.20	-129	21	1.00	56	**21**	20
	0.60	70	25	0.80	56	34				
His58	1.00	90	**14**	0.84	-76	20	0.10	-119	17	-130
				0.16	53	23	0.90	71	24	
His65	0.10	-111	21	0.10	-114	22	0.32	-97	26	-129
	0.90	96	18	0.90	58	29	0.69	100	22	
His66	0.10	-107	14	0.50	-103	23	0.28	-97	28	-107

	0.90	80	23	0.50	115	35	0.72	64	39	
His68	0.70	-94	23	0.06	-32	40	0.52	-79	27	139
	0.30	67	17	0.94	120	29	0.48	129	23	
His71	1.00	75	16	0.50	-45	34	1.00	105	15	-126
				0.50	87	13				
His75	0.10	-87	21	0.30	-93	22	1.00	84	**19**	27
	0.90	77	23	0.70	76	23				
His78	1.00	-81	**19**	0.76	-58	42	1.00	-89	27	-22
				0.24	93	30				
His79	1.00	79	21	0.25	-58	51	1.00	76	21	2
				0.75	65	24				
His88	1.00	-112	14	0.10	-118	18	1.00	-106	15	43
				0.90	51	24				
His89	0.90	-62	18	0.83	-66	24	0.05	-129	9	126
	0.10	109	54	0.17	90	40	0.95	77	25	
His90	0.65	-71	21	0.89	-63	33	1.00	-61	26	-100
	0.35	41	70	0.11	56	28				
His91	1.00	47	31	0.22	-155	20	1.00	-142	30	-103
				0.78	99	31				
His97	0.44	-105	31	0.99	-124	22	1.00	109	**16**	31
	0.56	99	31	0.01	162	10				
His98	0.64	-89	24	0.20	-117	25	1.00	-97	31	-99
	0.36	70	35	0.80	73	24				
His100	1.00	87	20	1.00	93	**12**	0.84	-68	33	-84
							0.16	126	24	
His106	1.00	112	**18**	1.00	64	29	0.70	-67	27	69
							0.30	110	20	
His107	0.05	-68	29	0.90	-90	18	1.00	-99	**14**	-131
	0.95	53	18	0.10	85	18				
His109	0.35	-36	48	0.10	-65	18	0.24	22	56	135
	0.65	100	30	0.90	74	22	0.76	116	20	

Figure 1.The three possible protonation states for the His residue, HID (left), HIE (middle), and HIP (right).

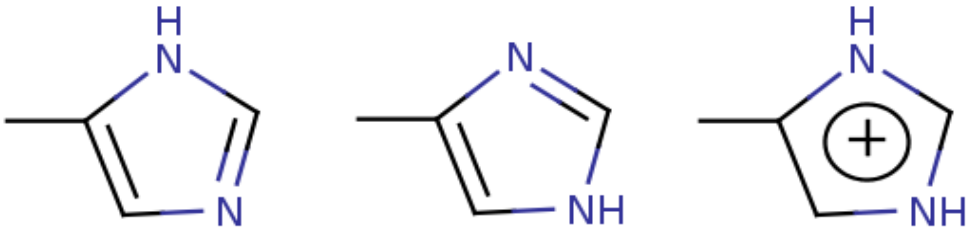

Figure 2. The three His residues in trypsin and their hydrogen-bond partners. Hydrogen atoms available in the pdb file are shown as well.

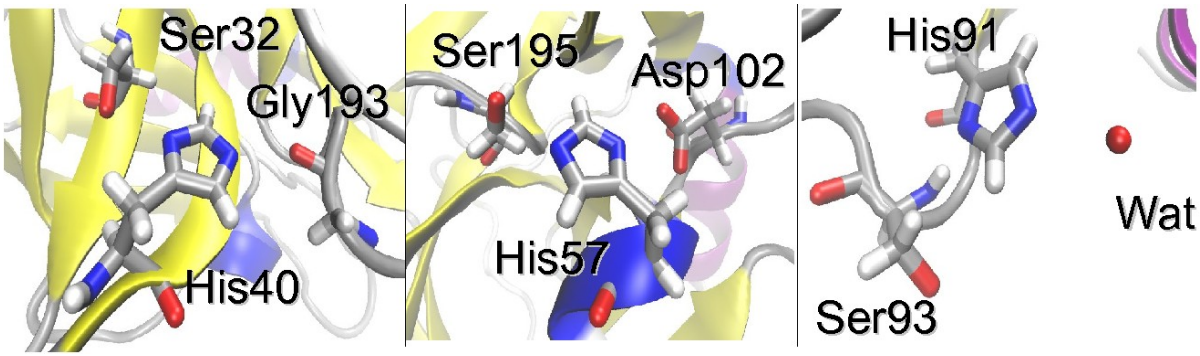

Figure 3. The four His residues in galectin-3 and their hydrogen-bond partners. The two alternative conformations of Glu205 are shown. The ligand is denoted L02.