Regular Article

Linear Discriminant Analysis for the *in Silico* Discovery of Mechanism-Based Reversible Covalent Inhibitors of a Serine Protease: Application of Hydration Thermodynamics Analysis and Semi-empirical Molecular Orbital Calculation

Yosuke Masuda,* Tomoki Yoshida, Noriyuki Yamaotsu, and Shuichi Hirono*

School of Pharmaceutical Sciences, Kitasato University; 5–9–1 Shirokane, Minato-ku, Tokyo 108–8641, Japan.

Received October 22, 2017; accepted January 22, 2018

We recently reported that the Gibbs free energy of hydrolytic water molecules (ΔG_{wat}) in acyl-trypsin intermediates calculated by hydration thermodynamics analysis could be a useful metric for estimating the catalytic rate constants (k_{cat}) of mechanism-based reversible covalent inhibitors. For thorough evaluation, the proposed method was tested with an increased number of covalent ligands that have no corresponding crystal structures. After modeling acyl-trypsin intermediate structures using flexible molecular superposition, ΔG_{wat} values were calculated according to the proposed method. The orbital energies of antibonding π^* molecular orbitals (MOs) of carbonyl C=O in covalently modified catalytic serine (E_{orb}) were also calculated by semi-empirical MO calculations. Then, linear discriminant analysis (LDA) was performed to build a model that can discriminate covalent inhibitor candidates from substrate-like ligands using ΔG_{wat} and E_{orb} . The model was built using a training set (10 compounds) and then validated by a test set (4 compounds). As a result, the training set and test set ligands were perfectly discriminated by the model. Hydrolysis was slower when (1) the hydrolytic water molecule has lower ΔG_{wat} ; (2) the covalent ligand presents higher E_{orb} (higher reaction barrier). Results also showed that the entropic term of hydrolytic water molecule ($-T\Delta S_{\text{wat}}$) could be used for estimating k_{cat} and for covalent inhibitor optimization; when the rotational freedom of the hydrolytic water molecule is limited, the chance for favorable interaction with the electrophilic acyl group would also be limited. The method proposed in this study would be useful for screening and optimizing the mechanism-based reversible covalent inhibitors.

Key words covalent inhibitor; Gibbs free energy; hydrolytic water molecule; linear discriminant analysis; hydration thermodynamics analysis; semi-empirical molecular orbital calculation

Covalent inhibitors are a distinctive class of drugs, owing to their covalent bonding capability with their target proteins when they present inhibitory efficacy. They exhibit several advantages over non-covalent inhibitors, such as strong inhibition in small doses, prolonged duration and less-frequent dosing, and potency for anti-drug resistance.¹⁾ Compounds with covalent inhibition potencies have been carefully excluded as drug candidates for a long time, mainly because of their safety concerns—electrophiles in covalent inhibitors often modify off-target proteins and cause serious idiosyncratic toxicity.²⁾

Mechanism-based reversible covalent inhibitors are a class of covalent inhibitors that reversibly inhibit target enzymes *via* catalytic mechanisms. While irreversible covalent inhibitors form permanent bonds, reversible covalent inhibitors form reversible bonds with their targets.²⁾ Reversible covalent inhibitors are expected to present lower risks of toxicity, owing to their small daily doses and their nature of reversible reactivity—off-target drug–protein adducts would not exceed levels required to induce a host immune response.²⁾ Several recent successful examples of mechanism-based reversible covalent inhibitors are known: boceprevir³⁾ (HCV protease inhibitor for hepatitis type C), vildagliptin⁴⁾ (Dipeptidyl Protease-4 inhibitor for diabetes), finasteride⁵⁾ (5- α -reductase inhibitor for prostatic hyperplasia), and dimethyl fumarate⁶⁾ (Keap1-nuclear factor-E2-related factor 2 (Nrf2) inhibitor for multiple sclerosis), *etc.*

Covalent inhibitors have often been found accidentally—

e.g., aspirin, penicillin, and omeprazole were not intended to be designed as covalent inhibitors.^{1,2)} Among the numerous examples of *in silico* discoveries of non-covalent inhibitors, the *in silico* design of covalent inhibitors has also been studied. In computer-aided drug discovery of covalent inhibitors, the affinity of non-covalent binding, which occurs first, as well as the reactivity of covalent bond formation, which occurs second, must be carefully optimized. Thus far, covalent docking simulation protocols have been developed to predict non-covalent binding affinities and binding poses.⁷⁾ However, few researchers have addressed the development of effective and efficient methods to predict the potency for covalent inhibition. The potency remarkably depends on covalent bond formation/cleavage kinetics^{1,2)}; therefore, the rates of bond forming/cleaving reactions should be estimated.

For estimation of the reaction rates, quantum mechanics calculations have revealed the energetics and kinetics of enzymes.^{8–10)} We recently reported that hydration thermodynamics analysis could be a useful method for estimating the catalytic rate constants of the serine protease trypsin.¹¹⁾ Hydration thermodynamics analysis can predict the locations for hydration on protein surfaces and the relative energies at the locations with simple procedures and reasonable computational time.¹²⁾ The WaterMap™ software is based on restricted molecular dynamics (MD) simulation and inhomogeneous fluid theory.¹³⁾ In WaterMap™, the enthalpy and entropy of a water molecule relative to bulk solvent (ΔH and ΔS , respec-

* To whom correspondence should be addressed. e-mail: yosuke.masuda@gmail.com; hironos@pharm.kitasato-u.ac.jp

Fig. 1. Flowchart of Procedures in This Study

The procedures adopted in the previous study¹¹⁾ are shown outside the dotted lines. The procedures adopted in this study are shown inside the dotted lines. Gray rectangular boxes: protein structures; white rounded rectangular boxes: calculation methods; squared bracket: database, software name, or calculation methods.

tively) are calculated from force field and degree of ordering of water molecules, respectively, and then, the Gibbs free energy of a water molecule relative to bulk solvent (ΔG) is calculated.¹⁴⁾ WaterMapTM has been applied to various targets, and the displacement of water molecules with unfavorable ΔG has been demonstrated to improve the affinity of non-covalent binding.^{15,16)} It has also been proposed that WaterMapTM would be useful in estimating non-covalent binding kinetics.¹⁷⁾

In our recent study, we focused on serine protease as a target enzyme. Serine proteases (EC 3.4.21) have been used as drug targets and are involved in various important biological processes.^{18,19)} Trypsin (EC 3.4.21.4) is one of the representatives of serine proteases, and has been an important model for understanding catalytic mechanisms.²⁰⁾ There are two steps in trypsin hydrolysis: acylation and deacylation. In the deacylation step, catalytic histidine first deprotonates a hydrolytic water molecule. The deprotonated water molecule then nucleophilically attacks the carbonyl carbon of the acyl group in an acyl-trypsin intermediate (C_{acyl}). After formation of the tetrahedral intermediate, cleavage of the C–O bond in the acyl group occurs, and the hydrolysis procedure is completed.²¹⁾ Mechanism-based reversible covalent inhibitors of trypsin often contain ester as a hydrolysis-susceptible group, and in most ester hydrolysis reactions by trypsin, the deacylation reaction is the rate-determining step.²¹⁾

As mentioned earlier, we have previously reported that ΔG of hydrolytic water molecule (ΔG_{wat}) in an acyl-trypsin intermediate calculated by WaterMapTM could be a useful metric to estimate the catalytic rate constants of the trypsin hydrolysis.¹¹⁾ However, covalent ligands tested in the previous study were slightly limited because there were only five crys-

tal structures of acyl-trypsin intermediates in the Protein Data Bank (PDB). In this study, we expanded the reported method from ligands with corresponding crystal structures towards ligands with no corresponding crystal structures. Acyl-trypsin intermediate structures of those ligands were modeled *via* flexible molecular superposition.^{22,23)} Furthermore, in order to complement ΔG_{wat} , the orbital energy of antibonding π^* molecular orbital (MO) of carbonyl C=O in covalently modified catalytic serine (E_{orb}) was calculated by semi-empirical MO calculation. Semi-empirical MO calculation has a long history and various examples applied for the analysis of enzyme catalysis.^{24–26)} Here, we demonstrate that the ΔG_{wat} descriptor is a good metric to discriminate covalent inhibitor candidates from substrate-like compounds when it is used with the E_{orb} descriptor in linear discriminant analysis (LDA). The LDA with ΔG_{wat} and E_{orb} is proposed to be an effective method for the screening of mechanism-based reversible covalent inhibitors. We also propose the possibility that k_{cat} of a covalent inhibitor candidate could be estimated by the entropic term of hydrolytic water molecule ($-T\Delta S_{wat}$), which may be related to the degree of rotational freedom of a hydrolytic water molecule. The entropic term could be a useful index for optimization of a covalent inhibitor candidate.

Experimental

Procedures The procedures used in this study are shown in Fig. 1. The methods to obtain the ΔG_{wat} descriptor—energy minimization, molecular dynamics simulation, and hydration thermodynamics analysis and related procedures—are the same as those used in our previous study (outside the dotted lines). The previous method was modified in the following

three ways: (1) acyl-trypsin intermediate structures of ligands with no corresponding crystal structures were modeled by flexible molecular superposition; (2) semi-empirical MO calculations were performed to estimate the orbital energy of antibonding π^* MO of carbonyl C=O in covalently modified catalytic Ser195; (3) linear discriminant analysis (LDA) with two descriptors (ΔG_{wat} and E_{orb}) was performed for the discrimination covalent inhibitor candidates from substrate-like compounds.

Preparation of Ligand Datasets The covalent ligands^{27–34)} were separated into two groups—inhibitor group and substrate group, according to their $\text{Log}(k_{\text{cat}})$ at a cut-off of -1.0 . The cut-off was set because it allowed clear differentiation between the inhibitor group—including clinically adopted drugs, nafamostat ($\text{Log}(k_{\text{cat}})=-4.5$) and gabexate ($\text{Log}(k_{\text{cat}})=-3.1$)—and the substrate group mainly consisting of substrate-analog compounds [e.g., K3 ($\text{Log}(k_{\text{cat}})=-0.11$)] and other substrate-analog peptides [e.g., succinyl-alanyl-alanyl-prolyl-lysyl (suc-AAPK) ($\text{Log}(k_{\text{cat}})=1.6$)]. The ligands were then separated into the training set and test set by clustering analysis based on the fingerprints of two-dimensional ligand structures by using Canvas^{35–37)} (Schrödinger suite 2015-3). In the clustering analysis, the linear daylight invariant atom types³⁸⁾ (precision: 32-bit) were used for the atom-typing scheme. Bonds were distinguished by bond order. Compound similarity indexed by the Tanimoto coefficient was used with the average linking method. The ligands for the test set were then selected from the centroid of each cluster. The test set was used for the validation of the LDA model.

Preparation of Acyl-Trypsin Intermediate Structures Bovine pancreatic trypsin was adopted as the target enzyme in this study. In the covalent ligand used for the study, four ligands—nafamostat,^{31,34)} gabexate,²⁹⁾ succinyl-alanyl-alanyl-prolyl-arginyl-*para*-nitroanilide (sucAAPR-*pNA*),³¹⁾ and succinyl-alanyl-alanyl-prolyl-lysyl-*para*-nitroanilide (suc-AAPK-*pNA*)³¹⁾—have corresponding crystal structures of acyl-trypsin intermediate in PDB. In those ligands, nafamostat has two corresponding crystal structures (1GBT and 2AH4). Acyl-trypsin intermediates of the other ligands with no corresponding crystal structures^{27,28,30)} were modeled by flexible molecular superposition using Phase^{22,23,39)} (Schrödinger suite 2015-3). Prolyl-arginyl-catalytic Ser195 moiety in the crystal structure of suc-AAPR trypsin (2AGE)³¹⁾ was used as the reference structure of covalent ligands in the acyl-trypsin intermediate. First, five kinds of property spheres—positive charge, hydrogen bond donor, hydrogen bond acceptor, hydrophobic, and aromatic—were extracted from the reference ligand structure. At the center of the five-membered ring in the proline, an aromatic property sphere was located manually. Then, property spheres extracted from the query ligands were flexibly superposed to those of the reference structure. Among the superposed poses of a query ligand, one pose was selected as follows: if the query ligand has the guanidine group, the pose that presents smallest root-mean squared distance (RMSD) to the guanidine moiety ($-\text{N}-\text{C}(=\text{N}^+)-\text{N}$) in the reference structure (suc-AAPR trypsin) was selected; if the query ligand has the ethyl ammonium group, the pose that has smallest RMSD to the ethyl ammonium moiety ($-\text{C}-\text{C}-\text{N}^+$) of prolyl-lysyl-Ser195 in suc-AAPK trypsin (2AGG)³¹⁾ superposed to 2AGE structure was selected. Selected ligand structures were then manually connected to catalytic Ser195 of 2AGE (covalent

ligand moiety was removed beforehand), and finally, the modeled structures of acyl-trypsin intermediates were obtained.

The crystal structures and modeled structures were prepared by using Protein Preparation Wizard^{40,41)} (Schrödinger suite 2015-3) in Maestro⁴²⁾ 10.3 according to the following procedure. Non-protein residues, except calcium ions and the covalent ligand moieties, were correctly removed. All crystallographic water molecules were also removed, except two coordinated water molecules of the divalent calcium ion. Disulfide bonds were added to their appropriate positions. Missing side chains and residues of protein were built using Prime^{43–45)} 4.1, if necessary. The protonation state of the ionizable amino acids at pH 8.0, the optimum pH for enzymatic activity of trypsin,⁴⁶⁾ were determined with PROPKA.⁴⁷⁾ His57 side chain was protonated only in the N_δ atom but not in the N_ϵ atom, because the His57 N_ϵ atom works as a general base and deprotonates the hydrolytic water molecule.²¹⁾ The structures were then relaxed in an implicit solvent by using the restrained energy minimization protocol of the Impref module of Impact^{48,49)} 6.8.

MD Simulation Bias for crystal packing may be remained in the prepared structures. Therefore, MD simulations using Desmond⁵⁰⁾ 4.3 were performed to generate an ensemble of energetically accessible structures in water solution and to eliminate the bias. Each system for the MD simulation was built by using the System Builder module of Desmond. The prepared structures were buffered with 10-Å distance from the edge of the water box. TIP3P water models⁵¹⁾—generally used to obtain protein structures in the solution with high computation efficiencies—were used to solvate the system. The system was neutralized by adding counter chloride ions. Parameters from the OPLS3 force field⁵²⁾ were assigned for protein, modified catalytic Ser195, and other ions. In the first step of the MD simulation, each system was relaxed by a series of energy minimization and short MD simulations using the standard equilibration protocol in Desmond.⁵⁰⁾ In the next step, the production MD simulation was performed (simulation time: 3.5 to 9.5 ns) at a constant pressure (1 atm) and temperature (300 K). We then determined and extracted an equilibrated 1-ns period which presents almost constant RMSD values throughout the time period from each MD trajectory—4.0–5.0 ns (2AH4), 3.0–4.0 ns (1GBT), 3.5–4.5 ns (Lactone3), 8.5–9.5 ns (Isocoumarin4), 6.5–7.5 ns (Isocoumarin3), 6.0–7.0 ns (Isocoumarin1), 3.5–4.5 ns (3VPK), 3.0–4.0 ns (K3), 4.0–5.0 ns (R3), 3.0–4.0 ns (K2), 3.0–4.0 ns (2AGG), 4.0–5.0 ns (R1), 4.0–5.0 ns (2AGE), 4.0–5.0 ns (R2), and 2.5–3.5 ns (K1) [where the modeled structures underlined]. From the extracted 1-ns trajectory, structures were extracted every 24 ps. From those structures, three representative structures were selected by the clustering algorithm based on RMSD of heavy atoms within 4 Å from the modified catalytic Ser195 by using the Conformer Cluster script in Maestro⁴²⁾ 10.3.

Hydration Thermodynamics Analysis To obtain the ΔG_{wat} descriptor, hydration thermodynamics analysis was then performed. First, a protocol of WaterMap^{TM53)} (Schrödinger suite 2015-3) on an acyl-trypsin intermediate structure (gabexate-trypsin intermediate, 3VPK) was validated. The crystal structure was prepared by Protein Preparation Wizard^{40,41)} (Schrödinger suite 2015-3) with the OPLS3 force field.⁵²⁾ All crystallographic water molecules were deleted prior to the

analysis. Covalently modified catalytic Ser195 was defined as the center of the region of interest, and water molecules $\leq 10 \text{ \AA}$ of the region were analyzed. The structure was truncated according to the standard protocol of WaterMapTM in order to reduce the computation time. After the 2-ns simulation of WaterMapTM with TIP4P water models⁵¹⁾—generally-used water model for WaterMapTM simulation because behaviors of water molecules can be estimated more accurately than TIP3P water models—the coordinates of the predicted water sites $\leq 5 \text{ \AA}$ from the C_{acyl} atom were compared to the coordinates of crystallographic water molecules in the same region.

As a result, the simulation reproduced all locations of the three crystallographic water molecules with one exceptional water site—the site had no corresponding crystallographic water molecule. The coordinates of the oxygen atoms of three crystallographic water molecules and the coordinates of the corresponding predicted water sites presented significant identity (RMSD: $0.81 \text{ \AA}$). Validation of WaterMapTM was also performed in the other four crystal structures. As a result, the locations of the predicted water sites reproduce the locations of corresponding crystallographic water molecules well—RMSDs are $0.57 \text{ \AA}$ (2AH4), $0.79 \text{ \AA}$ (1GBT), $1.35 \text{ \AA}$ (2AGG) and $1.28 \text{ \AA}$ (2AGE), respectively. Therefore, the protocol was used for the following analysis.

The representative solution structures obtained by MD simulations were then analyzed by WaterMapTM. The TIP3P water models⁵¹⁾ and ions for neutralization were deleted prior to the analysis. Among the several water sites predicted around the catalytic site, the hydrolytic water site had the smallest sum of distances d —given by $d = d_1 + d_2$ —where d_1 is the distance between the His57 N_ϵ atom and a water site around the catalytic site, and d_2 is the distance between the C_{acyl} atom and the water site—because the hydrolytic water site should be near both atoms. The calculated ΔG , ΔH , $-T\Delta S$, occupancy, and average number of hydrogen bonds between the acyl-trypsin intermediate and hydrolytic water molecule throughout the WaterMapTM trajectory (#HB) were averaged for each covalent ligand ($n=3$).

Semi-empirical MO Calculation Semi-empirical MO calculations were performed to obtain a metric of the electrophilicity of acyl-trypsin intermediates. First, covalently modified catalytic Ser195 moieties (C_α , C_β , and O_γ atoms of Ser195, acyl group, and covalent ligand moiety with hydrogen atoms) were extracted from the three representative structures of acyl-trypsin intermediates. In suc-AAPK trypsin (2AGG) and suc-AAPR trypsin (2AGE), only prolyl-lysyl-Ser195 and prolyl-arginyl-Ser195 moieties were extracted for the calculation, respectively. After adding hydrogen atoms to the terminal C_α atom of Ser195, restricted energy minimizations were performed by MacroModel⁵⁴⁾ (Schrödinger suite 2015-3) for the added hydrogen atoms. Semi-empirical MO calculations in an aqueous environment were then performed by MOPAC2016 package using the PM7 Hamiltonian with the COSMO solvation method.^{55–57)} The dielectric constant of water solvent was defined as 78.3. The electric charges of the ligand were set to +1, because guanidine or ammonium moieties are positively charged. Single self-consistent field (SCF) calculation was performed on each structure. Then, the orbital energy of antibonding π^* MO of carbonyl $C=O$ in covalently modified catalytic serine was investigated using the Jmol graphical software.⁵⁸⁾ Those orbital energy values were finally averaged for

Fig. 2. Dendrogram Obtained by Clustering Analysis of Covalent Ligands

Dotted line: classification line; light face text: training set ligands; bold face text with asterisk: test set ligands.

each ligand ($n=3$), and E_{orb} descriptors were obtained.

LDA LDA was then performed to confirm if covalent inhibitor candidates could be discriminated from the substrate-like compounds by ΔG_{wat} and E_{orb} . R⁵⁹⁾ version 3.4.0 (R Core Team 2017) and RStudio⁶⁰⁾ version 1.0.143 (RStudio Team 2016) were used for the analysis. The LDA model was built by the training set defined in the “Ligand Datasets” section. Prior probabilities were set to be equal in the inhibitor group (0.5) and substrate group (0.5). After the model building, the discriminant ability was evaluated by the percentage of correctly classified samples attained for the test sets.

Results

Preparation of Ligand Datasets Results of the clustering analysis are shown in Fig. 2. When the classification line (merge distance=0.9) was adopted and ligands in the centroid of each cluster were selected, two ligands were identified from each group as the test set ligands—nafamostat and isocoumarin4 from the inhibitor group, and suc-AAPR-*p*NA and K1 from the substrate group. Since the nafamostat-trypsin intermediate has different crystal structures (1GBT and 2AH4), three structures from the inhibitor group and two structures from the substrate group were selected as the test set. The training set ligands and the test set ligands are shown in Table 1 and Fig. 3, respectively.

Preparation of Acyl-Trypsin Intermediate Structures All the ligands with no corresponding crystal structures were modeled by flexible molecular superposition. Among the several poses obtained by the superposition, one pose per ligand was selected according to the norms mentioned in Experimental. The selected ligands were appropriately superposed to the reference structure (Fig. 4).

Hydration Thermodynamics Analysis After obtaining representative structures by MD simulations, WaterMapTM analyses were performed. As a result, several water sites were

Fig. 3. Test Set Ligands

Dotted rectangular boxes: leaving group.

Fig. 4. Structures of Covalent Ligands in Trypsin Modeled by Flexible Molecular Superposition

Stick: reference ligand, prolyl-arginyl-Ser195 moiety in suc-AAPR trypsin (2AGE); Line: covalent ligands superposed to the reference ligand (green: substrates; pink: inhibitors).

located around the catalytic sites of acyl-trypsin intermediates (an example of WaterMap™ result is shown in Fig. 5). ΔG , ΔH , $-T\Delta S$, occupancy, and #HB of hydrolytic water molecule were then investigated. Two nafamostat-trypsin intermediate structures (1GBT and 2AH4) exhibited almost the same ΔG_{wat} , ΔH_{wat} , $-T\Delta S_{\text{wat}}$, occupancy, and #HB values, indicating that the WaterMap™ simulations were reasonable. ΔG_{wat} of the training set and the test set are shown in Table 1 and Table S1, respectively.

Semi-empirical MO Calculations From the result of semi-empirical MO calculations, antibonding π^* MOs of carbonyl C=O in covalently modified catalytic serine were first identified. Then, the orbital energy values of the antibonding MO were investigated. Those values were averaged for each ligand. Results of the training set and the test set are shown in Fig. 6 and Fig. S1, respectively.

Fig. 5. Predicted Water Sites around the Catalytic Site in a Representative Structure of Acyl-Trypsin Intermediate of Gabexate

Gray stick: Catalytic residues (His57 and Ser195) and residues consists of oxy-anion hole; Orange stick: Covalent ligand (gabexate); Gray line: Trypsin residues around the covalent ligand; Spheres: Predicted water sites in 5 Å of the C_{acyl} atom (green: water site with favorable ΔG_{wat} ; orange: water site with unfavorable ΔG_{wat}); HW: Hydrolytic water site.

LDA All ΔG_{wat} values were positive, which indicates that hydrolytic water molecules would obtain unfavorable ΔG in acyl-trypsin intermediates. The relationships between k_{cat} and ΔG_{wat} were investigated, and it was found that when k_{cat} values were lower, the ΔG_{wat} values were also lower with a few exceptions. Because of these exceptions, the substrate group and the inhibitor group cannot be discriminated only by ΔG_{wat} (Fig. 7a). The relationships between k_{cat} and E_{orb} were also investigated, and it was found that when k_{cat} values were lower, E_{orb} values were mostly higher. This result is reason-

Table 1. Training Set Ligands and Related Data

No.	Ligand name/ligand structure ^{a)} / acyl intermediate structure	Group	PDB ID	k_{cat} (s ⁻¹)	ΔG_{wat} (kcal/mol)
1	Lactone3	Inhibitor	—	3.3×10^{-5} (ref. 27)	1.3
2	Isocoumarin3	Inhibitor	—	5.0×10^{-5} (ref. 28)	2.6
3	Isocoumarin1	Inhibitor	—	5.3×10^{-4} (ref. 28)	4.2
4	Gabexate	Inhibitor	3VPK (ref. 29)	8.0×10^{-4b}	2.2
5	K3	Substrate	—	7.7×10^{-1} (ref. 30)	3.3
6	R3	Substrate	—	6.7 (ref. 30)	4.2

able, because when antibonding π^* MO of carbonyl C=O in covalently modified catalytic serine has higher E_{orb} (higher energy of unoccupied orbital), the nucleophilic attack by the hydrolytic water molecule has a higher barrier. Nevertheless, the substrate water group and the inhibitor group cannot be fully discriminated only by E_{orb} (Fig. 7b).

To clearly discriminate the inhibitor group from the sub-

strate group, the LDA model was built with the training set ligands using ΔG_{wat} and E_{orb} . As a result, the following equation (f) that can discriminate the inhibitor group and the substrate group was obtained (Fig. 7c).

$$f = 0.70 \cdot \Delta G_{\text{wat}} - 0.34 \cdot E_{\text{orb}} - 0.58$$

Table 1. Continued

No.	Ligand name/ligand structure ^{a)} / acyl intermediate structure	Group	PDB ID	k_{cat} (s ⁻¹)	ΔG_{wat} (kcal/mol)
7	K2	Substrate	—	16.2 (ref. 30)	3.4
8	suc-AAPK-pNA	Substrate	2AGG (ref. 31)	41 (ref. 31)	3.0
9	R1	Substrate	—	54.3 (ref. 30)	3.2
10	R2	Substrate	—	106 (ref. 30)	4.2

Abbreviations: suc-AAPK-pNA, succinyl-alanyl-alanyl-prolyl-lysyl-*para*-nitroanilide; suc-AAPR-pNA, succinyl-alanyl-alanyl-prolyl-arginyl-*para*-nitroanilide. *a*) Ligand structures are inside the solid rectangular boxes. Leaving groups are shown in the dotted rectangular boxes. *b*) k_{cat} value of gabexate was estimated from the experimental catalytic constant (k_{cat}) of nafamostat³²⁾ and ratio of the experimental half-life ($t_{1/2}$) of nafamostat and gabexate in human plasma.³³⁾

If $f < 0$, the ligand is classified as the inhibitor group; if $f > 0$, the ligand is classified as the substrate group. The LDA model correctly classified all the training set ligands into those two groups with no exception—there were no “false inhibitors” and “false substrates.”

Then, the discriminant ability of the model was evaluated by the test set ligands (external validation). The LDA model correctly classified 100.0% of the ligands from the test set with no exception—there were no “false inhibitors” and “false substrates.” Consequently, LDA using ΔG_{wat} and E_{orb} was shown to be effective for the discrimination of covalent inhibitor candidates from substrate-like compounds.

Discussion

Computer-aided drug design generally includes virtual screening and *de novo* drug design. In virtual screening, the potency of large libraries of drug-like compounds to act against a target is tested. Then, the compounds expected to be drug candidates are experimentally tested. In the *de novo*

drug design approach, the 3D structure of the receptor is used to design novel molecules on the basis of ligand-growing programs and the intuition of the medicinal chemist.⁶¹⁾ The method presented in this study is applicable mainly for the former.

In the previous report,¹¹⁾ we reported that ΔG_{wat} could be a useful metric to discriminate a covalent inhibitor candidate from substrate-like ligands. In this study, semi-empirical MO calculation, a traditional method to analyze the energetics of an enzymatic reaction, was also used together with hydration thermodynamics analysis to complement the discriminant ability of ΔG_{wat} . Further, acyl-trypsin intermediate structures with no crystal structures were modeled by flexible molecular superposition to show that the proposed method is suitable for ligands with no crystal structures. Linear discriminant analysis by ΔG_{wat} and E_{orb} was then performed, and as a result, substrate-like compounds were completely excluded from covalent inhibitor candidates.

The number of “false inhibitors” are generally favored to be as low as possible, because this number represents “inac-

Fig. 6. Antibonding π^* Molecular Orbital of Carbonyl C=O in Covalently-Modified Catalytic Serine in Representative Training Set Ligand Structures of Acyl-Trypsin Intermediates

Grey arrows: antibonding π^* molecular orbital on C_{acyl} atom.

tive inhibitor candidates” that would be sent to biological assays. There were no incorrectly classified inhibitors in both the training set and the test set, which would be an ideal rate of failure—the time and the costs for the assays are to be reduced. Therefore, the LDA model built by ΔG_{wat} and E_{orb} should be an effective strategy for the screening of mechanism-based reversible covalent inhibitors. It was also demonstrated that covalent ligands with no corresponding crystal structures could be analyzed by using flexible molecular superposition. This indicates that our method can be also applied for wider diversities of compounds with no corresponding crystal structures.

It was found that k_{cat} is discriminated by ΔG_{wat} with few exceptions (for example, isocoumarin1). It suggests that ΔG_{wat}

acquired from the enzymatic environment must be one of the activation energy of hydrolytic water molecule for the deacylation reaction in a serine protease. In the catalysis of substrate-like compounds, hydrolytic water molecule must be “hot water”⁶²⁾—the water molecule must be activated by unfavorable ΔG_{wat} acquired from an enzymatic environment, and with highly activated water molecule, a higher rate of hydrolysis (k_{cat}) would be exhibited. Conversely, in the catalysis of covalent inhibitor candidates, the hydrolytic water molecule must be “cold water”⁶²⁾—the water molecule must not be activated with ΔG_{wat} in an enzymatic environment, and with hydrolytic water molecule that was not very activated, a lower rate of hydrolysis (k_{cat}) would be exhibited. There are two possible causes for the low ΔG_{wat} , *i.e.*, low ΔH_{wat} (stable water

Fig. 7. Linear Discriminant Analysis by ΔG_{wat} and E_{orb} Descriptors

(a) Linear discriminant analysis only with ΔG_{wat} descriptor (not successful); (b) linear discriminant analysis only with E_{orb} descriptor (not successful); (c) linear discriminant analysis with both ΔG_{wat} and E_{orb} descriptors (successful).

molecule) or low $-T\Delta S_{\text{wat}}$ (unconstrained water molecule). It can be easily understood that stable or unconstrained water molecules are cold water molecules that rarely attack the C_{acyl} atom—with such hydrolytic water molecules, hydrolysis would become slower.

In general, ligands of enzymes and other proteins are spontaneously bound to and stabilized by an enzyme ($\Delta G_{\text{bind}} < 0$). However, water molecules, the other factor of the hydrolysis in hydrolases, are not always stabilized on a protein surface—billiard states of the water molecules may squeeze water molecule onto the protein surfaces that are not always hydrophilic. This study clearly shows that all the hydrolytic water molecules in acyl-trypsin intermediates are “unhappy” ($\Delta G_{\text{wat}} > 0$) at the hydrolytic water site to some extent, and the ΔG_{wat} tends to be higher in the substrates than in the inhibitors. This is probably why hydrolytic water molecules play a unique role in the hydrolytic enzymes.

The MO theory has been applied for the enzymatic reactions. One example is the study on the structure–activity relationship of ketone protease inhibitors.⁶³⁾ In the study, the orbital energy of the π^* lowest unoccupied molecular orbital (LUMO) of the carbonyl of ketone-included inhibitors was calculated in the Hartree–Fock level. The metric was used together with the electric charge on the β -nitrogen atom in the cyclic systems bonding next to the ketone group, and with those two metrics, the inhibitor constant (K_i) of those inhibitors was successfully predicted. The other recent example is the study on α -chymotrypsin catalysis, in which the

Fig. 8. ΔH_{wat} , $-T\Delta S_{\text{wat}}$ and E_{orb} of Each Covalent Inhibitors

PM7 Hamiltonian²⁶⁾ with the COSMO solvation method (the same method for calculating E_{orb} in this study) was used. As a result, most of the results of the study agreed with those of previously reported studies including *ab initio* calculations. Considering the high-throughput *in silico* screening with a lot of compounds, semi-empirical MO calculations with less computational cost are more preferable than time-consuming *ab initio* calculations.

In this study, it was shown that, when the orbital energy of the antibonding π^* molecular orbital of carbonyl C=O (E_{orb}) is higher, hydrolysis tends to be slower. In the deacylation of

Fig. 9. Relationship between $-T\Delta S_{\text{wat}}$ and $\text{Log}(k_{\text{cat}})$ of Covalent Inhibitors

Approximate line: $\text{Log}(k_{\text{cat}}) = -0.79 \cdot (-T\Delta S_{\text{wat}}) - 2.2$.

serine proteases, nucleophilic attack by deprotonated hydrolytic water molecules is the rate-determining step,⁹⁾ and in this step, the MO of the lone pair of the hydrolytic water molecule and the antibonding π^* MO of carbonyl C=O first interacts. The difference in the energy of the two reactive MOs (ΔE) is a metric of chemical reactivity, and in this case, the orbital energy of the nucleophile—hydrolytic water molecule—is thought to be almost constant for all the ligands. Therefore, it is reasonable that the orbital energy of the antibonding π^* MOs (E_{orb}) was shown to be a good metric of hydrolytic reactivity.

Although this study mainly focused on the method for screening, a guideline for the optimization and modification of a covalent inhibitor candidate obtained through the screening is also required. By investigating ΔH_{wat} , $-T\Delta S_{\text{wat}}$, and E_{orb} , we found three key patterns to be a mechanism-based reversible covalent inhibitor (Fig. 8): (1) negative or nearly zero ΔH_{wat} , which was found in nafamostat (IGBT and 2AH4), lactone3, isocoumarin4, and isocoumarin3; (2) small $-T\Delta S_{\text{wat}}$, which was found in isocoumarin1 and gabexate (3VPK); (3) relatively high E_{orb} , which was found especially in gabexate (3VPK). From those patterns, we discuss that there are three strategies for optimizing covalent inhibitors: (1) decreasing ΔH_{wat} , (2) decreasing $-T\Delta S_{\text{wat}}$, and (3) increasing E_{orb} . The first strategy should mainly be realized by increasing the average number of hydrogen bonds between acyl-trypsin and hydrolytic water molecules (#HB), because ΔH_{wat} and #HB were found to present a negative correlation. The second strategy should be realized by decreasing the occupancy of the water molecule at the hydrolytic water site, because $-T\Delta S_{\text{wat}}$ and occupancy were found to present positive correlation. To decrease the occupancy, the number of “cul-de-sac” residues around the hydrolytic water site that would prevent escaping water molecules from the site should be decreased (the first and second strategy has also been suggested in the previous report¹¹⁾). The third strategy should be realized when electron-rich functional groups, such as aromatic groups or alkyl chains, are bonded directly or near the C_{acyl} atom. Those electron-rich groups would increase E_{orb} and decrease the electrophilicity of the C_{acyl} atom.

Among the various strategies, we should determine the most important index to identify a stronger inhibitor. Although at first glance, it may seem contradictory to the sec-

ond strategy, increasing $-T\Delta S_{\text{wat}}$ is suggested to be the most effective index for the optimization and modification of the covalent inhibitor candidates. It was found that when $-T\Delta S_{\text{wat}}$ is higher, $\text{Log}(k_{\text{cat}})$ tends to be lower (Fig. 9). This trend is interpreted as follows: when $-T\Delta S_{\text{wat}}$ is higher, the rotational freedom of the water molecule at the hydrolytic water site is limited. Limited rotational freedom leads to a limited chance for interaction between the highest occupied MO (HOMO) of the hydrolytic water molecule and the antibonding π^* MO of carbonyl C=O—which might be related to the slow rates of hydrolysis. In other words, considering entropy–enthalpy compensation⁶⁴⁾—the linear relationship between the enthalpies and entropies of activation—a stronger inhibitor seems to be realized when an inhibitor has stable hydrolytic water molecule (small ΔH_{wat}) rather than unconstrained hydrolytic water molecule (small $-T\Delta S_{\text{wat}}$). In this discussion, however, covalent inhibitor candidates for the analysis were slightly limited for the quantitative analysis; even so, this should be an important preliminary discussion for the future studies.

As a conclusion, LDA with ΔG_{wat} and E_{orb} would possibly be an effective method for the screening of covalent inhibitors in hydrolases; further, $-T\Delta S_{\text{wat}}$ could be an index to estimate covalent inhibitor potency. However, the developed method was validated only for trypsin, and the optimization method of covalent inhibitor candidates using $-T\Delta S_{\text{wat}}$ was evaluated with a slightly limited number of ligands. Future works should therefore clarify (1) if LDA with ΔG_{wat} and E_{orb} would also be useful for other hydrolases, and (2) if $-T\Delta S_{\text{wat}}$ would be useful for estimating covalent inhibitor potencies with increased number of ligands. Those investigations will append successful examples, and will increase reliability of the method for the discovery of mechanism-based reversible covalent inhibitors presented in this study.

Conflict of Interest The authors declare no conflict of interest.

Supplementary Materials The online version of this article contains supplementary materials. Figure S1. Antibonding π^* Molecular Orbital of Carbonyl C=O in Covalently Modified Catalytic Serine in Representative Test Set Ligand Structures of Acyl-Trypsin Intermediates. Gray arrows: antibonding π^* molecular orbital on C_{acyl} atom.

References

- 1) Singh J., Petter R. C., Baillie T. A., Whitty A., *Nat. Rev. Drug Discov.*, **10**, 307–317 (2011).
- 2) Bauer R. A., *Drug Discov. Today*, **20**, 1061–1073 (2015).
- 3) Poordad F., McCone J. Jr., Bacon B. R., Bruno S., Manns M. P., Sulkowski M. S., Jacobson I. M., Reddy K. R., Goodman Z. D., Boparai N., DiNubile M. J., Sniukiene V., Brass C. A., Albrecht J. K., Bronowicki J.-P., *N. Engl. J. Med.*, **364**, 1195–1206 (2011).
- 4) Villhauer E. B., Brinkman J. A., Naderi G. B., Burkey B. F., Dunning B. E., Prasad K., Mangold B. L., Russell M. E., Hughes T. E., *J. Med. Chem.*, **46**, 2774–2789 (2003).
- 5) Bull H. G., Garcia-Calvo M., Andersson S., Baginsky W. F., Chan H. K., Ellsworth D. E., Miller R. R., Stearns R. A., Bakshi R. K., Rasmussen G. H., Tolman R. L., Myers R. W., Kozarich J. W., Harris G. S., *J. Am. Chem. Soc.*, **118**, 2359–2365 (1996).
- 6) Linker R. A., Lee D. H., Ryan S., van Dam A. M., Conrad R., Bista P., Zeng W., Hronowsky X., Buko A., Chollate S., Ellrichmann G., Brück W., Dawson K., Goelz S., Wiese S., Scannevin R. H., Luka-

- shev M., Gold R., *Brain*, **134**, 678–692 (2011).
- 7) Zhu K., Borrelli K. W., Greenwood J. R., Day T., Abel R., Farid R. S., Harder E., *J. Chem. Inf. Model.*, **54**, 1932–1940 (2014).
- 8) Ishida T., Kato S., *J. Am. Chem. Soc.*, **126**, 7111–7118 (2004).
- 9) Topf M., Richards W. G., *J. Am. Chem. Soc.*, **126**, 14631–14641 (2004).
- 10) Hirano Y., Shinmoto N., Hata M., Tsuda M., Hoshino T., *J. Mol. Struct. THEOCHEM*, **757**, 19–27 (2005).
- 11) Masuda Y., Yamaotsu N., Hirono S., *Chem. Pharm. Bull.*, **65**, 889–892 (2017).
- 12) Bodnarchuk M. S., *Drug Discov. Today*, **21**, 1139–1146 (2016).
- 13) Lazaridis T., *J. Phys. Chem. B*, **102**, 3542–3550 (1998).
- 14) Young T., Abel R., Kim B., Berne B. J., Friesner R. A., *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 808–813 (2007).
- 15) Abel R., Young T., Farid R., Berne B. J., Friesner R. A., *J. Am. Chem. Soc.*, **130**, 2817–2831 (2008).
- 16) Beuming T., Farid R., Sherman W., *Protein Sci.*, **18**, 1609–1619 (2009).
- 17) Pearlstein R. A., Sherman W., Abel R., *Prot. Struct. Funct. Bioinf.*, **81**, 1509–1526 (2013).
- 18) Turk B., *Nat. Rev. Drug Discov.*, **5**, 785–799 (2006).
- 19) Bachovchin D. A., Cravatt B. F., *Nat. Rev. Drug Discov.*, **11**, 52–68 (2012).
- 20) Sprang S., Standing T., Fletterick R. J., Stroud R. M., Finer-Moore J., Xuong N. H., Hamlin R., Rutter W. J., Craik C. S., *Science*, **237**, 905–909 (1987).
- 21) Hedstrom L., *Chem. Rev.*, **102**, 4501–4523 (2002).
- 22) Dixon S. L., Smondyrev A. M., Knoll E. H., Rao S. N., Shaw D. E., Friesner R. A., *J. Comput. Aided Mol. Des.*, **20**, 647–671 (2006).
- 23) Dixon S. L., Smondyrev A. M., Rao S. N., *Chem. Biol. Drug Des.*, **67**, 370–372 (2006).
- 24) Daggett V., Schröder S., Kollman P., *J. Am. Chem. Soc.*, **113**, 8926–8935 (1991).
- 25) Brandt W., Ludwig O., Thondorf I., Barth A., *Eur. J. Biochem.*, **236**, 109–114 (1996).
- 26) Stewart J. J. P., *J. Mol. Model.*, **23**, 154 (2017).
- 27) Rai R., Katzenellenbogen J. A., *J. Med. Chem.*, **35**, 4150–4159 (1992).
- 28) Kam C. M., Fujikawa K., Powers J. C., *Biochemistry*, **27**, 2547–2557 (1988).
- 29) Masuda Y., Nitani Y., Mizutani R., Noguchi S., *Prot. Struct. Funct. Bioinf.*, **81**, 526–530 (2013).
- 30) Roberts D. V., Elmore D. T., *Biochem. J.*, **141**, 545–554 (1974).
- 31) Radisky E. S., Lee J. M., Lu C. J. K., Koshland D. E. Jr., *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 6835–6840 (2006).
- 32) Ramjee M. K., Henderson I. M. J., McLoughlin S. B., Padova A., *Thromb. Res.*, **98**, 559–569 (2000).
- 33) Takeda K., *J. Pancreas*, **8**, 526–532 (2007).
- 34) Mangel W. F., Singer P. T., Cyr D. M., Umland T. C., Toledo D. L., Stroud R. M., Pflugrath J. W., Sweet R. M., *Biochemistry*, **29**, 8351–8357 (1990).
- 35) Duan J., Dixon S. L., Lowrie J. F., Sherman W., *J. Mol. Graph. Model.*, **29**, 157–170 (2010).
- 36) Sastry M., Lowrie J. F., Dixon S. L., Sherman W., *J. Chem. Inf. Model.*, **50**, 771–784 (2010).
- 37) Schrödinger Release 2015-3: Canvas, Schrödinger, LLC, New York, NY, 2015.
- 38) Weininger D., Weininger A., Weininger J. L., *J. Chem. Inf. Comput. Sci.*, **29**, 97–101 (1989).
- 39) Schrödinger Release 2015-3: Phase, Schrödinger, LLC, New York, NY, 2015.
- 40) Sastry G. M., Adzhigirey M., Day T., Annabhimoju R., Sherman W., *J. Comput. Aided Mol. Des.*, **27**, 221–234 (2013).
- 41) Schrödinger Release 2015-3: Protein Preparation Wizard, Schrödinger, LLC, New York, NY, 2015.
- 42) Schrödinger Release 2015-3: Maestro, Schrödinger, LLC, New York, NY, 2015.
- 43) Jacobson M. P., Pincus D. L., Rapp C. S., Day T. J. F., Honig B., Shaw D. E., Friesner R. A., *Prot. Struct. Funct. Bioinf.*, **55**, 351–367 (2004).
- 44) Jacobson M. P., Friesner R. A., Xiang Z., Honig B., *J. Mol. Biol.*, **320**, 597–608 (2002).
- 45) Schrödinger Release 2015-3: Prime, Schrödinger, LLC, New York, NY, 2015.
- 46) Northrop J. H., Kunitz M., *J. Gen. Physiol.*, **16**, 295–311 (1932).
- 47) Olsson M. H. M., Søndergaard C. R., Rostkowski M., Jensen J. H., *J. Chem. Theory Comput.*, **7**, 525–537 (2011).
- 48) Banks J. L., Beard H. S., Cao Y., Cho A. E., Damm W., Farid R., Felts A. K., Halgren T. A., Mainz D. T., Maple J. R., Murphy R., Philipp D. M., Repasky M. P., Zhang L. Y., Berne B. J., Friesner R. A., Gallicchio E., Levy R. M., *J. Comput. Chem.*, **26**, 1752–1780 (2005).
- 49) Schrödinger Release 2015-3: Impact, Schrödinger, LLC, New York, NY, 2015.
- 50) Bowers K. J., Chow E., Xu H., Dror R. O., Eastwood M. P., Gregersen B. A., Klepeis J. L., Kolossvary I., Moraes M. A., Sacerdoti F. D., Salmon J. K., Shan Y., Shaw D. E., Proceedings of the ACM/IEEE Conference on Supercomputing (SC06), Tampa, Florida, November 11–17 (2006).
- 51) Jorgensen W. L., Chandrasekhar J., Madura J. D., Impey R. W., Klein M. L., *J. Chem. Phys.*, **79**, 926–935 (1983).
- 52) Harder E., Damm W., Maple J., Wu C., Reboul M., Xiang J. Y., Wang L., Lupyan D., Dahlgren M. K., Knight J. L., Kaus J. W., Cerutti D. S., Krilov G., Jorgensen W. L., Abel R., Friesner R. A., *J. Chem. Theory Comput.*, **12**, 281–296 (2016).
- 53) Schrödinger Release 2015-3: WaterMap, Schrödinger, LLC, New York, NY, 2015.
- 54) Schrödinger Release 2015-3: MacroModel, Schrödinger, LLC, New York, NY, 2015.
- 55) Klamt A., Schüürmann G., *J. Chem. Soc., Perkin Trans. 2*, **1993**, 799–805 (1993).
- 56) Stewart J. J. P., MOPAC2016, Version: 16.230, Stewart Computational Chemistry: <<http://OpenMOPAC.net>>, cited July 30, 2017.
- 57) Stewart J. J. P., *J. Mol. Model.*, **19**, 1–32 (2013).
- 58) “Jmol: an open-source Java viewer for chemical structures in 3D.”: <<http://www.jmol.org/>>, cited July 30, 2017.
- 59) R Core Team (2017). “R: A language and environment for statistical computing. R Foundation for Statistical Computing,” Vienna, Austria.: <<https://www.R-project.org/>>, cited July 30, 2017.
- 60) RStudio Team (2016). “RStudio: Integrated Development for R.” RStudio, Inc., Boston, MA.: <<http://www.rstudio.com/>>, cited July 30, 2017.
- 61) Lionta E., Spyrou G., Vassilatis D. K., Cournia Z., *Curr. Top. Med. Chem.*, **14**, 1923–1938 (2014).
- 62) Spyraakis F., Ahmed M. H., Bayden A. S., Cozzini P., Mozzarelli A., Kellogg G. E., *J. Med. Chem.*, **60**, 6781–6827 (2017).
- 63) Chan A. W. E., Golec J. M. C., *Bioorg. Med. Chem.*, **4**, 1673–1677 (1996).
- 64) Lumry R., Rajender S., *Biopolymers*, **9**, 1125–1227 (1970).