Chemistry & Biology, Vol. 9, 1085–1094, October, 2002, ©2002 Elsevier Science Ltd. All rights reserved. PII S1074-5521(02)00238-7

Small Molecule Affinity Fingerprinting: a Tool for Enzyme Family Subclassification, Target Identification, and Inhibitor Design

Doron C. Greenbaum,[1] William D. Arnold,[1] Felice Lu,[1] Linda Hayrapetian,[2] Amos Baruch,[2] Jennifer Krumrine,[1] Samuel Toba,[1] Kareem Chehade,[2] Dieter Brömme,[3] Irwin D. Kuntz,[1] and Matthew Bogyo[2,4]

[1]Department of Pharmaceutical Chemistry
[2]Department of Biochemistry and Biophysics
University of California, San Francisco
513 Parnassus Avenue
San Francisco, California 94143
[3]Department of Human Genetics
Mount Sinai School of Medicine
Fifth Avenue at 100 Street
New York, New York 10029

Summary

Classifying proteins into functionally distinct families based only on primary sequence information remains a difficult task. We describe here a method to generate a large data set of small molecule affinity fingerprints for a group of closely related enzymes, the papain family of cysteine proteases. Binding data was generated for a library of inhibitors based on the ability of each compound to block active-site labeling of the target proteases by a covalent activity based probe (ABP). Clustering algorithms were used to automatically classify a reference group of proteases into subfamilies based on their small molecule affinity fingerprints. This approach was also used to identify cysteine protease targets modified by the ABP in complex proteomes by direct comparison of target affinity fingerprints with those of the reference library of proteases. Finally, experimental data were used to guide the development of a computational method that predicts small molecule inhibitors based on reported crystal structures. This method could ultimately be used with large enzyme families to aid in the design of selective inhibitors of targets based on limited structural/function information.

Introduction

The recent genomics revolution has provided us with the first low-resolution roadmap of the human genome. However, the true challenge lies in using this raw sequence information to create a better understanding of the role of specific gene products in both normal and disease processes. Functional genomics efforts have begun to address this challenge using sequence-alignment algorithms and transcriptional profiling as a way to link biological functions to specific genes and gene products [1]. Indeed, this process has lead to the annotation of a substantial number of enzyme and protein families. In many cases, these families will serve as a starting point in the process of target selection for the development of preclinical drug candidates. However, many protein families are populated with dozens of closely related members. For example, the protease family alone comprises 1%–2% of the human genome and represents over 500 enzymes grouped within only a few distinct subfamilies. Therefore, potential drug targets such as these must be viewed not as single entities but as members of closely related protein networks. Therapeutic design must focus not only on issues of potency toward a single target but also, and often more importantly, on selectivity within the context of a target's nearest functional relatives.

Traditionally, the problem of specificity has been addressed using medicinal chemistry to generate compounds that have been optimized for a single protein target. Correlation of structural elements of small molecule leads with their inhibition potencies is used to generate structure activity relationships (SARs). These data can be used to rank individual compounds and ultimately to sort out the best candidates for further development. To aid in this process, several groups have developed complementary in silico methods to define molecular similarity among a class of protein targets [2, 3]. Additionally, computational methods have been developed that allow small molecule binding to be addressed by virtual docking to a protein active site [4–6]. From these computational SAR studies, a set of physicochemical descriptors can be generated that define the binding properties of many related small molecule inhibitors. Ultimately, such computational approaches allow a large number of theoretical compounds to be virtually assayed prior to embarking on costly and time consuming medicinal chemistry efforts.

In addition to providing a starting point for lead optimization, SAR data also provide information that can be used to generally define the topology of the small molecule binding pocket of a target protein. Furthermore, compilation of SAR data obtained from chemical library screening against a set of proteins provides affinity fingerprints for each target. As an increasing number of diverse compounds are assayed against these targets, the fingerprints that are generated become more refined. If these fingerprints become sufficiently unique, they can be used to establish subtle differences among members of a large protein family with a high degree of sequence homology.

Several methods for protein classification based on affinity fingerprints have been proposed. One such method relies upon a training set of inhibitors that is screened against a panel of disparate proteins to predict affinity fingerprints for other nonrelated proteins. Ultimately, this method could be used to allow chemists to quickly predict pharmacophores within a chemical library that will serve as lead compounds for further development [7, 8]. Yet another classification method has introduced structure activity relationship homologies (SARAH) as a means to cluster proteins within a

[4]Correspondence: mbogyo@biochem.ucsf.edu

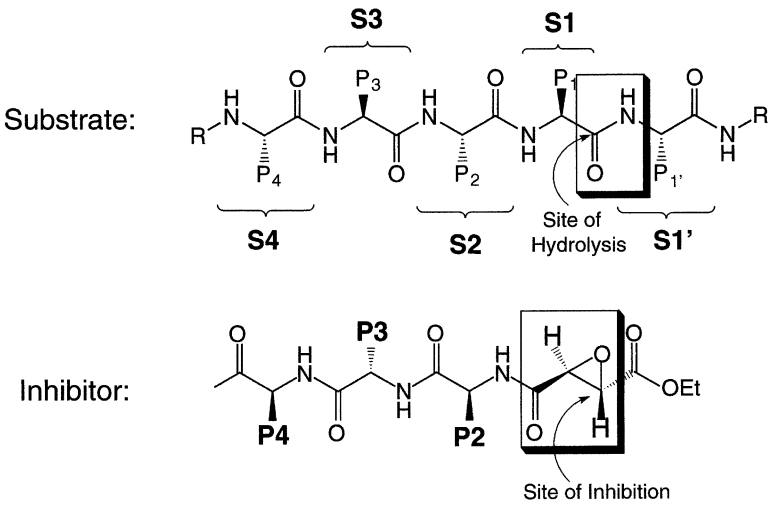

Figure 1. Comparison of Binding Mode of Peptidyl Epoxide Inhibitors and Peptide Substrates

Peptidyl epoxides bind to cysteine protease active sites in a manner analogous to a peptide substrate. The three amino acid side chains adjacent to the epoxide, termed the P2, P3, and P4 residues, align in the active site such that they occupy the S2, S3, and S4 binding pockets. Note that no side chain fills the S1 pocket due to the structure of the epoxide building block.

family. The kinase family of enzymes was used to highlight the utility of inhibitor fingerprinting as a rapid classification method for members of this large family of highly related proteins [9]. Once a functional classification is established based on SARAH, it becomes possible to group newly sequenced kinases into chemical subgroups to optimize the drug-screening process. Furthermore, this method of classification provides critical information concerning the "nearest neighbors" in the family that are likely to be of concern when trying to design a selective small molecule drug.

Here, we outline a combined chemical- and computational-based approach to generate and analyze affinity fingerprints for the papain family of cysteine proteases. An affinity labeling methodology has been employed to assess the inhibitory characteristics of a set of small molecule libraries toward this panel of closely related protease targets. This resulting inhibition data set is a compilation of affinity fingerprints for the set of purified targets and was used as a method to classify individual family members. In addition, the identity of proteases from crude cellular lysates could be determined by clustering affinity fingerprints of "unknown" targets with the data set of purified targets. A computational protocol was then developed and used to generate predictions for cysteine proteases based on experimentally determined crystal structures. Ultimately, this method could aid the process of development of small molecule inhibitors for families of related targets when only limited structural and functional information is available.

Results and Discussion

Inhibitor Library Design

We have previously described a set of positional scanning libraries (PSLs) based on the epoxide electrophile scaffold found in the natural product E-64 [10, 11]. This scaffold can be used to generate compounds that are mechanism-based irreversible inhibitors of the papain family of cysteine proteases [12]. The compounds in these libraries are made up of a primary tripeptide backbone linked to a reactive epoxide electrophile. The amino acids found adjacent to the epoxide moiety are expected to occupy the S2–S4 binding pockets of the protease targets (termed the P2, P3, and P4 amino acids; Figure 1). The S2 pocket has been shown to be the primary site of substrate discrimination for this family of proteases [13].

Initially, three sets of PSLs were synthesized by fixing each of the P2, P3, and P4 positions with each of the 20 possible natural amino acids (minus cysteine and methionine, plus norleucine as a mimetic of methionine). A mixture of the same natural amino acids was used in the remaining two amino acid positions, resulting in 19 P2, P3, and P4 sublibraries, with each made up of a mixture of 361 compounds.

Inhibitor Screening

The three sets of PSLs were assayed against purified protease targets by competition with the radiolabeled active-site-directed probe ^{125}I-DCG-04. Samples were analyzed by SDS-PAGE followed by phosphorimaging to determine the intensity of labeled bands using a commercial software package (Figure 2). Competition (i.e., loss of labeling) was indicative of inhibition by the unlabeled library member. Competition assays are performed by preincubation of protease targets with inhibitor libraries followed by labeling with the general probe. Since the extent of inhibition by the inhibitor libraries is a function of preincubation and labeling times, these parameters had to be carefully controlled, and assays were performed in triplicate to confirm the run-to-run reproducibility of the assay. Furthermore, for this method to provide a valid readout, final concentration of inhibitors (10–50 μM) must be held in excess over concentrations of the target protease (100–300 nM) throughout the assay. Using this method it was possible to determine a percent competition for each fixed position library by determining the ratio of intensity of labeled bands in the treated samples to the intensity of the untreated control. These data were subsequently used to generate affinity fingerprints.

Covalent irreversible inhibitors such as the peptide epoxides function mechanistically through a two-step process involving an initial reversible binding event (measured as an equilibrium constant, Ki) followed by an

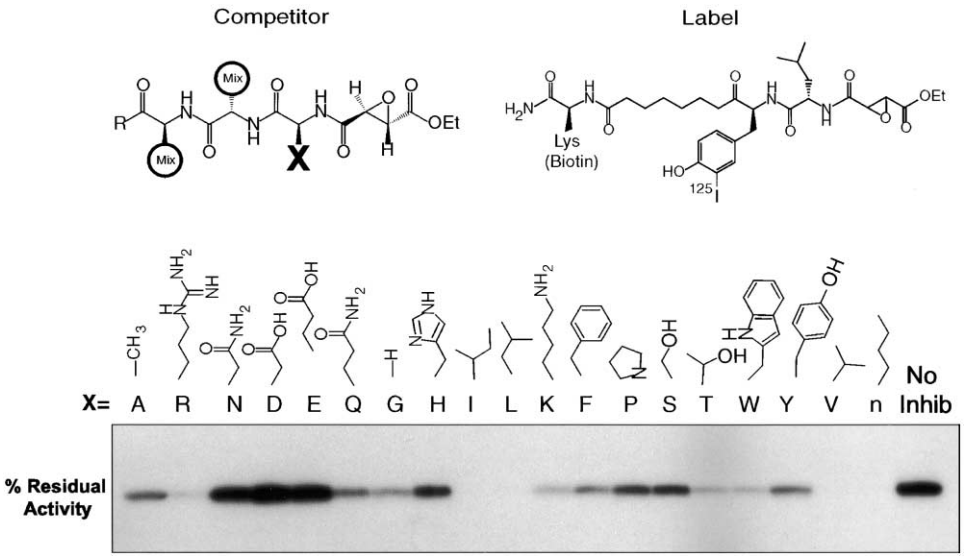

Figure 2. Methods for Generating Affinity Fingerprints

Example of an affinity fingerprint generated by screening of a P2 diverse peptide epoxide library. Purified cathepsin K was pretreated with individual constant P2 sublibraries (X position on competitor) followed by labeling with ^{125}I-DCG-04 (label). Samples were separated on a 12.5% SDS-PAGE gel and visualized by PhosphoImaging (Molecular Dynamics). Labeling intensity of each target relative to the control untreated sample was used to generate percent competition values. This method was used to generate competition values for multiple enzymes and for libraries with diversity at the P2, P3, and P4 positions on the inhibitor scaffold.

irreversible alkylation step (measured as rate constant kinact). Potency values for this class of inhibitors are expressed as a ratio of the Ki/kinact. Detailed kinetic studies of the peptide epoxides have shown that the rates of inactivation (kinact) remain relatively constant across structurally diverse inhibitor scaffolds [14]. As a result, competition data obtained for libraries of peptide epoxides provide mainly information that relates to the relative Ki values of an inhibitor for a given target. Furthermore, any small molecule that binds in the active site of a target will block the reversible binding step of the probe and will lead to loss of labeling (competition). Therefore, this method is suitable for screening of both reversible and irreversible inhibitors. In fact, similar screens with libraries of reversible cysteine protease inhibitors have been carried out for the parasitic protease target cruzain. These competition results were found to closely correlate with kinetic inhibition values obtained by standard substrate-based methods (D.C.G., M.B., and J. Ellman, unpublished results).

While substrate-based kinetic assays provide for high-throughput screening of targets, the competition-based method can be multiplexed to accommodate multiple targets in a single gel-based assay. Additionally, this screening method allows for rapid analysis of multiple related targets without the need to optimize substrate and kinetic conditions for each enzyme. Finally, the competition screen allows separation of the target from the substrates and small molecules in the screen, thereby eliminating problems of insoluble and intrinsically fluorescent compounds that can hinder an absorbance-based detection method. To increase the assay throughput, we have also designed a dot-blot-based readout for competition. In the case where a single protein target is screened, filtering of samples through a PVDF membrane provides a method to isolate and measure the amount of labeled target protein. This assay method circumvents the need for SDS-PAGE gels and allows the assay to be performed in a 96-well plate format (data not shown).

Affinity Fingerprint Analysis

This affinity-probe-based method of screening of PSLs has been validated by our laboratories in a representative crude proteome [11] and for a specific protease target [15]. These studies show that it is possible to use this screening method to rapidly identify selective inhibitors of protease targets. It was therefore of interest to apply the same set of PSLs to profiling the specificity of an expanded set of papain family enzymes. While this family of proteolytic enzymes has been extensively studied, most inhibitor SAR studies have been focused on a limited number of compounds screened against a small set of family members. It was therefore of interest to determine if a large data set could be used to classify this set of proteases into distinct subfamilies based on substrate/inhibitor binding.

PSLs were screened against a set of purified and recombinant papain family cysteine proteases that were obtained from commercial and public sources. To aid in the analysis of the data, numerical competition values were visualized by conversion to a color format using software developed by Eisen and coworkers designed for data generated from microarray analysis [16]. This software assigns colors based on the numerical competition values in the range from 0%–100%. Compounds that were potent inhibitors (i.e., 100% competition) were assigned a red (hot) color, while compounds that were weak inhibitors, showing little or no competition, were assigned a blue (cold) color. Compounds with intermediate activi-

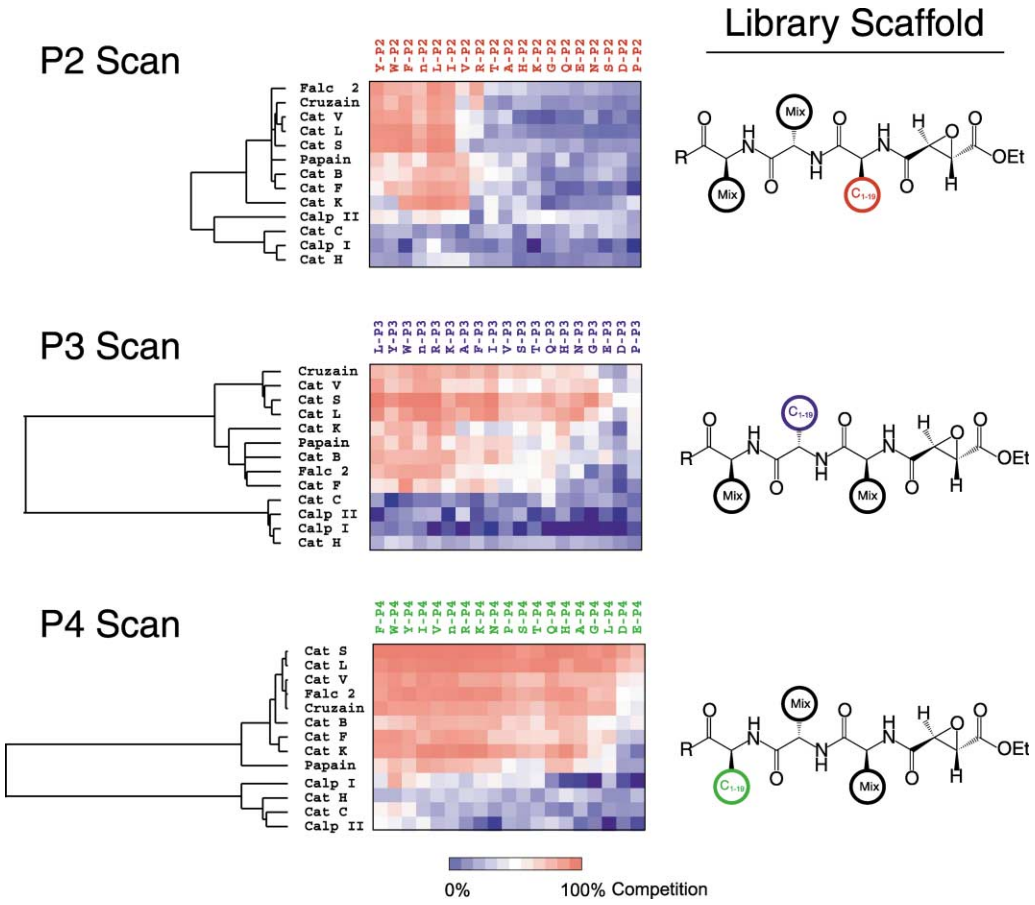

Figure 3. Cluster Analysis of Affinity Fingerprints for a Set of Papain Family Proteases: Subsite Specificities within the Active Sites

Inhibition data from screening of P2, P3, and P4 diverse inhibitor libraries (scaffold structures indicated on right of data panels). Sublibraries were composed of a single constant amino acid position that was varied through all natural amino acids (C_{1-19}) and two variable positions composed of a mixture of all 19 amino acids (mix). Competition data were obtained as describe in Figure 2 and were clustered and visualized using programs designed for analysis of microarray data (see Experimental Procedures). Colors indicate the potency of a sublibrary with the indicated fixed amino acid for a designated target protease. Potent (hot) inhibitors are assigned a red color, and weak or ineffective (cold) inhibitors are assigned a blue color. Target enzymes are arrayed along the *y* axis, and each of the constant amino acids is arrayed along the *x* axis. The tree structures at the left of the diagrams were obtained by hierarchical clustering and indicate the degree of similarity of enzymes as a function of the height of the lines connecting profiles. The color key is shown at the bottom. Amino acids are indicated by their single-letter code, with n used for norleucine.

ties were assigned lighter shades of red and blue, with white assigned to compounds with 50% inhibition. Furthermore, hierarchical clustering software was used to group the data based on similarities among profiles of enzymes (*y* axis) or small molecules (*x* axis).

Cluster analysis of inhibition data from each of the P2, P3, and P4 library sets against 12 papain family proteases revealed patterns of specificity for each of the three primary substrate binding pockets (Figure 3). The resulting specificity data agreed with previously reported findings identifying the P2 position as the primary site for enzyme-substrate interactions [13]. Furthermore, the S2 pocket of the papain family enzymes preferred many of the hydrophobic and aromatic amino acids, suggesting the need for a more diverse set of hydrophobic P2 residues in order to obtain distinct binding profiles for this class of enzymes.

A set of 41 hydrophobic nonnatural amino acids was selected and used to generate a nonnatural P2 library (for structures, see Supplemental Data). For this extended P2 library, each of the 41 nonnatural amino acids was held constant in the P2 position, while the P3 and P4 positions were composed of a mix of all possible natural amino acids. The mixture method was chosen rather than using general favorable binding P2 and P3 amino acids because this resulted in sublibraries that had greater overall utility for screening. These libraries were not biased in the P3 and P4 positions and therefore could be used to assay the contribution of the P2 element for virtually any cysteine protease target. In order to further increase the diversity of compounds for affinity fingerprinting, a second set of libraries was synthesized using the complete set of natural amino acid building blocks in the P2 position attached to the enantiomeric form of the epoxide electrophile (2R, 3R, versus 2S, 3S; Figure 4). Previous work has shown that this change in

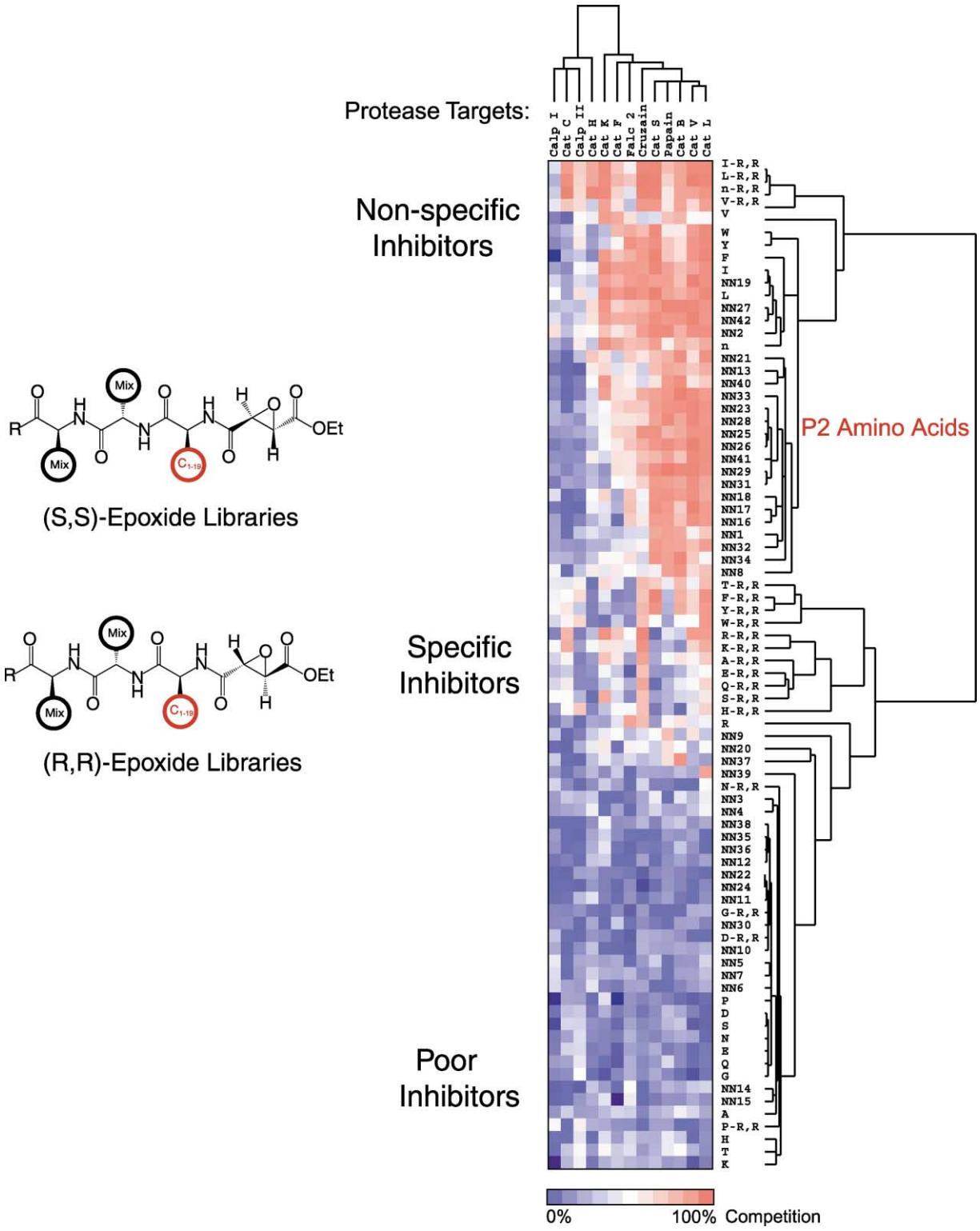

Figure 4. Cluster Analysis of an Extended P2 Diversity Library

A large set of P2 amino acids including the 19 natural amino acids and 41 nonnatural hydrophobic amino acids was selected and used to generate an extended P2 inhibitor library (structures and corresponding numerical assignments of the nonnatural amino acids can be found in the Supplemental Data). In addition to the set of 60 natural and nonnatural amino acids coupled to the epoxide moiety containing the (S,S) stereochemistry, the natural 19 amino acids were coupled to the enatiomeric form of the epoxide (R,R isomer; see structures at left). The resulting 79 sublibraries were assayed against the reference set of 12 papain family protease as described in Figures 2 and 3. Single-letter codes were used for natural amino acids, with n being assigned to norleucine. The 41 nonnatural amino acids were assigned arbitrary numbers (1–41) and listed with the NN prefix. Libraries containing the R,R enantiomer of the epoxide are listed with "R,R" following the single-letter amino acid code. Regions of weak binding, nonselective strong binding, and selective binding are labeled at the left.

stereochemistry is likely to favor binding of the inhibitors in the prime side of the active site, thus increasing the potential for finding binding pockets unique to each papain family protease [17].

The clustering of the extended P2 library data revealed underlying patterns of inhibition by grouping compounds with overall poor binding, promiscuous binding, or selective binding together (see annotation at left of clustergram in Figure 4). Grouping the data in this manner immediately identified P2 amino acids in the central region of the clustergram that conferred specificity for individual protease targets. Interestingly, the bulk of the amino acids found in this "specificity region" were non-natural amino acids and natural amino acids linked to the (R,R) enantiomer of the epoxide. These results suggest that changing the stereochemistry of the epoxide provided access to different binding sites in the protease active-site cleft. These differences are likely due to interactions of the R,R compounds with the prime-side binding pockets of the papain family proteases. This hypothesis will be confirmed through structural studies of inhibitor binding and will be the focus of future work.

This clustering methodology therefore shows that affinity fingerprinting data can be used to reveal information about the topology of each of these protease binding pockets. Ultimately a screen of a larger, more structurally diverse small molecule library is likely to provide a higher-resolution image of these inhibitor/enzyme interactions.

Identifying Enzymes from Crude Cellular Lysates

Another powerful application of this affinity-fingerprinting methodology is its ability to classify an unknown protease activity from a crude cell or tissue lysate by clustering its affinity fingerprint within a database of standard protease fingerprints. We have previously demonstrated the utility of activity-based probes as a means to profile cysteine protease activities within intact cells or crude cell lysates. This technology therefore allowed the extended P2 inhibitor library to be screened against several cysteine proteases in a crude cell extract [11].

The rat liver proteome was chosen for initial studies due to its high content of proteolytic enzymes and because the major protease activities in this sample were previously identified by purification and sequencing [11]. Total protein extracts were probed for cysteine protease activity using ^{125}I-DCG-04 (Figure 5A). Four major protease activities were observed by affinity labeling and SDS-PAGE analysis (Figure 5B). This profile exactly matched the results reported by our laboratory in an earlier publication [11], indicating that the labeling method is highly reproducible.

Affinity fingerprints were generated for each protease activity by pretreatment of extracts with inhibitor PSL sublibraries followed by affinity labeling. The resulting data sets were clustered with the database of extended P2 cysteine protease inhibition fingerprints (Figure 5C, black boxes). Protease band 2 clustered into a small subgroup of cathepsin proteases, with the greatest similarity to cathepsin B. The identity of this band was confirmed to be cathepsin B by isolation and sequencing by mass spectrometry [11]. Protease bands 3 and 4 had identical fingerprints and clustered together in the cluster tree as a distinct branch, which included cathepsin H. Again, this cluster-based assignment of bands 3 and 4 was confirmed by purification, sequencing, and identification of these two bands as differentially processed forms of cathepsin H [11]. Protease band 1, unlike the other proteases, clustered into its own branch and had no direct counterpart in the database. This protease activity was identified as cathepsin Z [11], an enzyme that was not fingerprinted and therefore had no reference points in the database. Thus, the clustering method was able to predict the identity of enzyme activities within a crude tissue lysate by virtue of their unique affinity fingerprints.

The results from this experiment highlight several strengths of combined inhibitor screening and clustering technology. First, the inhibitor libraries allow screening against cysteine proteases present in a crude cell and tissue proteome. The ability to use crude protein extracts, rather than recombinant or purified protein, greatly reduces the effort required to screen large inhibitor libraries and allows rapid lead identification for endogenously expressed enzymes. Second, the tight clustering of endogenous cathepsins with their recombinant counterparts suggests that this methodology could be used for rapid, crude characterization of unknown enzymes from complex protein samples without absolute knowledge of their identity.

Classification of Enzymes Based on Fingerprint Clustering

In addition to being useful for optimization of small molecule inhibitors, clustergrams of affinity fingerprints also yield functional information about the topology of the active site of the protein. The dendrogram that results from clustering of the library data using the programs Cluster and TreeView [16] pictorially describes the relationships amongst individual proteases. This dendrogram is analogous to homology trees that are generated through sequence alignments. However, it provides inhibitor-generated functional alignments, in contrast to traditional sequence alignments based on linear amino acid relationships.

For comparison, a dendrogram of proteases was generated using the sequence alignment program Clustal W and compared against the affinity-fingerprint alignment. The two dendrograms have overall similarities but upon closer inspection reveal significant differences (Figure 6). For example, cathepsin B and cathepsin C cluster together based on primary sequence alignments. Although these are both exoproteases, cathepsin B is a carboxypeptidase while cathepsin H is an aminopeptidase, and their true functions are highly divergent. The fingerprint clustering yields a more satisfying picture of the large functional difference between cathepsin B and C (Figure 6, red labels). On the other hand, sequence alignment of cathepsin K clusters it within a subfamily with cathepsins S, V, and L. However, affinity-fingerprint clustering identified cathepsin F as its closest neighbor and, therefore, the major concern for efforts to design cathepsin K selective inhibitors (Figure 6, green labels).

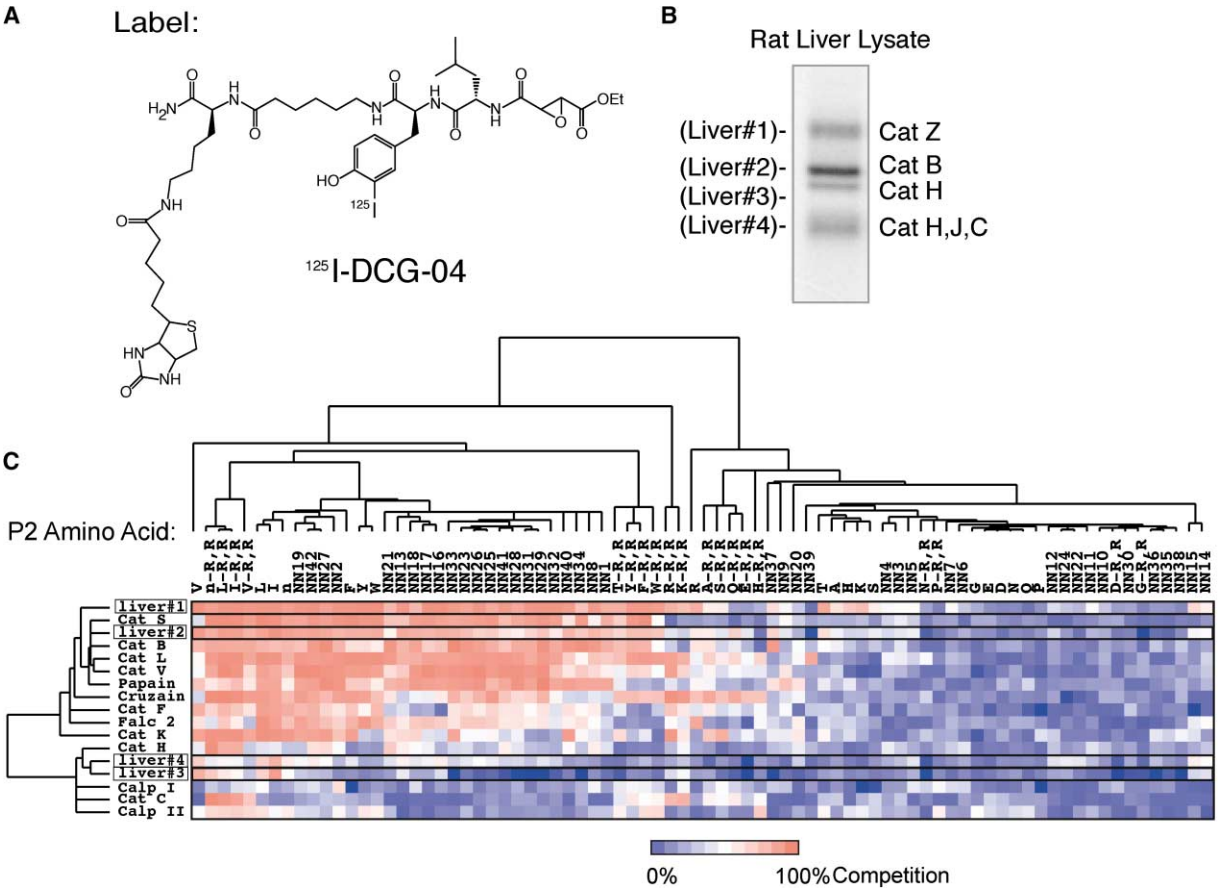

Figure 5. Identifying Unknown Proteases' Targets Using Fingerprint Clustering

(A) Structure of the activity based probe, ^{125}I-DCG-04.

(B) Profile of active papain family cysteine proteases in crude rat liver homogenates. "Unknown" proteases are labeled 1–4 (liver #1–liver #4 at left). The true identity of each protease was determined by mass-spectrometry-based sequencing and is listed for reference at right.

(C) Competition data obtained by treatment of crude homogenates with the extended P2 sublibraries described in Figure 4 followed by labeling with the probe. Competition data for each unknown (see black boxes) were added to the reference protease data, and the complete data set was clustered as described in Figures 3 and 4. Identity of the unknown proteases could be inferred by inspection of the closest neighbors in the vertical dendrogram shown on the left.

Furthermore, the fingerprint clustering identified cathepsins K, F, and H as the best candidates in this family of proteases for design of selective inhibitors due to the uniqueness of their specificity profiles (i.e., distinct branches in the clustering tree). Such information may also help to prioritize targets in large protein families based on the chances for successful development of selective inhibitors.

In Slico Generation of Affinity Fingerprints

The affinity fingerprints generated for a control set of cysteine proteases was also used to tailor the design of a computational protocol for generating in silico fingerprints based on structural data. A molecular docking scheme [18], which had proven successful for the design of both peptidic and nonpeptidic inhibitors in a series of serine proteases, was unable to distinguish specificity in the lysosomal cysteine proteases. We found that the covalent linkage between the inhibitor and the enzyme necessitated a complete molecular mechanical forcefield for proper inhibitor placement. The DOCK program, however, employs only the intermolecular van der Waals and coulombic terms as an energy scoring function. We therefore combined docking with molecular dynamics (MD) to develop a new strategy in the spirit of the MMPBSA (molecular mechanics Poisson-Bolzmann surface area) approach [19]. Relative binding free energies can be derived from MD trajectories using the theories of statistical thermodynamics. In this case, however, a simulation of each inhibitor for each enzyme would require over a hundred individual MD runs. In order to make the problem computationally tractable, we performed MD just once for each enzyme, using only the common portion of each inhibitor. Benzyl groups served as "dummy" side chains at the P2-P4 scaffold positions during these simulations and acted as placeholders in the enzyme pockets. Following the dynamics runs, full side chains at the P2 position were added in an incremental fashion and rank ordered according to the DOCK energy score [20]. The top 20 conformations of each side chain were then minimized in AMBER [21] and rescored using a PBSA solvation model [19]. Since the scaffold

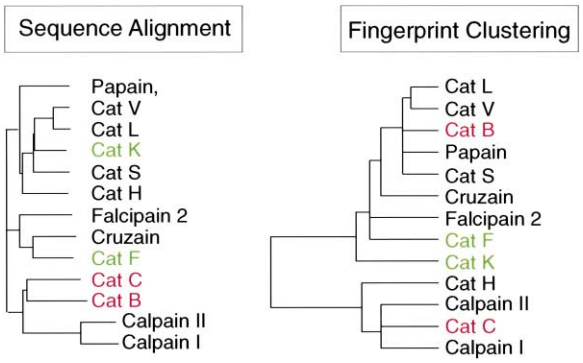

Figure 6. Comparison of Fingerprint Clustering and Sequence Alignment-Based Clustering

Hierarchical clustering of affinity fingerprints for the 12 reference cysteine proteases produced dendrogram trees that indicate the degree of functional similarity between enzymes as a function of the height of the lines connecting profiles. A dendrogram tree generated using affinity fingerprints was compared to a tree generated by primary sequence alignment using Clustal W, as described in the Experimental Procedures section. Examples of enzymes with divergent clustering based on sequence alignment but with similarities in affinity fingerprints are shown in green, while enzymes that show similar sequence alignment but dramatic differences in classification based on affinity fingerprinting are shown in red.

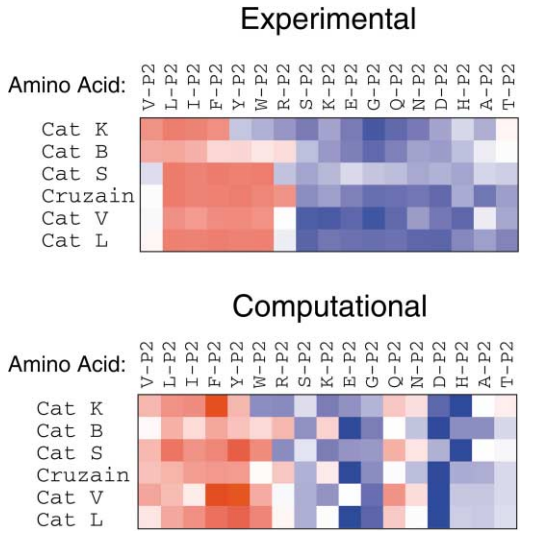

Figure 7. Comparison of In Silico Affinity Fingerprints with Experimental Fingerprints

The affinity-fingerprint inhibition data generated using a subset of the PSL P2 data were compared to data generated using a combination of molecular dynamics and DOCKing algorithms (see text). Computationally derived values for relative free energies were converted to color format similarly to experimentally obtained competition data. Cluster analysis highlights similarities between the two sets of data.

and enzyme conformational degrees of freedom were sampled during the dynamics runs, the resulting coordinates were preserved in subsequent steps. The side chain degrees of freedom were sampled using the less expensive incremental growth and energy minimization routines. Because we did not carry forth the thermodynamic ensemble of structures derived from the MD simulation, the results cannot be considered as time averaged free energies of binding. Although there is no physically rigorous way to isolate individual members of an MD ensemble for docking, we chose the member closest to a corresponding X-ray structure [22–27], which itself is part of a larger, physical ensemble.

The predictions derived from the six enzymes considered are in good qualitative agreement with the experimental data (Figure 7). Overall, the computational results accurately predict the general nature of favorable S2 sidechains for each enzyme. The computational results also agree with some of the fine discrimination seen between enzymes experimentally. Tryptophan, for example, is predicted to be a poor P2 sidechain for cathepsin K, and arginine is predicted to be poor for both cathepsin K and cathepsin S. These results demonstrate that qualitatively accurate results can be derived by DOCKing sidechains onto one member of an MD ensemble. It is reasonable to expect that individual predictions would improve as we averaged the docking results of more members of the scaffold-enzyme MD ensemble.

The largest differences between the in silico predictions and the experimental results are seen with the lysine, glutamine, and arginine residues (Figure 7). There are several differences between the conditions of the experiment and the assumptions of the models that could account for this. First, the experiment represents a measurement of relative residual enzymatic activity following treatment with each inhibitor sublibrary rather than a K_i. The calculations attempt to rank order the relative binding affinities of each P2 side chain. Second, the modeled inhibitors were constructed with alanine at the P3 and P4 sites, while the positional scanning libraries have equimolar mixtures of all amino acids at these sites. Third, the protonation states of the modeled acidic and basic residues were estimated based upon the experimental pH; the actual protonation states depend upon the local environments of each amino acid. Fourth, the inhibitor could adopt secondary structure in solution, thereby affecting its binding surface in a manner not considered during the simulations. Given these factors, it is reasonable that the theoretical predictions do not agree perfectly with the experimental results.

Ultimately, the computational protocol generated affinity fingerprints that can be used to predict most of the critical elements that control substrate specificity. Therefore, this method has the potential to be used to predict small molecule binding properties for other papain family proteases. Furthermore, the computational strategy allows for the screening of a virtual library of inhibitors to assist in the design of selective compounds for targets within a family of highly related enzymes.

Significance

In the post-genomic world, proteins are being conceptualized as members of families or networks, and this perspective should govern how all potential drug targets are analyzed. We have generated an affinity fingerprinting method to functionally characterize a family of cysteine proteases both chemically and com-

putationally. This method allows for the rapid visual analysis of inhibitor specificity and enzyme active site topology. Enzymes can then be subclassified based on functional relationships rather than simply by linear amino acid sequences. Furthermore, this method provides a direct readout of the overall inhibitory characteristics of compounds under a variety of assay conditions. This method will ultimately aid in the process of target selection, prioritization, and inhibitor design.

Experimental Procedures

Synthesis of Ethyl (2S,3S)-Oxirane-2,3-Dicarboxylate and Ethyl (2R,3R)-Oxirane-2,3-Dicarboxylate and DCG-04

The synthesis of (2R,3R)-oxirane-2,3-dicarboxylate is identical to that reported for the (2S,3S) isomer [28]. The synthesis of DCG-04 is reported elsewhere [10].

Synthesis of Positional Scanning Libraries

Synthesis of the PSL libraries was reported elsewhere [11]. Structures and corresponding number assignments for the 41 nonnatural amino acids used for the extended P2 library (Figure 4) are provided in the Supplemental Data.

Gel Electrophoresis

One-dimensional SDS-PAGE and two-dimensional IEF was performed as described [29].

Competition Labeling and Analysis of Data

Rat liver lysates (100 μg total protein in 100 μl buffer A: 50 mM Tris [pH 5.5], 5 mM $MgCl_2$, 2 mM DTT) or purified cathepsins (1 μg protein in 100 μl buffer A) were preincubated with 10 μM of each library member (diluted from 10 mM DMSO stocks) for 30 min at room temperature. Samples were then labeled by addition of ^{125}I-DCG-04 to each sample followed by further incubation at room temperature for 1 hr. Samples were quenched by the addition of 4$\times$ sample buffer, resolved by SDS-PAGE, and analyzed by PhosporImaging (Molecular Dynamics). Bands corresponding to each labeled protease were quantified. Intensities of inhibitor-treated samples were divided by the intensity of an untreated control sample to obtain a percent competition value. Numerical values for percent competition were analyzed as described previously [11, 15] using the programs Tree View and Cluster written by Eisen and coworkers [16]. These programs can be obtained from www.microarrays.org.

Cluster Analysis Based on Sequence Alignment

Amino acid sequences for all proteins were obtained from GenBank. All sequences used were human with the exception of falcipain 2 and cruzain. Sequence alignments were performed according to their primary structure similarity using the default settings for CLUSTAL W, version 1.5 (EMBL-EBI; www.ebi.ac.uk/clustalw/).

Computational Strategy

Initial geometries for the ligand-receptor structures were constructed by analogy to the E64-cathepsin K complex (Protein Data Bank ID code 1atk) [22, 30]. The structures of cathepsin L, V, B, K, S, and Cruzain were each aligned to the 1atk structure by matching the Cα atoms of four residues in the active site: Q19, C25, Y67, and W184 (papain numbering) [22–27, 30]. A minimal scaffold was then built into each receptor structure by analogy to the atomic coordinates of E-64. The resulting complexes were energy minimized for 500 steps using the AMBER program suite [21]. A 28 Å cap of TIP3P water [31] centered at the scaffold center-of-mass was added to each complex, and a subset of atoms 15 Å from the ligand as well as all water atoms were selected to be mobile during subsequent molecular dynamics simulations. Each complex was equilibrated to 300 K over 80 ps, and "snapshots" were then acquired every 4 ps over a 400 ps production MD run. From the resulting 100 scaffold-receptor poses, one was selected based upon minimum root-mean-square deviations from the original crystal structure and a minimum $C\alpha_{scaffold}$-$C\alpha_{E64}$ distance.

For each scaffold-receptor pose selected from the MD runs, side chains were incrementally grown away from the P2 scaffold position according to a previously reported methodology [18]. The resulting conformations of each side chain were then rank ordered by DOCK score [20], and the top twenty conformations of each added side chain on each scaffold-receptor pose were energy minimized using the AMBER program suite [21]. Cartesian restraints were applied to the scaffold and receptor atoms during the minimization. A 1 kcal/mol restraint was imposed upon the backbone atoms of the P2 residue, while a 500 kcal/mol restraint was imposed upon all other scaffold atoms and all receptor atoms. Only the P2 side chain atoms were allowed to move freely during 500 steps of minimization. Following minimization, each of the twenty conformations of each P2 side chain in each pose was rescored using a previously reported Poison-Bolzmann continuum solvation scheme [19]. Here, the free energy of binding is approximated by decomposition into molecular mechanical, solvation, and conformational entropy (ignored in this work) contributions.

Supplemental Data

The Supplemental Data contains the structures and number assignments of the 41 nonnatural amino acids used to generate the P2 diverse library. All compounds were obtained from commercial sources. Please write to chembiol@cell.com for a PDF.

Acknowledgments

We would like to thank Mark Rice and Paul Sprengeler (Celera, South San Francisco) for helpful discussion of the data and manuscript. We thank Dr. Vito Turk (Jozef Stefan Institute, Ljubljana, Slovenia) for the generous gift of purified recombinant human cathepsin L. This work was supported by funding from the Sandler Program in Basic Science (M.B., D.C.G., L.H., and K.C.) and the National Institutes of Health (GM31497 to I.D.K., GM64097 to W.D.A., GM56531 to P. Ortiz de Montellano, Principal Investigator, and CA72006 to M. Shuman, Principal Investigator).

Received: August 2, 2002
Revised: September 5, 2002
Accepted: September 6, 2002

References

1. Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O. (2000). Protein function in the post-genomic era. Nature *405*, 823–826.
2. Fetrow, J.S., and Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J. Mol. Biol. *281*, 949–968.
3. Hull, R.D., Singh, S.B., Nachbar, R.B., Sheridan, R.P., Kearsley, S.K., and Fluder, E.M. (2001). Latent semantic structure indexing (LaSSI) for defining chemical similarity. J. Med. Chem. *44*, 1177–1184.
4. Kick, E.K., Roe, D.C., Skillman, A.G., Liu, G., Ewing, T.J., Sun, Y., Kuntz, I.D., and Ellman, J.A. (1997). Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. Chem. Biol. *4*, 297–307.
5. Hopkins, S.C., Vale, R.D., and Kuntz, I.D. (2000). Inhibitors of kinesin activity from structure-based computer screening. Biochemistry *39*, 2805–2814.
6. Huo, S., Wang, J., Cieplak, P., Kollman, P.A., and Kuntz, I.D. (2002). Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: insight into structure-based ligand design. J. Med. Chem. *45*, 1412–1419.
7. Kauvar, L.M. (1995). Affinity fingerprinting. Biotechnology (N Y) *13*, 965–966.
8. Kauvar, L.M., Higgins, D.L., Villar, H.O., Sportsman, J.R., Engqvist-Goldstein, A., Bukar, R., Bauer, K.E., Dilley, H., and Rocke, D.M. (1995). Predicting ligand binding to proteins by affinity fingerprinting. Chem. Biol. *2*, 107–118.
9. Frye, S.V. (1999). Structure-activity relationship homology

(SARAH): a conceptual framework for drug discovery in the genomic era. Chem. Biol. *6*, R3–7.
10. Greenbaum, D., Medzihradszky, K.F., Burlingame, A., and Bogyo, M. (2000). Epoxide electrophiles as activity-dependent cysteine protease profiling and discovery tools. Chem. Biol. *7*, 569–581.
11. Greenbaum, D., Baruch, A., Hayrapetian, L., Darula, Z., Burlingame, A., Medzihradszky, K., and Bogyo, M. (2002). Chemical approaches for functionally probing the proteome. Mol. Cell Proteomics *1*, 60–68.
12. Barrett, A.J., Kembhavi, A.A., Brown, M.A., Kirschke, H., Knight, C.G., Tamai, M., and Hanada, K. (1982). L-trans-Epoxysuccinyl-leucylamido(4-guanidino)butane (E-64) and its analogues as inhibitors of cysteine proteinases including cathepsins B, H and L. Biochem. J. *201*, 189–198.
13. Turk, D., Guncar, G., Podobnik, M., and Turk, B. (1998). Revised definition of substrate binding sites of papain-like cysteine proteases. Biol. Chem. *379*, 137–147.
14. Meara, J.P., and Rich, D.H. (1996). Mechanistic studies on the inactivation of papain by epoxysuccinyl inhibitors. J. Med. Chem. *39*, 3357–3366.
15. Nazif, T., and Bogyo, M. (2001). Global analysis of proteasomal substrate specificity using positional-scanning libraries of covalent inhibitors. Proc. Natl. Acad. Sci. USA *98*, 2967–2972.
16. Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*, 14863–14868.
17. Schaschke, N., Assfalg-Machleidt, I., Machleidt, W., Turk, D., and Moroder, L. (1997). E-64 analogues as inhibitors of cathepsin B. On the role of the absolute configuration of the epoxysuccinyl group. Bioorg. Med. Chem. *5*, 1789–1797.
18. Lamb, M.L., Burdick, K.W., Toba, S., Young, M.M., Skillman, A.G., Zou, X., Arnold, J.R., and Kuntz, I.D. (2001). Design, docking, and evaluation of multiple libraries against multiple targets. Proteins *42*, 296–318.
19. Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., et al. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc. Chem. Res. *33*, 889–897.
20. Ewing, T.J., Makino, S., Skillman, A.G., and Kuntz, I.D. (2001). DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J. Comput. Aided Mol. Des. *15*, 411–428.
21. Pearlman, D.A., Case, D.A., Caldwell, J.W., Ross, W.S., Cheatham, T.E., III, DeBolt, S., Ferguson, D., Seibel, G., and Kollman, P. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energenic properties of molecules. Comp. Phys. Commun. *91*, 1–41.
22. Zhao, B., Janson, C.A., Amegadzie, B.Y., D'Alessio, K., Griffin, C., Hanning, C.R., Jones, C., Kurdyla, J., McQueney, M., Qiu, X., et al. (1997). Crystal structure of human osteoclast cathepsin K complex with E-64. Nat. Struct. Biol. *4*, 109–111.
23. Turk, D., Podobnik, M., Popovic, T., Katunuma, N., Bode, W., Huber, R., and Turk, V. (1995). Crystal structure of cathepsin B inhibited with CA030 at 2.0-A resolution: A basis for the design of specific epoxysuccinyl inhibitors. Biochemistry *34*, 4791–4797.
24. Somoza, J.R., Zhan, H., Bowman, K.K., Yu, L., Mortara, K.D., Palmer, J.T., Clark, J.M., and McGrath, M.E. (2000). Crystal structure of human cathepsin V. Biochemistry *39*, 12543–12551.
25. Guncar, G., Pungercic, G., Klemencic, I., Turk, V., and Turk, D. (1999). Crystal structure of MHC class II-associated p41 Ii fragment bound to cathepsin L reveals the structural basis for differentiation between cathepsins L and S. EMBO J. *18*, 793–803.
26. Brinen, L.S., Hansell, E., Cheng, J., Roush, W.R., McKerrow, J.H., and Fletterick, R.J. (2000). A target within the target: probing cruzain's P1' site to define structural determinants for the Chagas' disease protease. Struct. Fold. Des. *8*, 831–840.
27. McGrath, M.E., Palmer, J.T., Bromme, D., and Somoza, J.R. (1998). Crystal structure of human cathepsin S. Protein Sci. *7*, 1294–1302.
28. Bogyo, M., Verhelst, S., Bellingard-Dubouchaud, V., Toba, S., and Greenbaum, D. (2000). Selective targeting of lysosomal cysteine proteases with radiolabeled electrophilic substrate analogs. Chem. Biol. *7*, 27–38.
29. Bogyo, M., Shin, S., McMaster, J.S., and Ploegh, H.L. (1998). Substrate binding and sequence preference of the proteasome revealed by active-site-directed affinity probes. Chem. Biol. *5*, 307–320.
30. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. *28*, 235–242.
31. Jorgensen, W.L., Chandrasekhar, J., Madura, J., and Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. *79*, 926–935.