THE JOURNAL OF PHYSICAL CHEMISTRY B

ARTICLE

pubs.acs.org/JPCB

A Molecular Mechanism of Hydrolysis of Peptide Bonds at Neutral pH Using a Model Compound

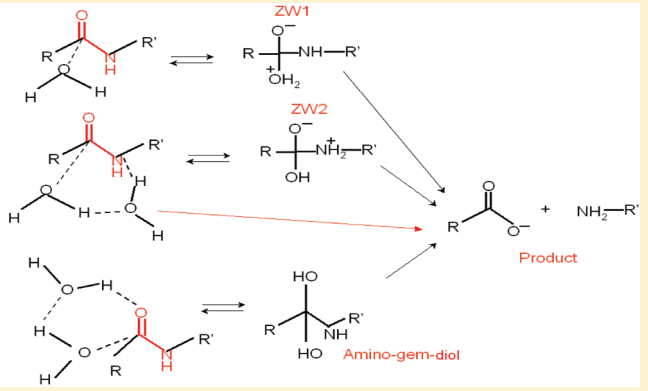
Bin Pan,[†] Margaret S. Ricci,[‡] and Bernhardt L. Trout*[,†]

[†]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

[‡]Department of Process & Product Development, Amgen Inc., Thousand Oaks, California 91320, United States

Supporting Information

ABSTRACT: The covalent stability of peptide bonds is a critical aspect of biological chemistry and therapeutic protein applications. In this computational study, the hydrolytic reaction of peptide bonds at neutral pH was studied using a model compound, *N*-MAA. The most probable reaction pathway and intermediate(s) involved are controversial in previous studies. In addition, most previous computational studies focus on the energetics of chemical species involved, rather than providing a dynamic picture of the reaction process in aqueous conditions. However, fluctuations at finite temperatures are quite important, as we show. Thus, a path sampling method was used to generate an ensemble of trajectories according to their statistical weights in trajectory space. An *ab initio* molecular dynamics technique was applied to advance the time of the reaction in order to collect trajectories. The likelihood maximization procedure and its modification were used in extracting dynamically relevant degrees of freedom in the system, and approximations of the reaction coordinate were compared. It was found that this hydrolytic reaction is very complex because it involves many degrees of freedom. The reaction coordinate C−O distance previously assumed was found to be inadequate in describing the dynamic progress of the reaction. In addition to affecting atoms directly involved in bond-making and -breaking processes, the water network also has determining effects on the hydrolytic reaction, a fact which is manifest in the expression of the best one-dimensional reaction coordinate that we found, which includes five geometric quantities. p_B histograms were computed to verify the results of the likelihood maximization and to evaluate the accuracy of our best reaction coordinate to the "true" reaction coordinate. The relation with previous suggested reaction pathways and intermediate(s) is discussed in terms of computational system, method, and accuracy.



INTRODUCTION

Hydrolytic reactions are ubiquitous in nature. For example, in biological systems, hydrolysis is important in fundamental processes such as the conversion among ATP, ADP, and AMP used in energy metabolism and storage and the decomposition of proteins, fats, oils, and carbohydrates, etc. A typical hydrolytic reaction involves the fragmentation of a parent molecule by the addition of a water molecule: one fragment obtains a hydrogen atom, and the other obtains a hydroxyl group. However, this picture is oversimplified; it does not reflect the influence of the changes of an interacting water network when a participating water molecule is split apart, nor does it explicitly consider the reorganized water network due to the resulting products whose polarity and solubility are different from their parent molecule. It is therefore important for us to study the fundamentals of hydrolytic reactions with the whole solvent network included.

Hydrolysis of peptide bonds on the polypeptide backbone,[1] in particular, is a major route of the covalent degradation of proteins. Therefore, it has both much biological significance and timely technological significance. Monoclonal antibodies have been reported by several different research groups to undergo nonenzymatic fragmentation in the hinge region.[2−4] This finding was identified mainly, for example, as the hydrolysis of several peptide bonds in the hinge.

There have been numerous experimental and computational studies on the hydrolysis of peptide bonds in aqueous solutions due to its important biological relevance. To elucidate enzyme proficiencies in catalyzing peptide bond hydrolysis, a number of groups extensively characterized the reaction rates of hydrolysis of peptide bonds.[1,5−7] It was found that, in general, the reaction rate at neutral pH is extremely slow, with a half-time corresponding to hundreds of years. However, this reaction occurs much faster in both acidic and basic conditions, as studied by Smith et al.[8] They extensively mapped the pH and concentration

Received: August 13, 2010
Revised: March 28, 2011
Published: April 19, 2011

dependencies of reaction rates for the hydrolytic reaction of a peptide bond in *N*-(phenylacetyl)glycyl-D-valine (PAGV) and identified three regimes of reaction mechanisms: an acid-catalyzed mechanism for pH < 4 with a first-order reaction rate in concentrations of both PAGV and H^+, a base-catalyzed mechanism for pH > 10 with a first-order reaction rate in concentrations of both PAGV and OH^-, and a water-mediated mechanism for pH values in between with a first-order reaction rate only in the concentration of PAGV.

A number of computational efforts[9−11] were devoted to understanding the energetics of reactants, products, and intermediate species involved in the hydrolytic reaction. However, these studies showed only a static picture of the reaction process. Many other analyses included dynamic simulations. Stranton et al.[12] used combined quantum mechanical and classical mechanical calculations to study the hydrolytic reaction of peptide bonds catalyzed by trypsin in solution; free energies of the reaction were reported. Cascella et al.[13] studied the hydrolytic reaction of formamide using steered CPMD, concluding that the reaction in neutral pH involves a tetrahedral intermediate characterized by a charge separation, essentially a zwitterionic species.

Zahn et al.[14−18] analyzed the hydrolytic reaction of *N*-MAA (*N*-methyl acetamide) in various solution conditions, namely, acidic, basic, and neutral pHs. In these *ab initio* calculations of the potential of mean force, reaction coordinates were assumed and the projection of the free energy hyper-surface onto these coordinates was performed using constrained molecular dynamics. However, the assumed reaction coordinates were never tested to determine if they were the correct ones. In the study of the hydrolytic reaction of *N*-MAA under neutral pH conditions,[17] a zwitterionic species different than Cascella et al.'s study[13] was found to be a stable intermediate formed by a concerted reaction process of water molecules attacking the peptide bond. However, no free energy barrier between this intermediate and the product state was reported therein. The author also noted that the scanning velocity used in Cascella et al.'s study[13] is probably too large to capture the rearrangement of the hydrogen-bonded network with a time-scale of picoseconds. However, essentially both studies used some scanning approach to study the high potential energy region of the phase-space in obtaining the free energy profile over an arbitrarily predetermined reaction coordinate.

Gorb et al.[19] summarized the various probable pathways of hydrolytic reaction for formamide, the simplest compound containing a peptide bond. They particularly noted that in the gas phase the unlikely zwitterionic species obtained in Zahn's study[17] can be greatly stabilized and thus is one possible intermediate in the liquid phase. However, they found that the amino-gem-diol species formed by the addition of a water molecule to the carbonyl double bond is more likely to be an intermediate for the stepwise water-assisted hydrolysis of formamide, rather than the energetically less favorable pathway of the stepwise water-assisted hydrolysis through the zwitterionic intermediate in Zahn's study. This study aims to resolve, or at least analyze in detail, some of those controversies in the literature about the most probable reaction pathway and intermediate(s) involved.

Even though the hydrolytic reaction occurs slowly at pH 4−10, hydrolysis of therapeutic antibodies can be significant in physiological conditions during circulation or within the formulation solution conditions over shelf life. Cordoba et al.[3] reported a several-percent hydrolytic degradation of residues in the hinge region of IgG1 antibody molecules over an incubation period of three months. Their study showed that the hydrolysis of the polypeptide backbone could occur at multiple sites in the hinge to varying extents. They also showed that the reaction was nonenzymatic around neutral pH. Using optimized reversed-phase methods coupled with mass spectrometry, Dillon et al. reached similar conclusions about the location and extent of hydrolytic cleavage of peptide bonds in the hinge region for IgG2 antibodies.[4] Furthermore, the resulting fragments subsequently associated to form high molecular weight species via a clip-mediated aggregation mechanism,[20,21] which can be the primary degradation pathway at elevated temperatures for IgG2 antibodies. Thus, from a scientific and practical standpoint, elucidation of the underlying mechanism of the hydrolytic reaction is essential.

Distinct mechanisms of the hydrolytic reaction have been proposed in the literature for different solution pHs. In both acidic and basic conditions, the hydrolytic reaction is accelerated by the protonation of the carbonyl oxygen or the addition of a hydroxyl group onto the carbonyl carbon in the peptide bond, respectively. However, under neutral pH conditions (with a pH range of 4−10, for example, on a dipeptide model), the hydrolytic reaction is extremely slow, having an exceptionally high reaction barrier of 27−30 kcal/mol, extrapolated from the experimental data.[3,6,8] We are interested in this "neutral pH" range (pH 4−10) for therapeutic proteins, since it represents the most relevant pH conditions for formulation applications and physiological environment.

In this work and a companion work,[22] we studied the hydrolytic reaction of a peptide bond under neutral pH and acidic pH conditions, respectively, using *N*-methyl acetyl acrylamide (*N*-MAA) as a model compound, in which two methyl groups are the minimal but computationally tractable constituents on the peptide bond −NH−CO−. *Ab initio* molecular dynamics simulations at finite temperature equipped with transition path sampling (TPS),[23−25] likelihood maximization,[26−28] and p_B histogram analysis[25] techniques were utilized to gain an understanding of the reaction mechanism, including identification and verification of the reaction coordinate.

What we mean by the reaction coordinate here is a descriptor of the real physical progress of the transition from the reactant state to the product state. The reaction coordinate as defined has to be contrasted with order parameters, which serve mainly to distinguish the two ending states. An order parameter is a quantity whose value can be used to distinguish different thermodynamic states, for example, reactant or product states, and crystalline, amorphous, or liquid states. A reaction coordinate has to be an order parameter, while the contrary is not necessarily true. Identification of the "correct" reaction coordinate from a collection of order parameters is a very challenging problem, especially when the transition is complex. However, the information about the "correct" reaction coordinate is necessary when computing quantities of interest, such as free energy barriers and reaction rate constants, from molecular simulations. Also, we believe that the knowledge of the "correct" reaction coordinate, and hence the reaction mechanism, can provide essential information on the molecular level for judicious engineering of complex systems.

The committor probability $p_B(\mathbf{x})$, which can be interpreted as the probability that a trajectory initiated with Maxwell−Boltzmann distributed momenta at the configurational

vector **x** reaches the product state B before reaching the reactant state A, can be used to best describe the mechanism of the transition during an activated process, thus serving as the "true" reaction coordinate.[29] As illustrated by Metzner et al.[30] and E et al.,[31] various quantities involving $p_B(\mathbf{x})$, such as the probability current of reactive trajectories and the average frequency of reactive trajectories, allow one to fully characterize the statistical properties of the transition trajectories in the trajectory space, and thus to compute quantities of interest such as reaction rate constants. However, in general, $p_B(\mathbf{x})$ is costly to compute and is a function of the highly dimensional configurational vector **x**; it provides no insight into the physical characteristics about the transition dynamics. Therefore, approximations of $p_B(\mathbf{x})$ with a lower dimensional (preferably a one-dimensional) descriptor, involving physical quantities such as bond lengths, dihedral angles, bonding number,[32] density fluctuation,[33] etc., are required to describe the transition process in providing essential physical insights.

We have addressed the problem of searching for an accurate reaction coordinate by combing the path sampling technique, developed originally by Chandler and co-workers,[23−25] and a statistical tool, likelihood maximization, recently developed by Peters and Trout.[26−28] These powerful techniques have allowed us to generate a large number of dynamic trajectories for the hydrolytic reaction, and to determine what physical degrees of freedom are important in describing the reaction progress. Our approach goes far beyond previous theoretical/computational studies of chemical reactions in complex systems, as we do not guess a putative reaction coordinate and assume that it is correct but systematically explore millions or more of possible reaction coordinates to find the best one. We have found that the assumed reaction coordinates that have been published for this system, and presumably related systems, are not good approximations to accurate reaction coordinates and have demonstrated our approach to be invaluable in providing physical insight into the reaction mechanism. Such insight can aid in better understanding complex chemical processes, in addition to addressing real-world problems, such as the degradation of biopharmaceuticals. Overall, our approach exemplified here in analyzing the complex chemical reaction in a condensed-matter system using path sampling techniques combined with techniques for the determination of the reaction coordinate is quite general; the approach can be directly applied to other types of transitions in which direct transition dynamics is infrequent and the transition states are short-lived.

OVERVIEW

At finite temperatures, no single transition trajectory, a series of points in the phase-space connecting the reactant and product states, can be used to describe the whole reaction process due to thermal fluctuations. In trajectory space, each transition path is associated with a statistical weight,[25] contributing to the experimentally observed transition process. It is therefore necessary to collect an ensemble of reactive trajectories and calculate the probability distribution according to their statistical weights in order to calculate quantities of interest that can be compared with experiments. The technique that implements this as a Monte Carlo procedure in trajectory space is TPS.[23−25]

Transition path sampling is well-suited for studying transition processes that have time-scale separation and exhibit a rough potential energy surface. Time-scale separation is often the characteristic of activated processes, where two stable states, the reactant state A and the product state B, are separated by a large free energy barrier. In rare events, such as chemical reactions, formation of critical nuclei during crystallization processes, and the protein-folding process, the time scale for the system to wander in the valley of the stable states is much longer than the time scale in which the transition dynamics occur. Therefore, a direct molecular simulation starting from a stable basin is very computationally inefficient in collecting reactive trajectories. Another issue that makes the study of rare events more complicated is the roughness of the potential energy landscape. For systems in the gas phase, the potential energy surface has only a few saddle points, which usually can be used to sufficiently characterize the transition process. In contrast, for rare events occurring in solution, the system of interest has a rugged potential energy surface on which myriads of small energy barriers with heights of the order of k_BT must be distinguished, with the true potential energy barrier often larger than k_BT. One way to circumvent the challenges posed by the time scale separation and the roughness of the potential energy surface is to focus, as the TPS technique does, on the dynamic bottleneck for the rare event which is defined as the transition state surface. TPS starts with a predetermined transition path connecting two stable states, and then by making shooting and shifting moves in the trajectory space using a Monte Carlo procedure, TPS can eventually lead to the true dynamics at transition states and an ensemble of physically meaningful pathways. Eventually, complete reaction profiles, transition states, free energies of reaction barrier, and reaction rates can be obtained.

The prerequisites to perform a transition path sampling are two: (1) definitions of two stable states and (2) an initial trajectory, or a series of points in phase space, which connect the two stable states. This initial trajectory may not be physical or at the same condition as the one of interest. The most appealing feature of TPS is that no prior knowledge about the reaction mechanism, the reaction coordinate, and the transition state is needed to begin with.

A more efficient sampling technique to explore the trajectory space is the aimless-shooting algorithm,[26,27] a variant of the TPS algorithm. As verified in this work, it has the advantage of having more decorrelation between successive trajectories than the original version of TPS and thus can explore the trajectory space more quickly. The detailed implementation of the algorithm can be found in Peters et al.[26,27]

METHODOLOGY

Pseudopentials and CPMD Validation. Even though a number of computational studies[13−18,34,35] used pseudopotentials and the Car−Parrinello MD with a plane-wave basis to study chemical transformations in molecular systems, no sufficient and detailed comparison was made between results obtained from CPMD simulations and those obtained using high-level large-basis-set quantum mechanical calculations in order to evaluate the suitability of applying CPMD to study these systems. Here, we performed calculations in order to make a detailed comparison. The transition state structure with a single imaginary frequency corresponding to the hydrolytic reaction in a molecular system with one *N*-MAA and two water molecules was first located using the Gaussian 03 (G03) quantum mechanical package.[36] Then, an internal reaction coordinate (IRC) calculation was performed to obtain a series of geometric configurations along the reaction pathway on the minimum potential energy pathway.

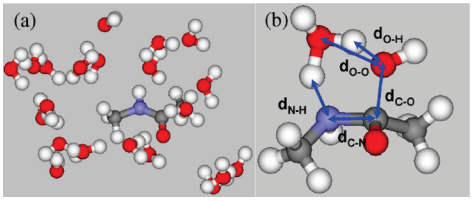


Figure 1. Simulation box (a) together with bond distances used to define basins of stable states (b).

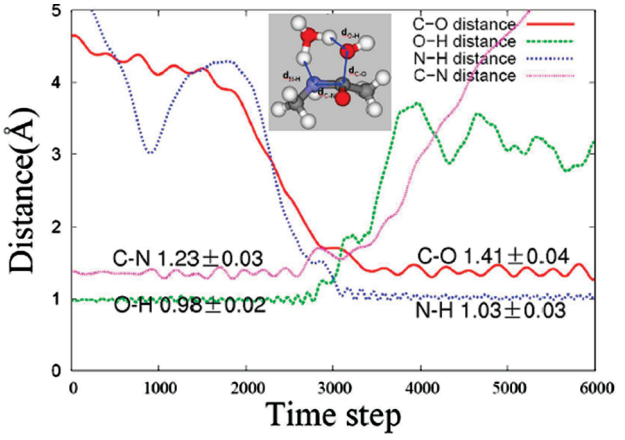


Figure 2. Transition trajectory with the associated changes in the OPs of bond distances.

The IRC calculations also served to verify that the TS structure indeed corresponds to the hydrolytic reaction, corresponding to the water-assisted concerted mechanism named in Gorb et al.'s study.[19] Out of consideration for accuracy and computational speed, the TS search and frequency calculations together with IRC calculations were all performed with the theory-level/basis-set B3LYP/6-31+G(d,p). Then, for each geometric configuration obtained in the IRC output, we performed quantum mechanical calculations using G03 with theory-level/basis-set B3LYP/cc-pVTZ and MP2/cc-pVTZ, as well as CPMD simulations for such an isolated system with the same exchange-correlation functional, pseudopotential, and wave function cutoff value as in the following section, and a cubic cell with dimensions 12 Å × 8 Å × 8 Å. No frequency calculations were obtained, due to the high memory requirement at the level of MP2 with the very large basis-set cc-pVTZ. Therefore, no zero-point energy corrections were included.

System Description. A simulation setup very similar to Zahn's potential-of-mean-force calculations[17] was chosen for our path sampling and analysis but with a few important differences. First, we performed an equilibration MD simulation by running long trajectories using the classical CHARMM force field[37] under the ambient temperature and pressure. This simulation was done using the CHARMM package[38] under the NPT ensemble. The force field parameters for *N*-MAA were taken from similar structures, and the geometry of *N*-MAA was fixed. The TIP3P force field parameters[39] were used for the water molecules in the system. The final equilibrated system had 1 *N*-MAA and 21 water molecules in a 12.27 Å × 8 Å × 8 Å simulation cell, as Figure 1 shows. We found that the value of the water density in Zahn's studies was too high, while the water density values used in these studies had the correct value at ambient temperature and pressure. Next, long (∼50 ps) constant-temperature MD trajectories were run with the C−O, O−H, and N−H distances as shown in Figure 1 constrained in the Car−Parrinello molecular dynamics (CPMD)[40] simulation using the CPMD package[41] in order to equilibrate other degrees of freedom.

Table 1. Ranges of Bond Distances in Figure 2 Used for Definitions of Basins of Stable States[a]

	fluctuation					
	reactant		product		definition	
	mean	std	mean	std	reactant	product
C−O dist (Å)			1.40	0.04	(2.07, +∞)	[1.27, 1.58]
O−H dist (Å)	0.98	0.05			[0.82, 1.14]	(1.68,+∞)
H−N dist (Å)			1.03	0.08	(1.79, +∞)	[0.79, 1.26]

[a] A configuration corresponds to a particular stable state (either reactant or product) only when three bond distances are simultaneously within the specified ranges.

Stable Basin Definitions. As shown in Figure 1, three distances (the C−O distance, the O−H distance, and the N−H distance) were used to describe to which stable state a particular configuration corresponds. Long molecular dynamic trajectories with neither constraints nor restraints were run using CPMD to obtain the fluctuations in these bond distances. Next, an appropriate range was taken by computing the variances of these bond distances in the stable state regions with adjustments such that, in the later aimless shooting procedure, no returning to indeterminate regions is made once the system reaches one stable basin (shown in Figure 2). The mean values of these bond distances and their fluctuations are listed in Table 1, which also gives the choice of basin definitions. During our path-sampling procedure, we observed no basin overlapping situation, in which a particular configuration satisfied both the reactant and product definitions.

Order Parameters. Order parameters were proposed based on extensive screening combined with physical intuition into the reaction mechanism. More specifically, a systematic procedure for including candidate OPs was used as follows. The set of candidate OPs includes the distances of all possible pairs within the system and cosine values of angles and dihedral angles for all possible triplets and quadruplets, respectively, selected from all atoms around the midpoint of the C−N bond within a 6.5 Å cutoff. The reason for using cosine values of these angles and dihedral angles instead of their absolute values is to avoid possible discontinuities when they take on values at the boundaries of their range. This procedure generated up to a total of 3 241 875 candidate order parameters. Other types of OPs used in previous studies, such as bonding number[32] and density fluctuation,[33] were not considered here. Our approach is to express the reaction coordinate as a function of configurational variables only and to validate that approach a posteriori using committor probability analyses. Exhaustive one-OP-variable screening in likelihood maximization was done for this vast set of candidate order parameters. A combinatorial problem arises even sooner when going to the two-OP-variable situation. The solution used in this study will be discussed shortly.

Aimless Shooting. The aimless shooting algorithm, a kind of transition path sampling, as described by Peters and Trout,[26,27] was applied to harvest an ensemble of independent trajectories according to their statistical weights. As with the other transition path sampling

algorithms,[23−25] the aimless shooting algorithm requires (1) accurate definitions of the basins of stable states and (2) an initial reactive trajectory, i.e., a trajectory that connects the stable basins.

In general, there are various ways to obtain the first reactive trajectory. It is also emphasized that different approaches to obtain the first reactive trajectory, which does not even need to be a physical one or one under the same conditions, do not matter in TPS algorithms. As long as the random walk sampling in the trajectory space approaches the "equilibrium" stage before one starts to save trajectory data used for analysis, the influence of the first reactive trajectory should have completely gone. Our approach to obtaining the first reactive trajectory is the following. We used the transition state structure (one *N*-MAA with two water molecules) determined in the G03 calculations described previously as a starting point. Then, we solvated this structure with water molecules and ran a long-enough (∼100 ps) classical MD on such a solvated system using CHARMM, as described previously, but keeping the key geometric quantities in the reaction core fixed, namely, the bond distances of the C−O, O−H, and N−H (see Figure 1). Afterward, the final configuration of the equilibrated system was used to perform a wave function optimization in CPMD. Then, a constrained MD simulation with a length of ∼2 ps was used to fully relax all the other degrees of freedom in the system in order to remove potential artifacts introduced when fixing these three distances. Both forward and backward trajectories shot from the equilibrated configuration were then obtained with a set of assigned velocities drawn from the Maxwell−Boltzmann distribution. Repeated shootings were performed until a reactive trajectory was obtained, since the basins of stable states were already defined. This resulting initial trajectory also provides some information on how fast the transition dynamics takes place, based on which the appropriate overall length of MD steps (600 with a time step of 4 au) can be derived.

The two-point version of the aimless shooting algorithm was carried out in the following way. Two configurations close to the hypothesized transition state were selected from the initial reactive trajectory, and one of the two was chosen randomly from which forward and backward half-trajectories were shot. Momenta for forward shooting were generated from a Maxwell−Boltzmann distribution such that there were no net linear and angular momenta for the whole system. Momenta for backward shooting were the reverse of those for forward shooting. The two configurations have a time displacement Δt, which is an adjustable parameter and needs to be carefully set. If the forward and backward half-trajectories are combined to give a reactive trajectory, this new trajectory was accepted and the two configurations with a time displacement Δt to the previous shooting point were recorded as a new two-point from which the shooting procedure was repeated.

As described by Beckham et al.,[42] the time displacement Δt has to be chosen appropriately to yield an acceptance ratio between 40 and 60%. If it is too large, the algorithm tends to go too far away from the transition-state region, leading to a low acceptance rate with many consecutive unaccepted trajectories; if it is too small, the aimless shooting algorithm will be very inefficient in exploring the shooting-point configuration space, and therefore more trajectories will be needed to obtain a good approximation to the reaction coordinate. For chemical reactions in which bond-breaking and -forming steps are involved, Δt is expected to be smaller than in more diffusive systems, since transitions driven by strong interactions are short in terms of transition duration.

Dynamic trajectories were collected using the CPMD package[41] in the NVT ensemble. A time step of 4 au (∼0.1 fs) and an electron fictitious mass of 400 au were used. A chain of four Nose−Hoover thermostats was used to control the temperature at 300 K. The molecular orbitals were described by a plane wave basis with an energy cutoff of 70 Ry. Norm-conserving Troullier−Martins pseudopotentials[43] and BLYP density functionals were used. We found that selecting from two points, $\mathbf{x}_{-\Delta t}$ or $\mathbf{x}_{+\Delta t}$, is sufficient to sample the transition state ensemble.

In the aimless shooting procedure, the trajectory length is set to be as short as possible in order to save computational time, resulting in the possibility of generating inconclusive trajectories, for which at least one end point in the forward and backward half-trajectories does not lie in any basin of stable state. A half-trajectory step of 600 was found to maintain the level of inconclusive trajectories at or below 10%. A time displacement, Δt, of 15 au was chosen to yield an acceptance rate of 44.7%. A total of 1650 trajectories were collected for later analysis.

In order to know how efficiently the trajectory space was sampled, the autocorrelation function was computed to describe how independent successive trajectories are. The configurations of shooting points in the aimless shooting procedure form an ordered series, which can be treated as a time series. Each configuration that led to a reactive trajectory was aligned with a reference configuration by a best-fit procedure on the reaction core part (*N*-MAA and the two attaching water molecules). Then, the C−O distance (no need to align) and the root-mean-square-deviation of the reaction core part were calculated, followed by the calculation of the normalized autocorrelation function. As Figure SI 4 of the Supporting Information shows, both descriptors show that essentially no memory exists in the following shooting trajectory. The fact that aimless shooting has essentially no correlation from one step to the next is presumably because of the complete renewal of momenta in each MC move in exploring the trajectory space.

Likelihood Maximization. As described in Peters and Trout,[27] the reaction coordinate, r, is modeled as a linear combination of candidate OPs, denoted as $\mathbf{q}$, with α_0 through α_m as adjustable coefficients:

$$r(\mathbf{q}) = \alpha_0 + \sum_{k=1}^{m} \alpha_k q_k \tag{1}$$

The choice of a linear combination is for purposes of convenience only, and a nonlinear reaction coordinate expression could be chosen.

The model for the committor probability $p_B(r)$ was chosen to be

$$p_B(r) = \frac{1}{2}[1 + \tanh(r)] \tag{2}$$

Note that we have found that the choice of eq 2 to model committor probability is not essential. The shape of the p_B function plotted against the reaction coordinate r just needs to satisfy the requirement that $p_B(r = 0) = 1/2$, $p_B(r \to -\infty) = 0$, and $p_B(r \to +\infty) = 1$.

This committor probability model was used to maximize the likelihood function with respect to the set of coefficients α_i's $(i = 0, \ldots, m)$

$$L(\alpha) = \prod_{\mathbf{x}_k \to B} p_B(\mathbf{x}_k) \prod_{\mathbf{x}_k \to A} [1 - p_B(\mathbf{x}_k)] \tag{3}$$

only using outcomes of forward half-trajectories. In principle, if an ensemble of candidate OPs is proposed, the maximization of likelihood (eq 3) should be performed over all combinations of OPs to determine the best reaction coordinate according to the models of eqs 1 and 2. However, when the set of candidate OPs is large, a combinatorial problem arises for exhaustive screening of the best reaction coordinate. One is forced to reduce the size of the set of candidate OPs when the number of OP variables m increases. One choice for overcoming this combinatorial problem is to do one-OP-variable exhaustive screening and use the best OPs for higher m searching, as was used in this study. For the best approximate reaction coordinate, the approximate transition state isosurface can be obtained by setting $p_B(r) = 1/2$. This occurs at $r = 0$, so setting $r(q) = 0$ defines the approximate transition state isosurface.

As shown in other examples, such as the alanine-dipeptide,[33] a very complex reaction coordinate might be involved in our system. In order to minimize or eliminate bias by any assumption about which OPs are important, an exhaustive but systematic approach was taken to find the best reaction coordinate model in likelihood maximization. Over 3 million candidate OPs were screened individually first. In order to tackle the algorithmic complexity problem when more than one OP variable $(d > 1)$ is included in the likelihood maximization procedure, the following systematic approach was adopted. Basically, it assumes that important OPs previously screened based on likelihood scores will also be important in comprising reaction coordinate models with a larger number of OPs. It starts with the best m one-OP-variable reaction coordinate models. Then, in each round for d-OP-variable optimization, every best ranked $(d - 1)$-OP-variable result is supplemented with every best m one-OP variable to give a d-OP-variable model. Next, only the n best d-OP-variable results are retained for $(d + 1)$-OP-variable screening. This way, each round has roughly $m \times n$ optimization problems to solve.

Uncertainty Analysis in Likelihood Maximization. The criterion with which to discriminate different reaction coordinate models $r(q)$ is the Bayesian information criterion (BIC),[27] which equals $\ln(N)/2$, where N is the number of accepted trajectories in aimless shooting. If the difference in two log-likelihood scores is greater than the BIC, the reaction coordinate model having a higher log-likelihood score is superior in describing the transition process; on the other hand, if the difference in two log-likelihood scores is within the BIC, the two reaction coordinate models are indistinguishable, at least from the data collected. However, due to finite sampling, there is statistical uncertainty present in the estimate of likelihood score in eq 3[26]

$$\sigma^2(\ln L) = \sum_{\mathbf{x}_k} p_B(\mathbf{x}_k)[1 - p_B(\mathbf{x}_k)]\{\ln p_B(\mathbf{x}_k) - \ln[1 - p_B(\mathbf{x}_k)]\}^2 \quad (4)$$

where the sum is over all shooting points, each of which is a p_B-realization.

Reaction Coordinate Validation. After the likelihood maximization generates an approximation to the reaction coordinate, its correctness must be checked. This can be done by computing the estimate of the probability of reaching the product basin (p_B) from the predicted transition state region obtained in likelihood maximization, commonly referred to as a committor distribution analysis, or p_B histogram computation. In this procedure, independent configurations are generated that all satisfy the requirement as transition state configurations according to the reaction coordinate model to be checked. Then, a number of trajectories are initiated with the momenta drawn from a Maxwell−Boltzmann distribution from these configurations and the estimate of p_B values for these configurations can thus be obtained. Next, a histogram of the number of configurations versus p_B values can be constructed.

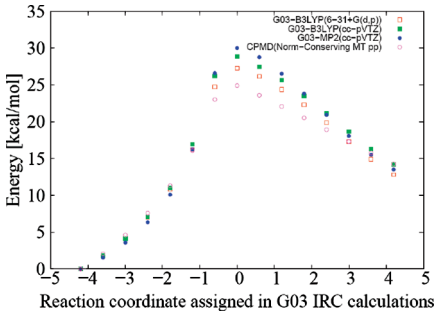


Figure 3. Comparison of energetics in quantum mechanical calculations in G03 with those obtained from CPMD calculations for an isolated system with one *N*-MAA and two water molecules. The points correspond to a series of configurations obtained from an IRC calculation starting from the TS structure determined in G03 with level/basis-set B3LYP/6-31+G(d,p). The energies are total potential energy without zero-point energy corrections, relative to the lowest values in the respective theory/basis-set or CPMD.

For complex reaction coordinates like the ones used in this study, generating independent configurations for the p_B histogram computation can be done efficiently by using the BOLAS algorithm.[44] In our work, shooting points were first examined and several of them were selected close to the predicted transition state region, as defined by $r(q) = 0$ in eq 1. Very short trajectories were fired randomly from each initial configuration, and the end points were evaluated to determine if they were within a narrow window on the transition state isosurface. If so, this configuration was accepted and became the next shooting point. This process was repeated until an adequate number of configurations were generated from which to shoot reactive trajectories to build a p_B histogram.

To construct the histogram, trajectories were shot from each configuration with a length corresponding to half the length of a reactive trajectory. The end points of the trajectories were evaluated, and a histogram was constructed of the probability of reaching basin B from the predicted transition state isosurface. The basin definitions for constructing the p_B histograms correspond to the same basin definitions used for the reactant and product basins in the aimless shooting simulations. An adequate approximation to the true reaction coordinate will yield a histogram that peaks sharply at $p_B = 0.5$.[25] Additionally, one can make a quantitative comparison of the histogram to the binomial distribution, which will have a mean value of $\mu = 0.50$ with a standard deviation, $\sigma = 0.050$.

The trajectories for the generation of new configurations are ~10 fs or 100 MD steps long, and the end point window width at $r = 0$ is constrained within a range of ±1% of the total configuration space sampled, as measured by Δr. For each histogram assembled in this study, 20 shooting points were collected. From each configuration collected, 20 trajectories were shot, corresponding to approximately 400 trajectories for each histogram. The trajectory length for calculating p_B values was ~60 fs or 600 MD steps, which was half the length of the

Table 2. Comparison of the Potential Energies Calculated Using G03 with the Level/Basis-Set B3LYP/6-31+G(d,p), B3LYP/cc-pVTZ, MP2/cc-pVTZ, and CPMD for Configurations Obtained from IRC Calculations with the Level/Basis-Set B3LYP/6-31+G(d,p)[a]

method or level/basis-set	E(au)	ΔE(TS) (kcal/mol)	ΔE(product) (kcal/mol)
B3LYP/6-31+G(d,p)	−401.4385	42.18	7.84
B3LYP/cc-pVTZ	−400.7138	41.73	9.90
MP2/cc-pVTZ	−401.5671	43.32	6.88
CPMD	−81.5010	40.37	7.36

[a] The system contains one *N*-MAA molecule with two water molecules. The absolute potential energies E (in atomic unit, au) of the reactant, the relative potential energies (in kcal/mol) of the transition state ΔE(TS), and the relative potential energies (in kcal/mol) of the product ΔE(product) to the reactant state are listed.

reactive trajectories in the aimless shooting simulations, again resulting in a low rate of inconclusive paths.

RESULTS AND DISCUSSION

Pseudopotentials and CPMD Validation. As mentioned in the previous section, several other studies used exactly the same BLYP functional and pseudopotentials with CPMD as in this work to study chemical reactions in molecular systems. However, a detailed analysis of the accuracy is required in order to provide a justification of the pseudopotentials and CPMD approach. Our comparison results are shown in Figure 3. Also, because the IRC calculation in G03 did not finish all the way from the TS to the reactant structure or from the TS to the product structure, a separate table is used to list the comparison of energies for the reactant, the TS, and the product configurations as in Table 2. In Table 2, optimized geometries for the reactant and product configurations obtained in the theory level/basis-set B3LYP/6-31+G(d,p) were used to obtain the comparison results but with no IRC reaction coordinates. (Therefore, they are not included in Figure 3.) Note that the zero-point-energy (ZPE) corrections are not included in these results, because they are not expected to make a large difference when the energies reported here are relative.

Table 2 shows that CPMD gave well matched relative potential energies with G03 calculations using various levels and basis-sets. Figure 3 shows that, as the basis-set becomes larger from 6 to 31+G(d,p) to cc-pVTZ, closer results of potential energies from B3LYP to MP2 are obtained, which is as expected, since MP2 is generally considered better in accounting for dispersion forces in a molecular system than DFT or hybrid methods in this case. The various quantum mechanical methods generate results comparable to those obtained from CPMD for configurations across the reactant, the TS, to the product states. In general, both Table 2 and Figure 3 show that larger differences between the various methods and basis-sets result in the TS region than in both stable state regions, which is also expected due to the larger degree of electron delocalization. From these results, it is inferred that CPMD gives energies with an accuracy within 5% of MP2 in the stable regions, and it can underestimate the energies closer to the TS region by 5 kcal/mol, or 15−20% when closer to the TS region. As pointed out in Cascella et al.'s study[13] and the work referenced therein, the BLYP functional has this error range in accounting for van der Waal's interactions. Nonetheless, considering the computational efficiency and reasonable accuracy, using CPMD with the BLYP functional and pseudopotentials are the best choice.

The BLYP functional together with norm-conserving pseudopotentials was widely used in the literature to study many different types of reactions, whether ionic or zwitterionic species are involved or not, and reported with success. Particularly among those references[13−19,45] in the relevance to the study of hydrolytic reaction of peptide bonds, Gorb et al.[19] contains a nice summary of the validity of using the BLYP functional together with norm-conserving pseudopotentials for different reaction routes involving zwitterionic species in addition to neutral ones. We believe these studies have proven DFT and pseudopotentials are very robust and transferable and therefore, valid and suitable for our study of *N*-MAA hydrolysis under neutral pH. It should not give severely biased sampling of one particular hydrolytic pathway relative to another or others. In addition, to provide a fair comparison to a previous theoretical study[17] and to evaluate the choice of the order parameter used for the projection of the free energy hypersurface therein, the same simulation techniques, i.e., CPMD with the BLYP functional and pseudopotentials were used in our study.

One must note that, in the CPMD calculations for these comparison results, the cell size is chosen to be the same as that used in our actual dynamic run. The differences in energetics between CPMD and MP2 are probably due to the limited cell size for the isolated system, because better results were obtained when a larger cell for this isolated system was used. However, in the actual dynamic CPMD run with filled solvent molecules, possible self-interaction energies due to the effect of the finite-size and the counter-interacting screening effect are expected to have a more subtle effect on the accuracy of the CPMD method. It also has to be noted that the ultimate validation of the applicability and accuracy of CPMD to study the hydrolytic reaction, or any reaction in general, should be done to compare forces or electronic density with high-level quantum mechanical calculations. Nonetheless, the user of the BLYP functional together with norm-conserving Troullier−Martins pseudopotentials should provide reasonably accurate energetic and dynamic results when applied to study the hydrolytic reaction of *N*-MAA.

Initial Trajectory. As mentioned previously, an initial reactive trajectory was obtained by first running long equilibration MD when fixing postulated bond distances of C−O, O−H, and N−H labeled in Figure 1 in order to remove potential artifacts when using constraints. Then, repeated forward and backward shootings were tried until the two half-trajectories combined to yield a reactive trajectory. Figure 2 shows how the bond distances changed in this particular trajectory, together with extended sampling in the stable states in order to see how these bond distances fluctuate. The transition dynamics has a time scale of

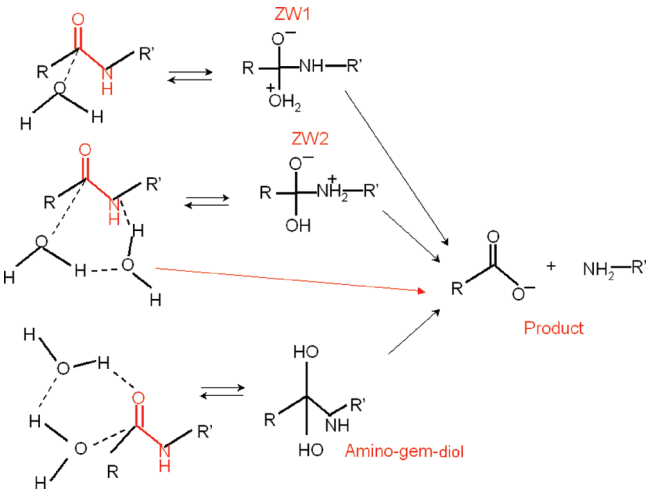


Figure 4. The various proposed reaction pathways with the water-assisted mechanism of the hydrolytic reaction of a peptide bond. Here, the proposed mechanisms involving the zwitterion TW1,[13] the zwitterion TW2,[17] and the amino-gem-diol[19] as intermediates are shown. A direct reaction pathway without any long-lived intermediate as shown in this study was also included.

~1000 MD steps. The values and fluctuations of these three bond distances provide a basis for choosing the MD steps in the aimless shooting procedure to reach the compromise of efficiency and minimal number of inconclusive trajectories.

Trajectory Characteristics. Snapshots from a typical reactive trajectory show what happens during the transition, as shown for the reaction core part in SI Figure 2 of the Supporting Information and for the whole system together with the changes in the fleeting hydrogen bond network in SI Figure 3 (see movie 1 in the Supporting Information for a better view on the reaction dynamics).

SI Figure 2 (Supporting Information) shows a series of configurational snapshots as a typical reaction process progresses. All of the reactive trajectories collected in the aimless shooting algorithm were visually inspected, and it was found that all of the reactive trajectories collected showed that the hydrolysis proceeds in a concerted fashion; i.e., the two proton-transfer processes and C−O bond formation occur simultaneously instead of forming any long-lived zwitterionic intermediate species (compared with Zahn's finding[17]). In all of the reactive trajectories collected, C−N bond breaking was observed to produce the final hydrolyzed product, a methylamine molecule and an acetic acid molecule, both in their nonionized forms due to the neutral pH condition. In SI Figure 3 (Supporting Information), significant changes in the hydrogen-bond pattern close to the reaction core can be seen, indicating the effects of solvent degrees of freedom.

Discussions of Possible Reaction Pathways. In the work by Gorb et al.,[19] a number of reaction pathways for hydrolytic reaction in neutral pH was summarized, as shown in Figure 4. In Figure 4, only the water-assisted mechanisms are shown, due to the expectation that, in neutral pH and a liquid environment, water assistance can significantly lower the reaction barrier and thus favorably initiate the reaction. The study of Cascella et al.[13] concluded that the hydrolytic reaction of formamide in neutral pH involves a zwitterionic intermediate TW1 after the oxygen in a water molecule attacks the carbonyl carbon; in contrast, Zahn's study[17] found that the reaction involves two or more water molecules attacking both carbon and nitrogen atoms in the peptide bond in a concerted fashion, forming a different zwitterionic intermediate TW2. However, in Gorb et al.'s study,[19] the reaction proceeds with an amino-gem-diol intermediate, with two or more water molecules attacking the carbonyl bond in a concerted fashion. Our work found no intermediate in the reaction dynamics, and showed that the concerted attacking of carbon and nitrogen atoms by water molecules directly results in the products.

As pointed out by Zahn, the steering MD used in Cascella et al.'s study probably has a too fast scanning velocity to capture the rearrangement of the hydrogen-bonded network occurring on a time scale of picoseconds. It should be emphasized that the studies of both Cascella et al. and Zahn's used some kind of constraining or restraining forces on the system in order to observe the reaction occurring; these forces may introduce some artifacts into the reaction dynamics. Another difference in these studies is the use of different reactants. The studies of both Cascella et al. and Gorb et al. used formamide, while ours and Zahn's studies used *N*-MAA. The methyl group, because it is bulkier than a hydrogen atom, is likely to disfavor the attacking of the carbonyl bond by water molecules.

Observing the different results, we note that Cascella et al., Zahn, and we used the same DFT functional and pseudopotentials in CPMD, while Gorb et al. used the QM/MM approach where their QM treatment used DFT with the BP functional and a double-ζ basis-set with polarization functions. Gorb et al. fail to mention how accurate the QM/MM methodology is. One other point to note is that Gorb et al. found a reaction free energy barrier (37.1 kcal/mol), which is very close to Zahn's simulation (35.2 ± 3 kcal/mol), suggesting the possibility of equally probable reaction pathways. However, neither found the intermediate involved in the other's study. Most importantly, Gorb et al. located stationary points to approximate the reaction's free energy barrier, while Zahn's study projected free energy onto a probably wrong RC; both of the reported numbers of the free energy barrier are an approximation to the "true" one, in addition to the fact that there are inaccuracies in the computational techniques.

In principle, TPS algorithms should be able to autodetect all relevant reaction intermediates and reaction pathways, given that the sampling of path space is sufficient. However, it is still possible that higher potential energy regions prohibit the sampling random walker in the path space from escaping local minima, in this case, particular reaction pathways. Our study determined the sufficiency of the path ensemble by looking at the convergence of the reaction coordinate produced from likelihood maximization, rather than determining whether the whole path space has been sufficiently trespassed. (A problem prohibitively expensive if not impossible.) Of course, the latter is in reality rather difficult to know. Other computational techniques such as replica exchange, simulated annealing, etc., need to be applied to further analyze this problem. It should be noted that both our results and those from Zahn's study show that the hydrolytic reaction occurs by the concerted water-assisted mechanism; however, the two differ in showing whether there is a long-lived intermediate involved or not. Our simulation results show that some of the structures (like ZW1 and ZW2) appeared shortly in some reactive trajectories collected by our aimless shooting algorithm, but these structures never lived long enough before the bond between the carbonyl carbon and the nitrogen finally broke to form the final products.

Table 3. Likelihood Maximization Results for N = 1650 Aimless Shooting Paths, with BIC = log(N/2) = 3.704[a]

number of OP variables in the RC model	OPs in best ranked RC models	log-likelihood Score (ln(L))
1	phi(O13−C26−H35−H56)	−913.613
	d(O13−C26)[b]	−1111.496
2[c]	phi(C25−O1−C26−H56), phi(H31−H56−H75−H43)	−846.562
3[+]	phi(C25−O1−C26−H56), phi(H31−H56−H75−H43), phi(H54−O19−H56−H55)	−819.826
4[+]	phi(C25−O1−C26−H56), phi(H52−O6−H74−H56), d(H54−H55), d(O13−H56)	−810.862
5[+]	d(O13−H56), d(H54−H55), phi(C25−O1−C26−H56), a(O13−O3−H56), phi(H52−O6−H74−H56)	−803.535
6[+]	phi(C25−O1−C26−H56), phi(H31−H56−H75−H43), phi(H54−O19−H56−H55), phi(C23−O5−H52−H54), phi(O11−H56−H71−H27), phi(H47−O6−H62−H56)	−800.130

[a] The order parameters (OPs) have the following meaning: d(n1, n2) is the distance between atom numbers n1 and n2; a(n1, n2, n3) is the angle comprised of atom numbers n1, n2, and n3; and phi(n1, n2, n3, n4) is the dihedral angle comprised of atom numbers n1, n2, n3, and n4. Refer to Figure 1 in the Supporting Information for atom labels. The column α_i's gives the vector $\alpha = (\alpha_0, \alpha_1, ..., \alpha_n)$ corresponding to reduced and normalized OP $q_i \in [0, 1]$. [b] This order parameter was used in previous potential-of-mean-force calculations.[17] [c] Results of likelihood maximization with more than one OP variable, with m = 100, n = 100. Convergence was achieved at d = 5. See the text for the modified algorithm of likelihood maximization.

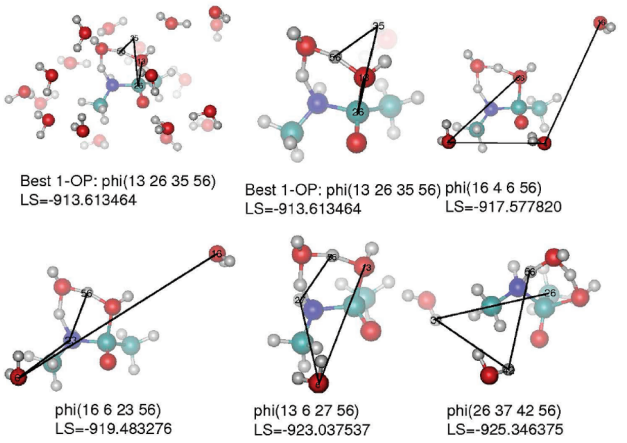


Figure 5. Illustration of the OPs in the best ranked one-OP-variable reaction coordinate models in the likelihood maximization procedure. The OP is given as a dihedral among quadruple atoms (denoted as phi(atom_index_1, atom_index_2, atom_index_3, atom_index_4)), as an angle among triple atoms (denoted as a(atom_index_1, atom_index_2, atom_index_3)), or as a bond distance between pair atoms (denoted as d(atom_index_1, atom_index_2)). The associated likelihood scores (LS) are also given. One observation is that almost all of these best-ranked reaction coordinate models involve the hydrogen atom indexed as 56:H.

Therefore, all of the above referenced studies, including ours, may have analyzed one of the several probable reaction pathways for the hydrolytic reaction of peptide bonds. However, if there were a significant lowering of free energy in the reaction pathway involving the gem-diol intermediate, the aimless shooting procedure would have resulted in many inconclusive trajectories. However, this is not the case in our simulation. Besides, structures similar to zwitterionic species and gem-diols were indeed observed in our simulation, but they did not persist long enough for the reaction process to be trapped. We think this is firm evidence that they should be irrelevant in the reaction pathway. However, we still we believe that the correct approach to resolve the differences in the reaction mechanism and their relative importance in the actual reaction system is to put all possible reaction mechanisms together and obtain their relevant weights in the actual reaction system. TPS algorithms combined with other higher level techniques are required for this purpose, of course coming with a much higher cost. This approach also requires further detailed analysis of the accuracy of the computational method. However, to the best of our knowledge, no one has applied global optimization techniques to systems of the magnitude in this study. If there were other pathways that are not separated by high barriers and are in very different regions of phase space, our method should have found them, but it did not.

Likelihood Maximization. The results of likelihood maximization appear in Table 3. A few of the best reaction coordinate models with one up to six OP variables are listed. (See Table SI in the Supporting Information for more information.) In addition, the best one-OP-variable reaction coordinate models are also pictorially shown in Figure 5. On the basis of likelihood scores, these one-OP-variable reaction coordinate models are much better in describing the hydrolytic reaction in the statistical sense of likelihood maximization than the order parameter of the distance between O(13)−C(26), which was used in the calculation of the potential-of-mean-force for the rate-limiting step of the hydrolysis of N-MAA at neutral pH.[17] As more OPs were included in the linear combination expression of the reaction coordinate model in eq 1, higher and higher log-likelihood scores were obtained, suggesting a complex reaction mechanism involving many physical degrees of freedom. In our approximation scheme for the likelihood maximization procedure, when including more OP variables in the reaction coordinate models, the log-likelihood score fell within the BIC criterion after five OP variables were included, indicating a convergent result.

Both best reaction coordinate models with three OP variables and five OP variables were checked against the aimless shooting data, as shown in Figure 6. All accepted shooting points for which the forward and backward shootings led to a conclusive trajectory were used to construct two histograms of the reaction coordinate determined from likelihood maximization. The two histograms were based only on the half-trajectories of the forward shooting from each accepted shooting point whether it ends in the reactant basin or in the product basin. Then $p_B(r)$ data values in Figure 6 were computed as the ratio

$$p_B(r)|_{\text{data},\, i^{\text{th}}\text{ bin}} = \frac{N_{fA,i}}{N_{fA,i} + N_{fB,i}} \qquad (5)$$

where $N_{fA,i}$ and $N_{fB,i}$ stand for the number of shooting points whose configurations give the reaction coordinate value r in the ith bin and

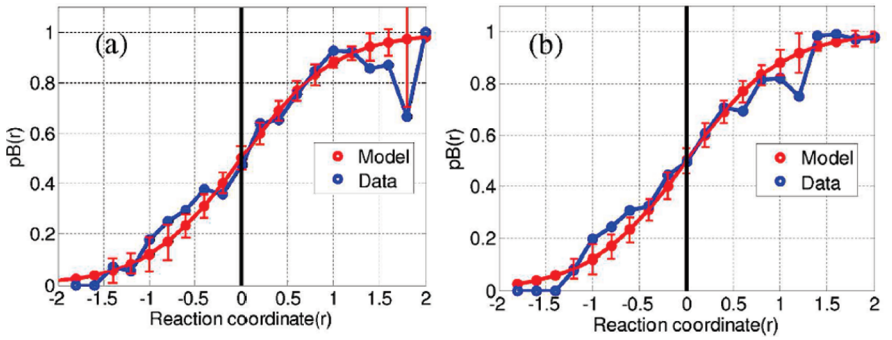


Figure 6. Comparison of the $p_B(r)$ model vs aimless shooting data. Here, a half-trajectory $p_B(r)$ model was used, i.e., $p_B(r) = [1 + \tanh(r)]/2$. Note that the error bars appear on the model, not on the data. The error bars show how far shooting point data should deviate from the probabilities $p_B(r)$ for a perfect reaction coordinate model. (a) 3-OP-variable reaction coordinate model. (b) 5-OP-variable reaction coordinate model.

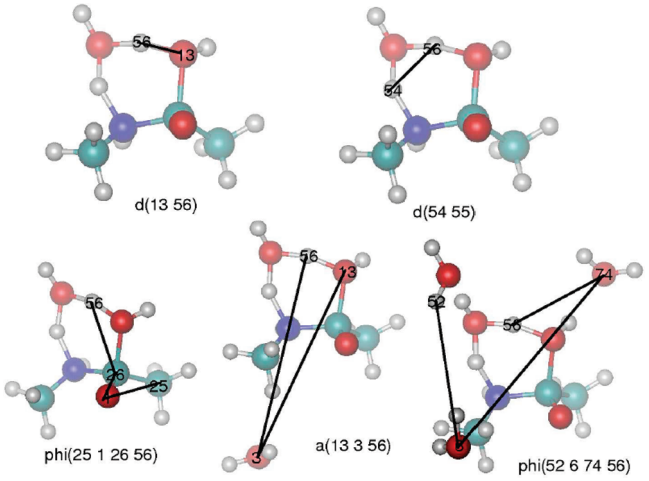


Figure 7. Illustration of constituent OPs in the best 5-OP-variable reaction coordinate model. Naming of these OPs is the same as in Figure 5.

from which the forward shooting half-trajectory that ends in A or B, respectively. Thus, the comparison of model vs data provides a measure of how well aimless shooting collects information about transition paths and shooting points, as well as how well the likelihood maximization is performed. Figure 6 shows that both reaction coordinate models satisfactorily fit the aimless shooting data.

In eq 4, the statistical uncertainty in the log-likelihood score was computed with the collection of accepted shooting points to be $\sigma^2(\ln L) = 14.86$ and $\sigma^2(\ln L) = 12.60$ for best the three-OP- and five-OP-variable reaction coordinate models, respectively. These values provide reasonable confidence in our log-likelihood score value.[28]

Figure 7 shows the OPs which comprise the best reaction coordinate model with five OP variables. These include the local bonding pattern changes, such as proton H56 being transferred between the two attacking water molecules, and the newly formed N23−H54 bond (refer to Figure SI 1 of the Supporting Information for labeling). They also reflect some influence of the solute *N*-MAA itself on the reaction dynamics, such as the dihedral angle between C25−O1−C26−H56. Solvent degrees of freedom in affecting reaction dynamics are also needed, as seen by the presence of an angle O13−O3−H56 and a dihedral angle H52−O6−H75−H56. The inclusion of both local OPs close to the reaction center and the OPs describing the solvent networks suggests the importance of the solvent in determining reaction dynamics. Consistent with the findings of other studies, such as proton transfer in liquid water,[32] protein folding,[46] and the solvation of small hydrophobic molecules,[47,48] aqueous phase reactions can critically depend on solvent degrees of freedom.

In this study, the five geometric quantities used in the best reaction coordinate model also involve many atoms, actually 11 in total (33 coordinate variables), and while the best reaction coordinate model does not express the important collective variables in terms of bonding number, density fluctuation, or tetrahedrality as used in other studies, it does provide insight into the multiple degrees of freedom and solvent effects. The insight gained here is based on finding the right reaction coordinate systematically by introducing trial coordinates systematically and testing them systematically. Additional variables, such as the number of hydrogen bonds or the radial, angular, and relative orientational distribution of solvent molecules with respect to the reaction center, etc., could be included in the likelihood maximization procedure for analysis. However, we did not pursue this direction because we took a different approach that could be applied systematically to and with more likelihood of yielding convergent results. Our approach is to assume many geometric quantities, such as distances, angles, and dihedral angles, in comprising the reaction coordinate. After screening millions of different reaction coordinate models, we verified whether the best reaction coordinate can serve as a good descriptor for describing the reaction process by calculating the pB histograms.

Our results emphasize that the convergent RC determined here contains information about the reaction dynamics and solvent influence compared with the IRC reaction coordinate determined in G03 calculations for the reaction in the gas phase. Even though the coordinates of other solvent molecules do not explicitly appear in the final expression of the convergent RC, their influence is implicitly included. We also note that the RC determined as described in this work is specific to the simulation system and computational techniques employed and a different solvated simulation system with a different number of water molecules, for example, may produce a different RC due to the different levels of approximation to the "true" RC. However, the essential feature should be contained in different RC expressions, to reflect the common feature of the underlying reaction mechanism.

Reaction Coordinate Validation. Four p_B histograms were computed using the method described above. These include

dx.doi.org/10.1021/jp1076802 |*J. Phys. Chem. B* 2011, 115, 5958–5970

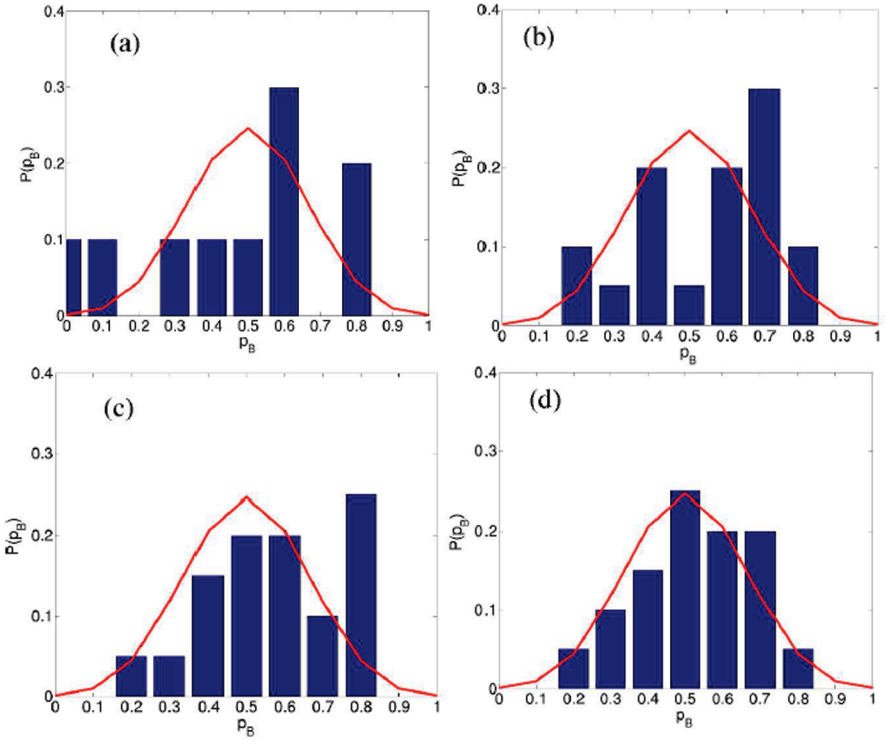


Figure 8. Committor probability histogram using C(26)−O(13) as a reaction coordinate model (a), best one-OP-variable reaction coordinate model (b), best two-OP-variable reaction coordinate model (c), and best five-OP-variable reaction coordinate model (d), compared with binomial distribution (red line). Quantification of means and standard deviations for these histograms following the procedure in Peters[28] is shown in Table 3.

Table 4. Maximal Likelihood Estimates for Means and Standard Deviations in the p_B Histograms Shown in Figure 7 [a]

reaction coordinate model/distribution	μ_h	σ_h
C(26)−O(13)	0.470	0.066
best one-OP variable	0.550	0.034
best two-OP variable	0.575	0.031
best five-OP variable	0.525	0.024
binomial distribution	0.500	0.025
ideal $P(p_B)$	0.500	0.000

[a] The procedure was used as in Peters.[28]

using the C−O distance reaction coordinate model, the best one-OP reaction coordinate model, the best two-OP reaction coordinate model, and the best five-OP reaction coordinate model. The results are shown in Figure 8 and Table 4. The poor description of the reaction process by the C−O distance reaction coordinate model can be seen in this p_B histogram, since its distribution is very skewed. Both the best one-OP reaction coordinate model and the best two-OP model show a histogram inferior to the best five-OP-variable reaction coordinate model.

CONCLUSIONS

In this study, the mechanism of the hydrolytic reaction of peptide bonds at neutral pH was examined using a model compound, *N*-MAA. Due to fluctuations at finite temperatures, the path-sampling method was used to generate an ensemble of trajectories according to their statistical weight in trajectory space. The *ab initio* molecular dynamics technique was applied to advance the time evolution of the reaction and collect trajectories. Likelihood maximization and its modification were used in extracting physically important degrees of freedom in the system, and approximations of the reaction coordinate were compared. It was found that this hydrolytic reaction is very complex in nature, and involves many degrees of freedom. The specific conclusions obtained in our study are the following:

- Hydrolysis of *N*-MAA at neutral pH occurs in a concerted fashion; no stable or long-lived intermediate was found in our path-sampling simulations.
- The likelihood maximization procedure was extended to screen reaction coordinate models with more than one OP variable; a reaction coordinate with five constituent geometric variables was found to be the best in describing the path ensemble generated, in that including greater model complexity did not increase the log likelihood score above the BIC.
- In the best reaction coordinate model, both geometric quantities that reflect bond-making and -breaking dynamics and those that reflect the solvent network changes are included, suggesting a complicated reaction involving many degrees of freedom. The inclusion of both local OPs close to the reaction center and the OPs describing the solvent networks is manifest of the importance of solvent in determining reaction dynamics. Our study shows consistency with the conclusions from other studies, such as proton transfer in liquid water,[32] protein folding,[46] and the solvation of small hydrophobic molecules,[47,48] that aqueous phase reactions can critically depend on solvent degrees of freedom. Here, we further acknowledged this by explicitly constructing the best approximate RC. In this

logic, the participation of the solvent degrees of freedom necessitates the use of explicit water molecules in studying chemical reactions such as this one.
- Several p_B histograms were computed to verify the results of likelihood maximization, and the relative ranking of the top reaction coordinate models is in accord with their respective likelihood scores.

The technique of likelihood maximization is a very powerful statistical tool in determining a reaction mechanism, especially when it is complex. However, new problems arise when the set of candidate order parameters is large, such as the combinatorial problem in exhaustive screening for the best reaction coordinate model. This study provides an approach to conduct a curtailed optimization procedure to determine the reaction coordinate. In addition, our approach presented here critically depends on the proper choice of the set of candidate OPs, as does the approach of genetic algorithm and neutral network.[33] This requirement has to be supplemented with one's chemical intuition into the system being studied. One can take the advantage that the computational cost of a single such OP variable is fast in order to perform the screening over a large number of candidate variables, thus partially overcoming the problem of dependence on the set of candidate OPs. Even though it may be difficult to relate a concise physical picture with the likelihood maximization using many OPs, with respect to the p_B histogram and the calculation of quantities of interest such as free energy profile and reaction rates, this improvement has advantages.

Overall, our study shows that combining path sampling and reaction coordinate determination techniques gave us direct information about the nature of the hydrolysis reaction under acidic pH. In particular, the reaction itself involves many degrees of freedom, including solvent degrees of freedom. There does not seem to be a simple geometric coordinate that governs the reaction. The reaction coordinate found here should be physically relevant in that rates of reaction and free energy barriers to reaction that could be calculated using this reaction coordinate should be physically accurate. This reaction coordinate could also be used to understand the reason why rates of hydrolysis of the same pairs of amino acids depend on the structure of the protein in which they reside. Both of these would be separate studies and outside the scope of the present work.

■ ASSOCIATED CONTENT

Supporting Information. A movie showing the MD trajectory (from a CPMD simulation) of a typical reactive transition for the hydrolytic reaction of the peptide bond in N-MAA. It also contains the figures for the atom labeling scheme referenced in the paper, the snapshots describing the reaction process, changes of the hydrogen network in the system, and the autocorrelation function showing the dependence of successive shooting moves in the aimless shooting algorithm. A table showing the best RC's with detailed comprising OP's is also included. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: trout@mit.edu.

■ ACKNOWLEDGMENT

We thank Amgen Inc. for their generous financial support of this study. We thank Dr. Gregg Beckham, Dr. Baron Peters, Dr. Erik Santiso, and Nicholas Musolino for insightful discussions. We are grateful to Tom Dillon, Dr. Pavel Bondarenko, Dr. David Brems, and Jeff Abel at Amgen Inc. for their support of this work and our previous work. Special thanks to Elizabeth Fox at MIT's writing center for editing of this manuscript.

■ REFERENCES

(1) Kahne, D.; Still, W. C. *J. Am. Chem. Soc.* **1988**, *110*, 7529–7534.
(2) Cohen, S. L.; Price, C.; Vlasak, J. *J. Am. Chem. Soc.* **2007**, *129*, 6976–+.
(3) Cordoba, A. J.; Shyong, B. J.; Breen, D.; Harris, R. J. *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2005**, *818*, 115–121.
(4) Dillon, T. M.; Bondarenko, P. V.; Rehder, D. S.; Pipes, G. D.; Kleemann, G. R.; Ricci, M. S. *J. Chromatogr., A* **2006**, *1120*, 112–120.
(5) Brown, R. S.; Bennet, A. J.; Slebockatilk, H. *Acc. Chem. Res.* **1992**, *25*, 481–488.
(6) Bryant, R. A. R.; Hansen, D. E. *J. Am. Chem. Soc.* **1996**, *118*, 5498–5499.
(7) Radzicka, A.; Wolfenden, R. *J. Am. Chem. Soc.* **1996**, *118*, 6105–6109.
(8) Smith, R. M.; Hansen, D. E. *J. Am. Chem. Soc.* **1998**, *120*, 8910–8913.
(9) Krug, J. P.; Popelier, P. L. A.; Bader, R. F. W. *J. Phys. Chem.* **1992**, *96*, 7604–7616.
(10) Antonczak, S.; Ruizlopez, M. F.; Rivail, J. L. *J. Am. Chem. Soc.* **1994**, *116*, 3912–3921.
(11) Bakowies, D.; Kollman, P. A. *J. Am. Chem. Soc.* **1999**, *121*, 5712–5726.
(12) Stanton, R. V.; Perakyla, M.; Bakowies, D.; Kollman, P. A. *J. Am. Chem. Soc.* **1998**, *120*, 3448–3457.
(13) Cascella, M.; Raugei, S.; Carloni, P. *J. Phys. Chem. B* **2004**, *108*, 369–375.
(14) Zahn, D. *J. Phys. Chem. B* **2003**, *107*, 12303–12306.
(15) Zahn, D. *Chem. Phys. Lett.* **2004**, *383*, 134–137.
(16) Zahn, D. *Chem. Phys.* **2004**, *300*, 79–83.
(17) Zahn, D. *Eur. J. Org. Chem.* **2004**, 4020–4023.
(18) Zahn, D.; Schmidt, K. F.; Kast, S. M.; Brickmann, J. *J. Phys. Chem. A* **2002**, *106*, 7807–7812.
(19) Gorb, L.; Asensio, A.; Tunon, I.; Ruiz-Lopez, M. F. *Chem.—Eur. J.* **2005**, *11*, 6743–6753.
(20) Buren, N. V.; Rehder, D.; Matsumura, H. G. M.; Jacob, J. *J. Pharm. Sci.* **2009**, *98*, 3013–3030.
(21) Perico, N.; Purtell, J.; Dillon, T.; Ricci, M. S. *J. Pharm. Sci.* **2009**, *98*, 3031–3042.
(22) Pan, B.; Ricci, M. S.; Trout, B. L. *J. Phys. Chem. B* **2010**, *114*, 4389–4399.
(23) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.
(24) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964–1977.
(25) Dellago, C.; Bolhuis, P. G.; Geissler, P. L. *Adv. Chem. Phys.* **2002**, *123*, 1–78.
(26) Peters, B.; Beckham, G. T.; Trout, B. L. *J. Chem. Phys.* **2007**, *127*, 034109.
(27) Peters, B.; Trout, B. L. *J. Chem. Phys.* **2006**, *125*, 054108.
(28) Peters, B. *J. Chem. Phys.* **2006**, *125*, 241101.
(29) Weinan, E.; Ren, W. Q.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2005**, *413*, 242–247.
(30) Metzner, P.; Schutte, C.; Vanden-Eijnden, E. *J. Chem. Phys.* **2006**, *125*, 084110.
(31) E, W.; Vanden-Eijnden, E. *J. Stat. Phys.* **2006**, *123*, 503–523.
(32) Geissler, P. L.; Dellago, C.; Chandler, D.; Hutter, J.; Parrinello, M. *Science* **2001**, *291*, 2121–2124.

(33) Ma, A.; Dinner, A. R. *J. Phys. Chem. B* **2005**, *109*, 6769–6779.
(34) Blumberger, J.; Ensing, B.; Klein, M. L. *Angew. Chem.* **2006**, *118*, 2959–2963.
(35) Gunaydin, H.; Houk, K. N. *J. Am. Chem. Soc.* **2008**, *130*, 15232–15233.
(36) Frisch, M. J.; et al. *Gaussian 03*, revision B.05; Gaussian, Inc.: Wallingford, CT, 2004.
(37) MacKerell, A. D., Jr.; *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
(38) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
(39) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
(40) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.
(41) (Please see http://www.scc.acad.bg/downloads/programs/CPMD/CPMD_V3.13.2_Manuel.pdf page 8): CPMD, http://www.cpmd.org/, Copyright IBM Corp 1990−2008, Copyright MPI fur Festkorperforschung Stuttgart 1997−2001.
(42) Beckham, G. T.; Peters, B.; Starbuck, C.; Variankaval, N.; Trout, B. L. *J. Am. Chem. Soc.* **2007**, *129*, 4714–4723.
(43) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.
(44) Radhakrishnan, R.; Schlick, T. *J. Chem. Phys.* **2004**, *121*, 2436–2444.
(45) Pelmenschikov, V.; Blomberg, M. R.; Siegbahn, P. E. *J. Biol. Inorg. Chem.* **2002**, *7*, 284–298.
(46) Bolhuis, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12129–12134.
(47) Miller, T. F.; Vanden-Eijnden, E.; Chandle, D. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 14559–14564.
(48) van Erp, S. T.; Meijier, J. E. *Angew. Chem., Int. Ed.* **2004**, *43*, 1660–1662.