Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D

Ellen K Kick*[1], Diana C Roe*[2], A Geoffrey Skillman[2], Guangcheng Liu[1], Todd JA Ewing[2], Yaxiong Sun[2], Irwin D Kuntz[2] and Jonathan A Ellman[1]

Background: The identification of potent small molecule ligands to receptors and enzymes is one of the major goals of chemical and biological research. Two powerful new tools that can be used in these efforts are combinatorial chemistry and structure-based design. Here we address how to join these methods in a design protocol that produces libraries of compounds that are directed against specific macromolecular targets. The aspartyl class of proteases, which is involved in numerous biological processes, was chosen to demonstrate this effective procedure.

Results: Using cathepsin D, a prototypical aspartyl protease, a number of low nanomolar inhibitors were rapidly identified. Although cathepsin D is implicated in a number of therapeutically relevant processes, potent nonpeptide inhibitors have not been reported previously. The libraries, synthesized on solid support, displayed nonpeptide functionality about the (hydroxyethyl)amine isostere. The (hydroxyethyl)amine isostere, which targets the aspartyl protease class, is a stable mimetic of the tetrahedral intermediate of amide hydrolysis. Structure-based design, using the crystal structure of cathepsin D complexed with the peptide-based natural product pepstatin, was used to select the building blocks for the library synthesis. The library yielded a 'hit rate' of 6–7% at 1 μM inhibitor concentrations, with the most potent compound having a K_i value of 73 nM. More potent, nonpeptide inhibitors (K_i = 9–15 nM) of cathepsin D were rapidly identified by synthesizing and screening a small second generation library.

Conclusions: The success of these studies clearly demonstrates the power of coupling the complementary methods of combinatorial chemistry and structure-based design. We anticipate that the general approaches described here will be successful for other members of the aspartyl protease class and for many other enzyme classes.

Addresses: [1]Department of Chemistry, University of California, Berkeley, CA 94720-4160, USA and [2]Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143-0446, USA.

*Indicates an equal contribution to this work

Correspondence: Irwin D Kuntz and Jonathan A Ellman
E-mail: kuntz@cgl.ucsf.edu and jellman@uclink.berkeley.edu

Key words: aspartyl proteases, cathepsin D, combinatorial chemistry, structure-based design

Received: **12 February 1997**
Revisions requested: **4 March 1997**
Revisions received: **26 March 1997**
Accepted: **26 March 1997**

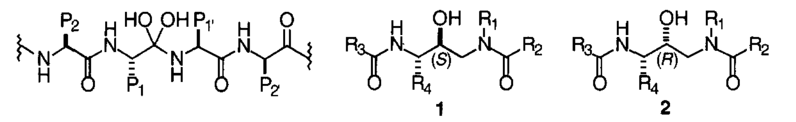
Chemistry & Biology April 1997, **4**:297–307
http://biomednet.com/elecref/10745521004 00297

Introduction

A cherished goal of chemists is to design and synthesize compounds with a specific set of properties. This goal is particularly urgent in biological and medicinal chemistry as a part of the drug-discovery process. Two powerful new tools that can be used in this effort are structure-based design [1,2] and combinatorial chemistry [3,4]. Structure-based design uses information gleaned from crystallographic and magnetic resonance experiments on a target macromolecule, frequently an enzyme, to guide the selection or design of inhibitors. Computation is important in this process [2,5]. Combinatorial chemistry is based on general chemical transformations that allow different building blocks to be combined in high yield. These transformations can be performed in parallel, to synthesize libraries of related compounds rapidly and efficiently [3,4]. Nonetheless, the discovery of a new lead compound or the improvement of the properties of an existing lead are still demanding tasks. Here, we integrate computational and combinatorial methods in a design protocol to produce libraries of compounds directed against specific macromolecular targets.

Combinatorial approaches to ligand identification initially focused on biopolymer libraries prepared by either chemical or biological methods [6]. For these libraries, all possible combinations of the building blocks are typically used because there are only four natural nucleotide building blocks for nucleic acid libraries and 20 proteinogenic amino-acid building blocks for peptide libraries. Both the structures of the compounds and the theoretical number of compounds in the library are determined by the length of the biopolymer chain. Recently, considerable efforts have been directed towards the preparation of libraries of compounds that encompass a wider spectrum of chemical transformations, to produce compounds with a broader range of

Figure 1



Intermediate of peptide hydrolysis

(Hydroxyethyl)amine-based inhibitors

Mechanism-based inhibitor design. The (hydroxyethyl)amine isostere is a stable mimetic of the tetrahedral intermediate of aspartyl protease catalyzed peptide hydrolysis.

properties than found in peptides or oligonucleotides [3,4]. These new approaches introduce significant challenges in library design.

A crucial element of any library design is the procedure for selecting which compounds to synthesize. This includes the choice of the scaffold, the basic reactions and the nature of the building blocks. If the building blocks are readily available components such as amines, aldehydes or carboxylic acids, the number of potential compounds to be considered can be quite large. For example, combining three building blocks with thousands of components at each position leads to over one billion compounds. Although different strategies have distinct practical limits, typically one is prepared to synthesize only thousands of spatially separate compounds and tens of millions of compounds in mixtures. Furthermore, evaluation and deconvolution of a very large library become rate-limiting activities [7]. Thus, there would be significant advantages to a method that reduces the synthetic effort to a small subset of compounds biased towards the desired properties.

How can the potential choices be efficiently reduced? Two standard strategies are diversity selection and directed selection. Diversity approaches attempt to maximize the sampling of chemical and biological properties in a fixed number of compounds [8]. In directed libraries the size and often the diversity of the library is reduced by selecting those building blocks that are predicted to have favorable interactions with the target, or by eliminating candidates that are believed to have unfavorable interactions. A directed library can be designed on the basis of substrate preferences, information about known inhibitors, or, in the work described here, an assessment of the potential interaction of specific functional groups with the target. Both diversity and directed strategies permit a multistage attack with secondary libraries generated from active compounds found in the first round.

The development of general and efficient approaches to identify small nonpeptidic inhibitors of proteases continues to be of interest because proteases have important roles in therapeutically relevant processes [9–12]. Proteases have also proven to be excellent targets for structure-based approaches [13,14]. Our target, cathepsin D, has been implicated in tumor metastasis in breast cancer, melanoma metastasis [15] and Alzheimer's disease [16,17]. Potent nonpeptide inhibitors of cathepsin D have not been reported previously [18]. Here we describe the efficient development of a combinatorial library with functionality that is selected using structure-based design. These studies resulted in the identification of potent inhibitors of cathepsin D, that do not contain amino acids and have molecular weights under 800 Da.

Results

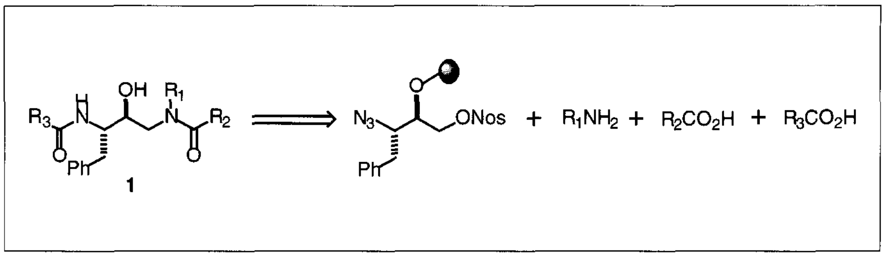
Specific approach

One powerful strategy to target an enzyme class is to incorporate a stable mimetic or isostere of the transition state or of an intermediate of the enzyme-catalyzed reaction [19]. The libraries for potential cathepsin D inhibitors are based upon the well-known (hydroxyethyl)amine isostere (Fig. 1). For our initial libraries, the P_1 sidechain (R_4) is the benzyl substituent, on the basis of the X-ray crystallographic structure of cathepsin D complexed with the natural peptide inhibitor pepstatin [20] and inhibition constants of peptide-based inhibitors [21,22].

In a pilot study, both *S* and *R* epimers at the hydroxyl carbon (Fig. 1; structures **1** and **2**) were prepared because both diastereomers have been found in potent inhibitors of other aspartic acid proteases [19]. Because inhibition of cathepsin D activity at 1 μM was only found with compounds of scaffold **1** in the pilot study, further syntheses of libraries used only scaffold **1**. Computer modeling (see below) predicted that structure **1** (Fig. 1) would provide the most potent inhibitors.

The solid-phase synthesis, which we reported previously [23], introduces diversity in three positions: a primary amine for the R_1 substituent and acylating agents for the R_2 and R_3 substituents (Fig. 2). The library synthesis was designed to use commercially available compounds for incorporation of the functionality at R_1, R_2, and R_3. We began our library design by selecting amines, carboxylic acids, sulfonyl chlorides and isocyanates with MW < 275 Da from the Available Chemical Directory

Figure 2



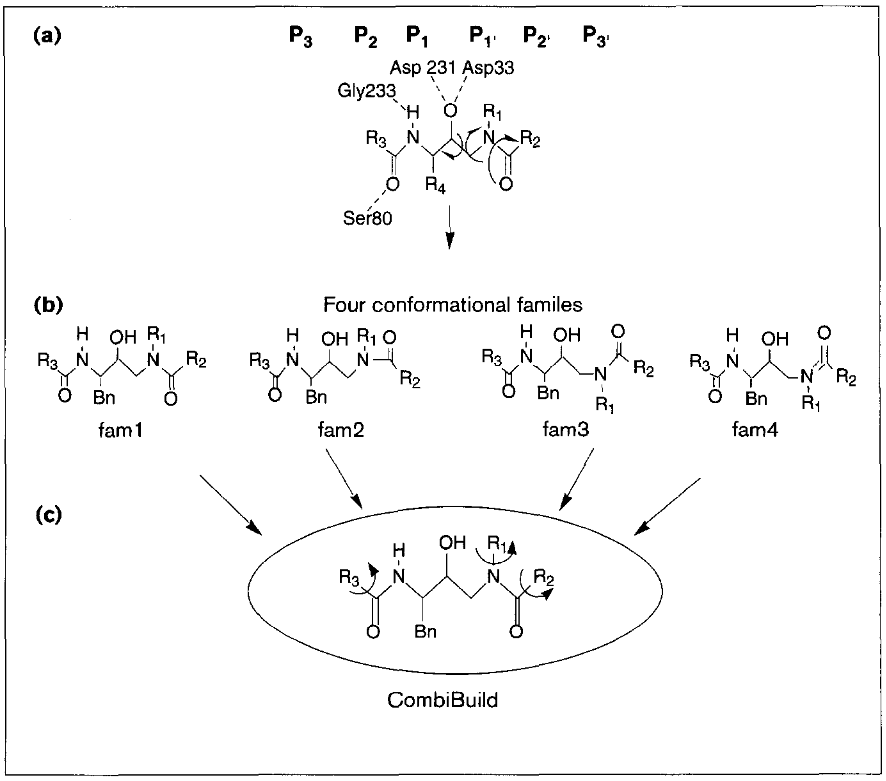
Components employed to prepare the libraries targeting cathepsin D. The same disconnections provide scaffold **2**. Isocyanates and sulfonyl chlorides, which can be used to incorporate R_2 and R_3, provide ureas and sulphonamides, respectively. For the synthesis, the scaffold is attached to polystyrene resin using a tetrahydropyran linker (indicated by the sphere).

(ACD, version 93.2; from MDL Information Systems, San Leandro, CA, USA). We eliminated compounds with functionality that was obviously incompatible with our synthesis. The resulting list included ~700 amines and ~1900 acylating agents, which would provide access to more than one billion compounds. To reduce the number of compounds in a sensible way, we turned to a computational screening process.

Directed library design

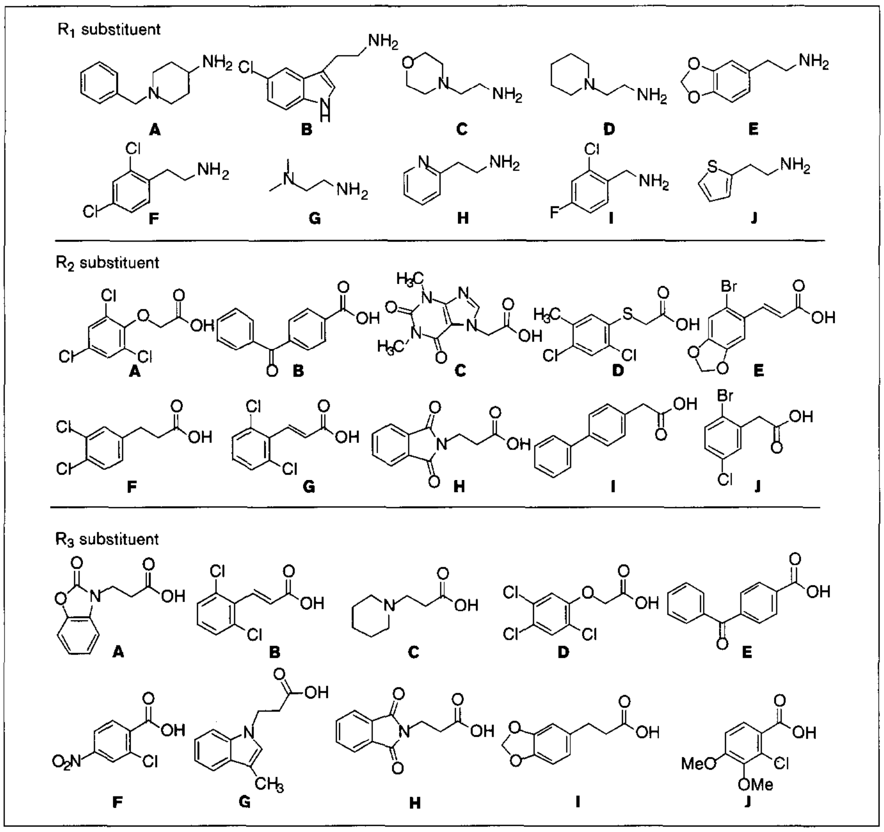
We chose a structure-based screening process using a new feature for our BUILDER molecular modeling program [24,25], called CombiBuild [26] (see Materials and methods section). To begin the design process, the scaffold was modeled in the active site with the assumption that the binding orientation of the scaffold would be similar to pepstatin (Fig. 3a). A conformational search of the scaffold identified a number of conformations with comparable energies. We clustered these conformations into four families on the basis of geometric similarity (Fig. 3b). CombiBuild [26] was used to position each of the R_1, R_2 and R_3 components onto the scaffold and to perform a full conformational search for the torsion angles of the substituent. In order to reduce the combinatoric problem, the R_1, R_2 and R_3 components were examined independently, but a probability-based clash grid was constructed to identify R_1 and R_2 conformations that might overlap. For example, if an R_1 conformation clashed with

Figure 3



Designing the combinatorial library with CombiBuild. **(a)** Modeling the scaffold. Coordinates and P_1–P_3 conformations of the pepstatin inhibitor were used as the starting geometry for the (hydroxyethyl)amine scaffold (see Materials and methods section). Methyl groups were placed at each of the scaffold's R_1–R_4 positions. **(b)** Scaffold conformation. A conformational search about the three torsion angles of the scaffold yielded four conformational families. A benzyl sidechain (Bn) was added to each of these families at the R_4 position. **(c)** Evaluating library components. The program CombiBuild performed a conformational search on all possible components at each variable position (R_1–R_3) on each family and scored the components by their potential interaction with cathepsin D. The top scoring candidates for each family were merged.

Figure 4



Directed library components. Each component is labeled with a letter code. EHA is defined as R_1=E, R_2=H and R_3=A.

more than 50% of the R_2 components, that conformation was discarded. Components from all of the conformational families were merged. Then, based on AMBER score, the fifty best components at each position were selected (see Materials and methods section). The 50 best-scoring components for all conformational families were merged for each of the three variable positions. Components with cost above $35 per gram were removed. The remaining compounds were hierarchically clustered to maximize the diversity of the top-ranking compounds that were selected for library synthesis. For R_1, R_2 and R_3, the ten best-scoring compounds from unique clusters were selected for each position.

Diverse library design

A diverse library, the same size as the directed library, was prepared as a control to provide a 'hit' rate when structure-based methods are not employed. The diverse library was designed to maximize the variety of functional groups and structural motifs of the library components. The sidechains for this library were selected by clustering the original list of components on the basis of their similarity to each other. Components were clustered with the Jarvis–Patrick algorithm [27] using the Daylight connectivity measure of similarity (Daylight Clustering Toolkit, v. 4.42; Daylight Chemical Information Systems, Inc., Santa Fe, NM, USA) and a binary Tanimoto metric [28,29] (see the Materials and methods section).

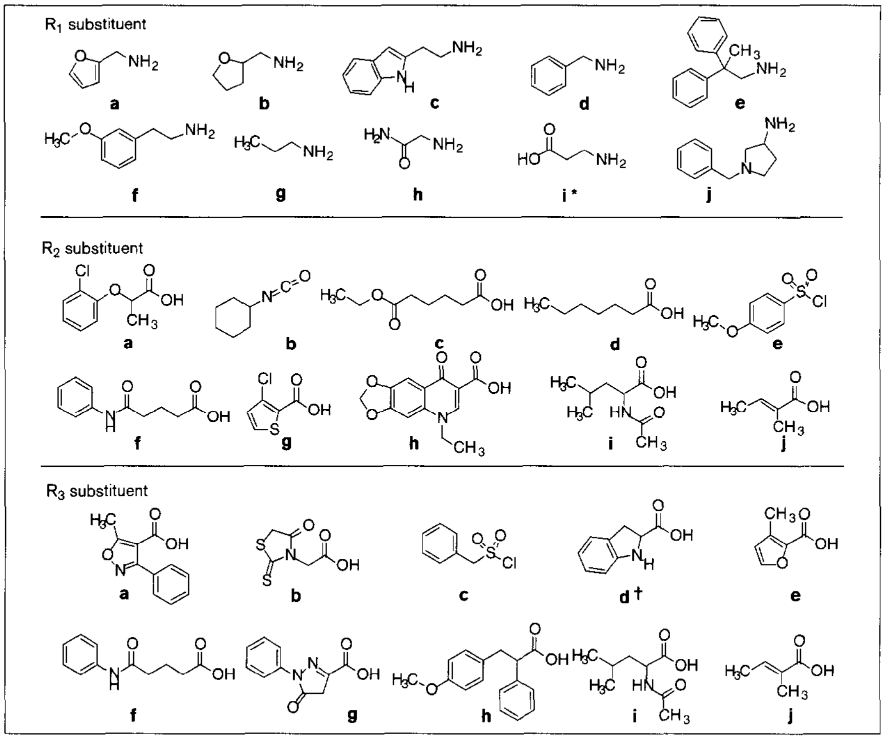
Library synthesis and screening

The directed and diverse libraries (1000 compounds each) were prepared using diastereomer **1** of the (hydroxyethyl)amine scaffold with the components used in library syntheses shown in Figures 4 and 5, respectively. Because the pilot study with *R* and *S* epimers only showed activity at 1 μM inhibitor concentration for the *S* epimers, only the *S* epimers of the directed and diverse library were synthesized. The libraries of spatially separate compounds were prepared by multiple parallel synthesis, and were screened in a high through-put fluorometric assay for inhibitory activity against cathepsin D [30].

Assay results

Using ~1 μM of crude compound, the directed library yielded 67 compounds that inhibited cathepsin D activity ≥50%. Further dilution to 333 nM and 100 nM inhibitor concentrations afforded 23 and 7 compounds, respectively, that inhibited cathepsin D activity ≥50% (Table 1). The data for the diverse library are also in Table 1. There are many uncertainties that can influence the results of a high

Figure 5



Diverse library components. Each component is labeled by a lower case letter code in a similar manner to Figure 4. *The *t*-butyl ester of R_1 = i was used in the coupling reaction. †The Boc protected amine of R_3 = d was used in the coupling reaction. These protecting groups are removed during the cleavage step.

through-put fluorescence assay, including the purity of each compound, the concentration of the compounds, and the experimental errors associated with the microtiter fluorescence assay. From repetitive experiments we estimate these errors to be ~30%, expressed as enzyme activity.

In order to obtain accurate inhibition constants (K_i), several compounds judged to be potent inhibitors on the basis of the library screening were synthesized on a larger scale, purified by chromatography, and characterized by nuclear magnetic resonance (NMR) and mass spectrometry. The K_i values were calculated from IC_{50} determinations (Table 2). From the compounds that were fully characterized, we obtained one compound from the directed library with a K_i below 100 nM, whereas the diverse library contained inhibitors that were three to four times less potent.

Second generation library

As a separate experiment, we explored a simple optimization strategy to identify compounds with improved binding affinity. We chose the directed library for this test. In the design of the directed library, we had selected derivatives by applying a clustering algorithm (see Directed library design section). We re-examined these clusters to expand upon the important structural elements of the most active compounds. In particular, we synthesized and screened a second generation library of 39 compounds from the clusters for the R_1, R_2 and R_3 positions that provided the most active compounds (Fig. 6). At 1 μM, 92% of the compounds that were screened inhibited cathepsin D

Table 1

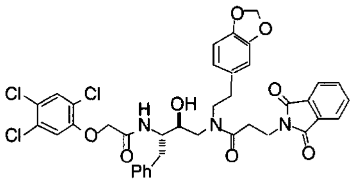
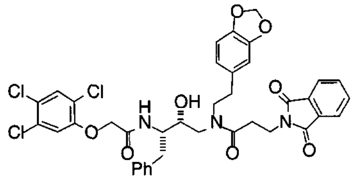
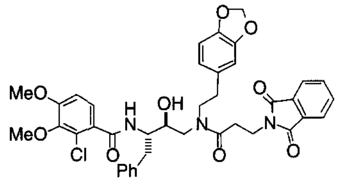
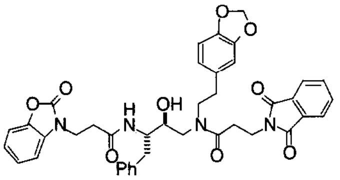
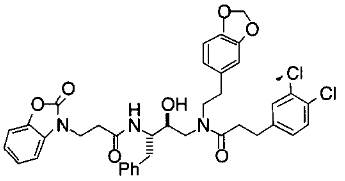
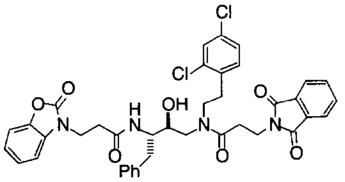
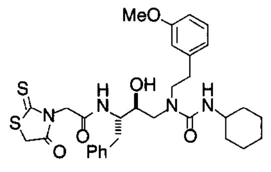
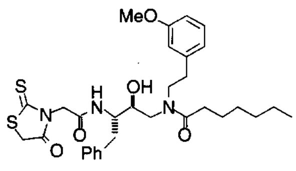
Number of compounds inhibiting cathepsin D.

	Library	
[Inhibitor]	Directed	Diverse*
100 nM	7†	1§
330 nM	23‡	3¶
1 μM	67	26
10 μM	11/95#	

†Compounds that inhibit ≥50% of cathepsin D activity at respective concentrations: EAA, EFA, EHA, EHD, EHI, EHJ, and FHA. An additional six compounds provided 40–50% inhibition of cathepsin D. ‡EAA, EFA, EHA, FAA, FFA, FHA, EHB, EFD, EHD, EEF, EHF, FHF, EFH, EHH, FAH, FFH, EFI, EHI, EAJ, EFJ, EGJ, EHJ, and FHJ. An additional 30 compounds provided 40–50% inhibition of cathepsin D. #100 compounds were selected by random number generation for testing at 10 μM. Five compounds were highly fluorescent at these concentrations, so that accurate assay data could not be obtained in these cases. §fbb. ¶fba, fbb, and fcb. Four compounds (fca, fdb, fib, and hhb) provided 40–50% inhibition of cathepsin D; with the experimental error in the assay, this activity is similar to the activity for the three that are listed. *The diverse library was not tested at 10 μM.

Table 2

Inhibition constants for selected compounds.

Inhibitor	Cmp code	K_i (nM)
	EHD	73 ± 9
	(*R*)- EHD	>5000
	EHJ	111 ± 8
	EHA	131 ± 12
	EFA	171 ± 25
	FHA	231 ± 31
	fbb	356 ± 31
	fdb	595 ± 66

Inhibition constants (K_i) were determined for several of the 'hits' from the designed and diverse libraries. The K_i values were determined from the IC_{50} values (see Materials and methods section).

≥50%, and 18% of the compounds at 100nM inhibited cathepsin D ≥50%. Inhibition constants were determined for selected compounds (Table 3), providing several potent inhibitors (K_i ≤15 nM) of cathepsin D.

Discussion

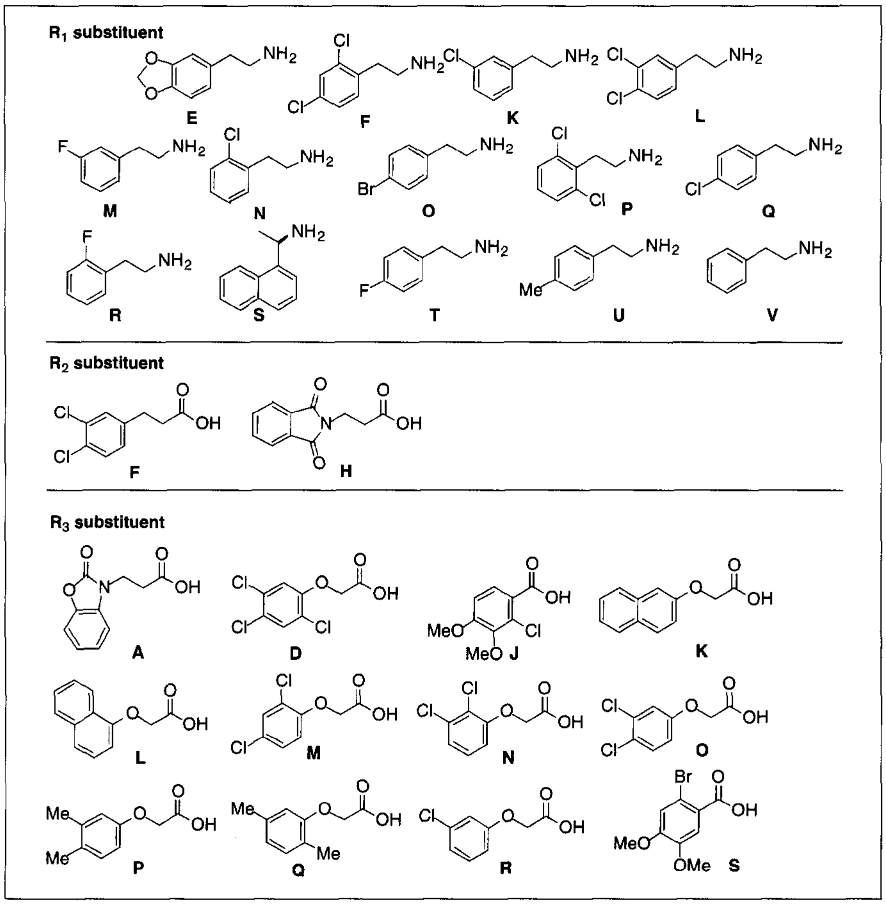
Novel low nanomolar inhibitors of cathepsin D were identified rapidly using combinatorial chemistry coupled with two different computational strategies. The diverse and directed libraries together, not including the optimization experiment, yielded over 90 compounds that were active at 1 μM and 26 compounds that were active in the submicromolar range. The 'hit rate' for activity at 1 μM is 6–7% for the directed library and 2–3% for the diverse library. When screening was performed at concentrations below 1 μM, there were seven times more 'hits' in the directed library than the diverse library. The most potent inhibitors from the directed library are threefold to fourfold better inhibitors than those in the diverse library. For the first round of synthesis and screening the number and potency of the active compounds were higher when the structural information of cathepsin D was used.

The experimental data collected in this project give a first picture of the distribution of activity within a library (Fig. 7). Although we have extended results only for the directed library, the pattern, as one might expect, shows a relatively steep reduction in the number of compounds that are active at low concentrations, roughly a factor of ten for each decade reduction in IC_{50}. Because the limiting slope for the diverse library is similar to that for the directed library, these results suggest that roughly ten times the number of compounds would have to be synthesized for the diverse library to yield the same results as the directed library. These data also provide a basis for testing different theoretical models of ligand binding.

A strength of the structure-based procedure is that it leads directly to testable geometrical hypotheses. In this study, there are three hypotheses. First, *S* epimers are predicted to bind better than the *R* epimers; second, there are two energetically reasonable scaffold conformations (Fig. 3; family 1+2, family 3+4), which place R groups into different pockets; third, all the inhibitors are assumed to bind in approximately the same orientation as pepstatin. The first hypothesis was directly tested in pilot experiments in which no inhibitors that were based upon the *R* epimer had activity at 1 μM. In addition, the *R* epimer of one of the most potent compounds had a K_i that was no better than 5 μM whereas the K_i of the *S* epimer was 15nM (Table 3). This conclusion and the inhibitor orientations in the cathepsin D complex will be examined by X-ray crystallography.

The computational approach outlined in this paper is most applicable when the scaffold orientation can be restricted

Figure 6



Components in each of the clusters (see Directed library design section) that contained the most active sidechains, R_1 = E, F; R_2 = F, H; R_3 = A, D, J. 39 compounds incorporating these sidechains were synthesized on resin as described previously: EFD, EHD, FFD, FHD, KFD, KHD, LFD, LHD, MFD, MHD, NFD, NHD, OFD, OHD, PFD, PHD, QFD, QHD, RFD, RHD, SFD, SHD, TFD, THD, UFD, UHD, VFD, VHD, EHA, EHJ, EHK, EHL, EHM, EHN, EHO, EHP, EHQ, EHR, and EHS. The compounds were assayed at 1 μM, 333 nM, 100 nM and 33 nM in high through-put screening. The most active compounds were synthesized on a large scale and the K_i values were determined (see Table 3).

using information from the structures of complexes. We are developing methods that will work even if the scaffold orientation is unknown or uncertain (Y.S., T.J.A.E., and I.D.K., unpublished observations). One of these methods docks scaffolds and sidechains simultaneously. Another method docks sidechains separately and then links them. With these methods we overcome the combinatorial explosion normally associated with generating all possible combinations in advance. Ultimately, the complexities of combinatorial chemistry may require different computational strategies for the design of the full range of oligomeric and small molecule libraries.

The work presented here is seen as the first stage of a process in which active compounds are identified and then, in later stages, the activity is optimized. The optimization criteria can include improved potency, selectivity, pharmacokinetic properties, or reduced toxicity. Each of these issues appears amenable to library design. For example, compounds with fivefold to sixfold improved potencies were rapidly identified by synthesizing and screening a small second generation library that explored variants of the most active compounds. This strategy is straightforward and can be readily applied to both directed and diverse libraries.

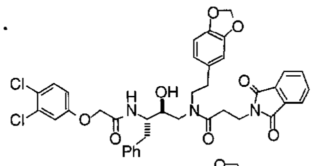
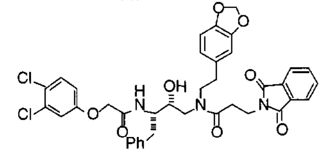
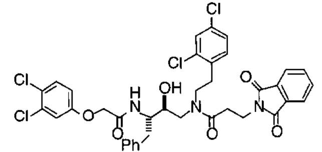
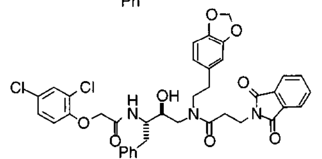
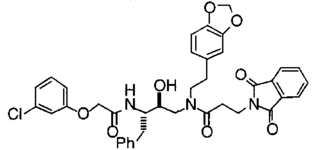
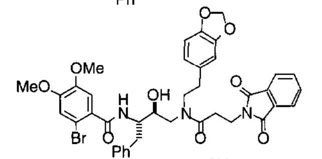
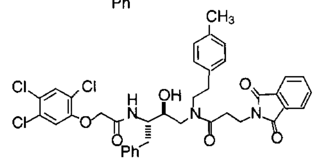
The success of the directed library towards finding potent inhibitors demonstrates the power of coupling combinatorial libraries with structure-based design. Combinatorial libraries allow a larger area of molecular space to be explored with the functionality selected by the structure-based design, removing the need to identify in advance a single 'best' target. Similarly, computational methods allow the rapid examination of extremely large virtual regimes (> 10^{10} compounds) and focus the chemical efforts into productive regimes. Diverse libraries remain an important strategy in the absence of target information and always offer the potential advantage of a wider range of lead candidates.

Significance

The identification of potent small molecule ligands for receptors and enzymes is one of the major goals of chemical and biological research. Two powerful new tools in these efforts are structure-based design and combinatorial chemistry. In the present work, we have integrated

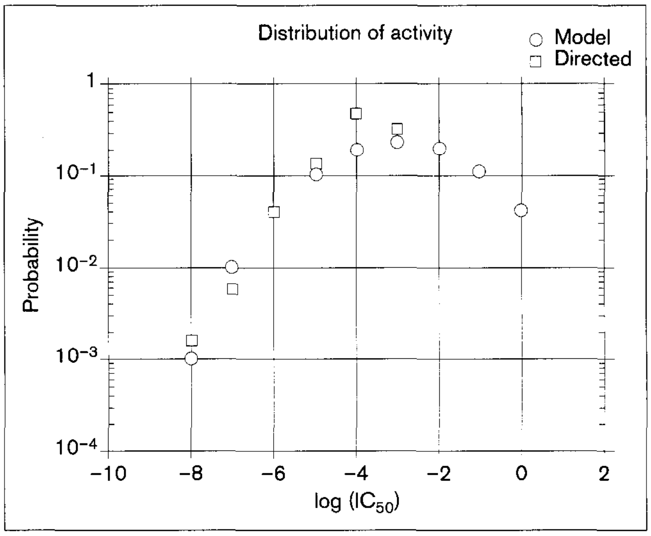
Table 3

Second generation library inhibition constants (K_i).

Inhibitor	Cmp code	IC_{50} (nM)	K_i (nM)
	EHO	19 ± 2	15
	(*R*)-EHO	>5000	
	FHO	18 ± 2	14
	EHM	14 ± 2	9
	EHR	20 ± 2	15
	EHS	64 ± 6	59
	UHD	229 ± 44	224

Inhibition constants (K_i) were determined for selected compounds from the second generation library. The K_i values were determined from the IC_{50} values shown (see Materials and methods section).

Figure 7



Normalized number of active compounds at each activity in directed and pilot libraries (squares) as a function of log (activity). The point at 10^{-3} is an upper limit. A very simple model is consistent with these data and illuminates some characteristics of the distribution of activity (circles). In the model, each molecule is assumed to be made of M_0 small units. If each unit has a probability, p, of yielding a favorable binding interaction of strength ε, and a 1 − p chance of an interaction strength 0, then the number of molecules, N(E), having an interaction energy of $E = M \times \varepsilon$, where M lies between 0 and M_0, is given by a binomial distribution. If p = 0.5, $N(E) = M_0! / (M! \times (M_0 - M)!)$. This distribution has a maximum affinity ($E = M_0 \times \varepsilon$), followed by a rapid increase in the number of molecules with diminishing activity. Interestingly, the distribution has a maximum probability and decreases as the binding affinity approaches 0. If M_0 is chosen as 10, $\varepsilon = 1.4\,\text{kcal mol}^{-1}$, N(E) is converted to a probability by normalization, and the model matched to the directed library data by an offset constant of $2.8\,\text{kcal mol}^{-1}$, the model tracks the experimental results fairly closely. The model as described has only three parameters that determine the shape of the curve. It is far too simple to be used in a quantitative manner. It may serve to stimulate further measurements and analysis, however.

these approaches to develop a general method for the rapid identification of potent enzyme inhibitors. We have demonstrated this method by identifying potent, nonpeptide inhibitors (K_i = 9–15 nM) of cathepsin D, a prototypical aspartyl protease that has been implicated in a number of therapeutically relevant processes, but for which potent nonpeptide inhibitors have not previously been reported [18].

The significance of this study is threefold. First, this general method should be directly applicable to the rapid identification of potent inhibitors of the other members of the aspartyl protease class. Second, with careful selection of the appropriate scaffold, we anticipate that the general approaches described here will also be successful for many other enzyme classes. Third, the success of this work clearly demonstrates the power of coupling the complementary approaches of combinatorial chemistry and structure-based design.

Material and methods

Directed library design

The structure-based design process began with coordinates for pepstatin in a complex with cathepsin D [20]. Using the standard nomenclature [31], the scaffold is identical to pepstatin on the P_1–P_3 side, but differs on the $P_{1'}$–$P_{3'}$ side and cannot form the same hydrogen bonds with the enzyme (Fig. 3a). Thus, we used the pepstatin positions for the P_1–P_3 side and systematically rotated the three scaffold torsion angles on the $P_{1'}$–$P_{3'}$ side. Each rotation was followed by energy minimization within the cathepsin D active site, using the AMBER [32] force field in the program Sybyl, a molecular modeling software package (Tripos

Associates, St Louis, MO, USA). During minimization, the enzyme was kept rigid but full flexibility of the scaffold was allowed. Both *S* and *R* epimers, structures **1** and **2**, were modeled using methyl groups for each of the R_1–R_4 groups. The conformational energies of the *R* epimers were generally ~2 kcal higher than for *S* epimers, leading us to predict that the *S* epimers would bind more tightly than the *R* epimers. All minimized conformations of *S* epimers within a 2 kcal mol^{-1} range were collected and clustered into four families based on geometric similarity (Fig. 3b). A benzyl group was added to each family at the R_4 position. The processed list of compounds from the ACD was passed through Sybyl to obtain Gasteiger and Marsili partial atomic charges for each component [33,34]. To reduce the computational time for searching the components, compounds with more than 4 torsional bonds were identified and removed. A new feature for our BUILDER molecular modeling program [24,25], called CombiBuild [26], was used to position each of the R_1, R_2 and R_3 components onto the scaffold and to perform a full conformational search for the torsion angles of the substituent at 15 degree increments. In order to reduce the combinatoric problem, the R_1, R_2 and R_3 components were examined independently, but a probability-based clash grid was constructed to identify R_1 and R_2 conformations that might overlap. For example, if an R_1 conformation clashed with more than 50% of the R_2 components, that conformation was discarded. Each rotation was then examined for intramolecular clashes with the scaffold and overlap with cathepsin D. Each accepted conformation was rigid-body minimized [35] and scored with a force-field grid [36]. The total computer time required to evaluate all torsion angles for all sidechains attached to four different scaffold conformations was 16 h on a Silicon Graphics Iris R4400. The 50 best scoring components for all families were merged for each of the three variable positions, and sorted by overall lowest score. Components with cost above $35 per gram were removed, leaving 34, 35 and 41 components at R_1, R_2 and R_3, respectively. Each remaining component was structurally fingerprinted using the Daylight software (Daylight Clustering Toolkit, v. 4.42, Daylight Chemical Information Systems, Inc., Santa Fe, NM, USA) and hierarchically clustered (similarity cutoff=0.63) [37] using the Tanimoto similarity metric [28,29]. For R_1, R_2 and R_3, the ten best-scoring components from unique clusters were selected for the directed library.

Diverse library design

Components from the original ACD list were clustered with the Jarvis–Patrick algorithm [27] using the Daylight connectivity measure of similarity and a binary Tanimoto metric [28,29]. In the Jarvis–Patrick method, two compounds are placed in the same cluster if they are neighbors of one another and share at least p neighbors from a list of q nearest neighbors, where p and q are adjustable parameters. The compound nearest the cluster centroid was chosen as the cluster representative.

The R_1 (amine) components were clustered directly as the primary amines. The R_2 and R_3 acylating agents were each attached to a portion of the scaffold before clustering to yield the proper chemical context at the linkage site. The first round of clustering yielded 47, 154 and 162 clusters using p/q = 4/11, p/q = 4/12, and p/q = 4/12 for R_1, R_2 and R_3, respectively. The representative R_2 and R_3 components were clustered a second time (p/q = 4/7 for R_2 and p/q = 4/7 for R_3), resulting in 23 R_2 and 35 R_3 components. We note that it is not practical to condense a large number of compounds into an arbitrarily small number of clusters because the cluster membership can become very diverse. Final selection of ten compounds from each list was based upon size, cost, availability and synthetic feasibility. Additionally, we sought a balance of functional groups for each set of sidechains. A comparison of the directed and diverse libraries (Figs 4,5) shows the much greater range of functionality spanned in the diverse library.

Library synthesis

We have previously reported the optimization of the solid-phase synthesis sequence to prepare the (hydroxyethyl)amine inhibitors and the demonstration of reaction generality [23]. Further testing was performed to establish that the different functionality to be displayed at R_1, R_2 and R_3 would be successfully incorporated into the potential inhibitors. First, all the amines and acylating agents to be incorporated in both the diverse and directed libraries were treated with trifluoroacetic acid for 2 h at room temperature to ensure stability to the support-cleavage conditions, which are by far the harshest reaction conditions in the synthesis sequence. Second, components that might pose difficulties on chemical or steric grounds were evaluated by incorporating them into syntheses of potential (hydroxyethyl)amine inhibitors on the basis of the synthetic protocol previously described [23]. These compounds were then purified by chromatography and analyzed by NMR and mass spectrometry. Five amines and four carboxylic acids that did not provide the expected final compound in high yields or purity were discarded. The following amines and acylating agents were successfully tested in the synthesis sequence: R_1 = B, C, E, F, a, e, h, i, j; R_2 = B, C, D, E, H, a, e, f; R_3 = A, D, E, H, a, b, e, g, i. The remaining components were assumed to be compatible with the synthesis sequence due to their similarity to sidechains that previously had been successfully incorporated.

The library synthesis was performed on polystyrene beads (20–40 mesh) prepared in our laboratory. The library was synthesized in a spatially separate array using a 96-well filter apparatus. Transfer of the resin to the individual wells was performed using an isopycnic mixture of *N,N*-dimethylformamide (DMF) and 1,2-dichloroethane. Incorporation of R_1 was carried out using 1.0 M amine in *N*-methylpyrrolidinone (NMP) at 80°C for 36 h. Incorporation of R_2 was carried out using stock solutions of 0.3 M carboxylic acid, 0.3 M benzotriazole-1-yl-oxy-tris-pyrrolidino-phosphonium hexafluorophosphate (PyBOP), 0.3 M 7-aza-1-hydroxybenzotriazole (HOAt), and 0.9 M *i*Pr_2EtN in NMP overnight. The coupling reactions were performed twice to ensure that complete coupling had occurred. After azide reduction with $SnCl_2$, PhSH and Et_3N, incorporation of R_3 was carried out as reported above for R_2. Carboxylic acid R_2 = E was coupled using *O*-(7-azabenzotriazol-1-yl)-1,1,3,3-tetramethyl-uronium hexa-fluorophosphate (HATU) instead of PyBOP due to formation of a precipitate under the standard coupling procedure. The isocyanate R_2 = b was coupled at 0.3 M in NMP overnight, and the sulfonyl chlorides R_2 = e and R_3 = c were coupled at 0.3 M in NMP that was 0.9 M in *i*Pr_2EtN. Cleavage of the material from support was achieved by subjecting the resin to trifluoroacetic acid : H_2O (95:5) for 30 min, followed by rinsing the resin and concentration of the filtrates using a Jouan 10.10 centrifugation concentrator. Toluene was added to form an azeotrope with trifluoroacetic acid during the concentration step. After concentration, the libraries were stored at −20°C.

Compounds from each library, picked by random number generation, were analyzed by mass spectrometry in a matrix of α-cyano-4-hydroxy-cinnamic acid on a Perseptive Biosystems MALDI instrument. For the diverse library the expected molecular ion peaks were observed for 46 of 49 compounds (poor ionization was obtained for the other three). Molecular ion peaks were obtained for 44 of 49 compounds from the directed library. In addition, the synthesis has been validated by the reasonable correlation of the approximate IC_{50} values of the crude material from the libraries with purified material that was synthesized on large scale for a number of compounds (see Table 2).

High through-put cathepsin D assay

A fluorometric high through-put assay for inhibitor activity towards human liver cathepsin D (Calbiochem) was performed in 96-well microtiter plates [30]. The peptide substrate (Ac–Glu–Glu(Edans)–Lys–Pro–Ile–Cys–Phe–Phe–Arg–Leu–Gly–Lys(Methyl Red)–Glu–NH_2) used in the assay has been previously reported ($K_m \approx 6\ \mu M$) [20]. The assay was performed in DYNATECH Microfluor fluorescence microtiter plates, and readings were taken on a Perkin-Elmer LS-50B with an attached 96-well plate reader. The excitation wavelength was 340 nm. A 340 nm interference filter (Hoya, U-340) for excitation and a 430 nm cut-off filter for emission were used. For the microtiter-based assays the substrate concentration was 5 μM and the cathepsin D concentration was 3–4 nM in a 0.1 M formic acid buffer (pH=3.7). Dimethylsulphoxide (DMSO; 10%) was used to ensure complete dissolution of the inhibitors. The fluorescent unit readings were taken at three time points within the linear region of the substrate cleavage, and percentage activity of the enzyme was determined by comparing the change of fluorescent units (FU) for each well to the average change in FU for six control wells without inhibitor. Each library was screened at approximately 1 μM inhibitor with the concentration based on the assumption that 50% of the theoretical yield was obtained for each inhibitor. All compounds that inhibited ≥50% cathepsin D activity were screened at subsequent threefold dilutions. All active compounds that inhibited ≥40% cathepsin D activity at 1 μM or lower inhibitor concentrations were assayed in duplicate.

Synthesis of inhibitors

Several of the most potent compounds from the designed and diverse libraries were synthesized on a 30 milligram scale on the solid-support following the previously reported method [23]. These compounds were purified by column chromatography, and characterized by ^{1}H NMR and either mass spectrometry or elemental analysis. The characterization data are listed below after the appropriate compound code. The ^{1}H NMR is only reported for the major amide rotomer for each compound.

EHA. ^{1}H NMR (300 MHz, $CDCl_3$) δ 2.43 (m, 2H), 2.55 (t, *J* = 7.2 Hz, 2H), 2.62–2.83 (m, 4H), 2.88 (dd, *J* = 2.3, 14.3 Hz, 1H), 3.27–3.36 (m, 2H), 3.48 (dd, *J* = 8.7, 14.3 Hz, 1H), 3.70 (d, *J* = 8.0 Hz, 1H), 3.85 (t, *J* = 7.2 Hz, 2H), 4.03–4.14 (m, 3H), 5.91 (s, 1H), 5.92 (s, 1H), 6.44 (dd, *J* = 1.6, 7.9 Hz, 1H), 6.52 (d, *J* = 1.6 Hz, 1H), 6.66 (d, *J* = 7.9 Hz, 1H), 7.03–7.26 (m, 9H), 7.70 (dd, *J* = 3.0, 5.4 Hz, 2H), 7.81 (dd, *J* = 3.0, 5.4 Hz, 2H). LRMS (MALDI-TOF) α-cyano-4-hydroxycinnamic acid matrix: mass calc'd for $C_{40}H_{39}N_4O_9$ (MH$^+$) 719.3, found 718.7. Anal. calc'd for $C_{40}H_{38}N_4O_9$: C, 66.84; H, 5.33; N, 7.79. Found: C, 66.67; H, 5.20; N, 7.97.

EHD. ^{1}H NMR (300 MHz, $CDCl_3$) δ 2.43–2.54 (m, 2H), 2.76 (t, *J* = 7.6 Hz, 2H), 2.92–2.98 (m, 2H), 3.02 (dd, *J* = 1.9, 14.2 Hz, 1H), 3.34 (m, 2 H), 3.68 (dd, *J* = 8.9, 14.2 Hz, 1H), 3.82–4.01 (m, 3H), 4.19 (apparent q, *J* = 8.0 Hz, 1H), 4.40 (d, *J* = 14.3 Hz, 1H), 4.51 (d, *J* = 14.3 Hz, 1H), 5.91 (d, *J* = 1.5 Hz, 1H), 5.92 (d, *J* = 1.5 Hz, 1H), 6.45 (dd, *J* = 1.6, 7.9 Hz, 1H), 6.53 (d, *J* = 1.6 Hz, 1H), 6.66 (d, *J* = 7.9 Hz, 1H), 6.95 (s, 1H), 7.17–7.28 (m, 5H), 7.49 (s, 1H), 7.71 (dd, *J* = 3.1, 5.5 Hz, 2H), 7.82 (dd, *J* = 3.1, 5.5 Hz, 2H). LRMS (MALDI-TOF) α-cyano-4-hydroxycinnamic acid matrix: mass calc'd for $C_{38}H_{35}N_3O_8Cl_3$ (MH$^+$) 766.1, found 767.1.

EHJ. ^{1}H NMR (300 MHz, $CDCl_3$) δ 2.37–2.56 (m, 2H), 2.63 (t, *J* = 6.8, 2H), 2.98–3.01 (m, 2H), 3.14 (d, *J* = 11.2, 1H), 3.41 (m, 2H), 3.82 (s, 3H), 3.86 (s, 3H), 3.76–3.90 (m, 4H), 4.33 (apparent q, *J* = 8.0 Hz, 1H), 5.87 (d, *J* = 1.3 Hz, 1H), 5.89 (d, *J* = 1.3 Hz, 1H), 6.45 (dd, *J* = 1.6, 7.9 Hz, 1H), 6.49 (d, *J* = 1.6 Hz, 1H), 6.63 (d, *J* = 7.9 Hz, 1H), 6.77 (d, *J* = 9.2 Hz, 1H), 6.78 (d, *J* = 8.7 Hz, 1H), 7.18 (d, *J* = 8.7 Hz, 1H), 7.24–7.28 (m, 5H), 7.69 (dd, *J* = 3.0, 5.4 Hz, 2H), 7.79 (dd, *J* = 3.0, 5.4). LRMS (MALDI-TOF) α-cyano-4-hydroxycinnamic acid matrix: mass calc'd for $C_{39}H_{39}N_3O_9Cl$ (MH$^+$) 728.2, found 727.9.

EFA. ^{1}H NMR (300 MHz, $CDCl_3$) δ 2.10–2.33 (m, 2H), 2.49 (t, *J* = 6.7 Hz, 2H), 2.47–2.86 (m, 7H), 3.17–3.33 (m, 2H), 3.58–3.69 (m, 2H), 4.02–4.14 (m, 3H), 5.90 (s, 2H), 6.38 (dd, *J* = 1.6, 7.9 Hz, 1H), 6.47 (d, *J* = 1.6 Hz, 1H), 6.67 (d, *J* = 7.9 Hz, 1H), 6.88 (dd, *J* = 2.0, 8.2 Hz, 1H), 7.02–7.26 (m, 10H), 7.29 (d, *J* = 8.2 Hz, 1H). LRMS (MALDI-TOF) α-cyano-4-hydroxycinnamic acid matrix: mass calc'd for $C_{38}H_{38}N_3O_7Cl_2$ (MH$^+$) 718.2, found 719.0.

FHA. ^{1}H NMR (300 MHz, $CDCl_3$) δ 2.46 (m, 9H), 2.94 (dd, *J* = 2.3, 14.3 Hz, 1H), 3.33 (apparent q, *J* = 7.0 Hz, 1H), 3.46 (dd, *J* = 8.8, 14.3 Hz, 1H), 3.73 (d, *J* = 8.2 Hz, 1H), 3.89 (t, *J* = 7.3 Hz, 2H), 4.02–4.16 (m, 3H), 6.43 (d, *J* = 9.2 Hz, 1H), 6.98–7.26 (m, 11H), 7.32 (d, *J* = 2.1 Hz, 1H), 7.70 (dd, *J* = 3.2, 5.4 Hz, 2H), 7.82 (dd, *J* = 3.2, 5.4 Hz, 2H). LRMS (MALDI-TOF) α-cyano-4-hydroxycinnamic acid matrix: mass calc'd for $C_{39}H_{37}N_4O_7Cl_2$ (MH$^+$) 743.2, found 743.0.

fbb. ^{1}H NMR (300 MHz, $CDCl_3$) δ 0.64–1.89 (m, 10H), 2.67 (t, *J* = 6.4 Hz, 2H), 2.79 (t, *J* = 6.7 Hz, 1H), 2.86–2.96 (m, 3H), 3.26 (t, *J* = 6.4 Hz, 2H), 3.43 (t, *J* = 6.8 Hz, 1H), 3.77–3.87 (m, 5H), 4.06 (d, *J* = 7.2 Hz, 1H), 4.09 (d, *J* = 7.2 Hz, 1H), 4.63 (s, 2H), 6.45 (d, *J* = 9.2 Hz, 1H), 6.64–6.82 (m, 3H), 7.18–7.31 (m, 6H). FABHRMS: 613.2519 (M$^+$ + H, $C_{31}H_{41}N_4O_5S_2$ requires 613.2518).

fdb. ^{1}H NMR (300 MHz, $CDCl_3$) δ 0.87 (t, *J* = 6.8 Hz, 4H), 1.20–1.26 (m, 8H), 1.47–1.58 (m, 2H), 1.95–2.14 (m, 2H), 2.67 (t, *J* = 7.2 Hz, 2H), 2.90–3.01 (m, 2H), 3.34–3.55 (m, 2H), 3.64–3.90 (m, 2H), 3.79 (s, 3H), 4.08 (m, 2H), 4.63 (s, 1H), 6.40 (d, *J* = 9.0 Hz, 1H), 6.62–6.78 (m, 3H), 7.18–7.31 (m, 6H). FABHRMS: 600.2555 (M$^+$ + H, $C_{31}H_{42}N_3O_5S_2$ requires 600.2566).

EHM. ^{1}H NMR (300 MHz, $CDCl_3$) δ 2.48 (m, 2H), 2.60 (m, 2H), 2.94 (m, 2H), 3.03 (dd, *J* = 1.7, 14.8 Hz, 1H), 3.37 (m, 2H), 3.68 (dd, *J* = 8.9, 14.2 Hz, 1H), 3.83–4.00 (m, 3H), 4.20 (apparent q, *J* = 8.2 Hz, 1H), 4.42 (d, *J* = 14.5 Hz, 1H), 4.51 (d, *J* = 14.5 Hz, 1H), 5.91 (s, 2H), 6.46 (dd, *J* = 1.6, 7.9 Hz, 1H), 6.53 (d, *J* = 1.6 Hz, 1H), 6.65 (d, *J* = 7.9 Hz, 1H), 6.76 (d, *J* = 8.8 Hz, 1H), 7.17 (dd, *J* = 2.6, 8.8 Hz, 1H), 7.20–7.26 (m, 5H), 7.40 (d, *J* = 2.6 Hz, 1H), 7.71 (dd, *J* = 3.1, 5.4 Hz, 2H), 7.83 (dd, *J* = 3.1, 5.4 Hz, 2H). LRMS (MALDI-TOF) α-cyano-4-hydroxycinnamic acid matrix: mass calc'd for $C_{38}H_{36}N_3O_8Cl_2$ (MH$^+$) 732.2, found 732.5.

EHO. ^{1}H NMR (400 MHz, $CDCl_3$) δ 2.47 (m, 2H), 2.62 (t, *J* = 7.1 Hz, 2H), 2.86–2.99 (m, 3H), 3.32–3.41 (m, 2H), 3.68 (dd, *J* = 8.8, 14.4 Hz, 1H), 3.79 (d, *J* = 8.6 Hz, 1H), 3.88–3.95 (m, 2H), 4.19 (apparent q, *J* = 8.1 Hz, 1H), 4.38 (d, *J* = 14.7 Hz, 1H), 4.47 (d, *J* = 14.7 Hz, 1H), 4.83 (s, 1H), 5.92 (d, *J* = 1.4 Hz, 1H), 5.93 (d, *J* = 1.4 Hz, 1H), 6.45 (dd, *J* = 1.5, 7.9 Hz, 1H), 6.53 (d, *J* = 1.5 Hz, 1H), 6/67 (d, *J* = 7.9 Hz, 1H), 6.77 (dd, *J*E = 2.9, 8.8 Hz, 1H), 6.91 (d, *J* = 9.5 Hz, 1H), 7.03 (d, *J* = 2.9 Hz, 1H), 7.20–7.28 (m, 5H), 7.35 (d, *J* = 8.8 Hz, 1H), 7.72 (dd, *J* = 3.0, 5.4 Hz, 2H), 7.83 (dd, *J* = 3.0, 5.4 Hz, 2H). FABHRMS: *m/e* 732.1879 (M$^+$ + H, $C_{38}H_{36}N_3O_8Cl_2$ requires 732.1879).

EHR. ^{1}H NMR (300 MHz, $CDCl_3$, CD_3OD) δ 2.40 (m, 2H), 2.58 (t, *J* = 7.0 Hz, 2H), 2.82 (m, 2H), 2.98 (dd, *J* = 3.4, 14.1 Hz, 1H), 3.31–3.46 (m, 3H), 3.74–3.88 (m, 3H), 4.13 (m, 1H), 4.32 (d, *J* = 14.8 Hz, 1H), 4.41 (d, *J* = 14.8 Hz, 1H), 5.84 (s, 2H), 6.41 (dd, *J* = 1.6, 7.9 Hz, 1H), 6.49 (d, *J* = 1.6 Hz, 1H), 6.60 (d, *J* = 7.9 Hz, 1H), 6.74 (dd, *J*= 1.9, 8.3 Hz, 1H), 6.89 (m, 1H), 6.94 (d, *J* = 8.0 Hz, 1H), 7.07–7.22 (m, 6H), 6.67 (dd, *J* = 3.1, 5.4 Hz, 2H), 7.78 (dd, *J* = 3.1, 5.4 Hz, 2H). LRMS (MALDI-TOF) α-cyano-4-hydroxycinnamic acid matrix: mass calc'd for $C_{38}H_{37}N_3O_8Cl$ (MH$^+$) 698.2, found 698.0.

EHS. ^{1}H NMR (300 MHz, $CDCl_3$) δ 2.49 (m, 2H), 2.65 (t, *J* = 7.1 Hz, 2H), 3.02 (m, 2H), 3.15 (d, *J* = 12.7 Hz, 1H), 3.41 (m, 2H), 3.75–3.99 (m, 4H), 4.36 (apparent q, *J* = 8.3 Hz, 1H), 5.90 (d, *J* = 1.5 Hz, 1H), 5.91 (d, *J* = 1.5 Hz, 1H), 6.47 (dd, *J* = 1.6, 7.9 Hz, 1H), 6.55 (d, *J* = 1.6 Hz, 1H), 6.66 (d, *J* = 7.9 Hz, 1H), 6.71 (d, *J* = 9.3 Hz, 1H), 6.90 (s, 1H), 6.98 (s, 1H), 7.17–7.28 (m, 5H), 7.70 (dd, *J* = 3.1, 5.6 Hz, 2H), 7.82 (dd, *J* = 3.1, 5.6 Hz, 2H). LRMS (MALDI-TOF) α-cyano-4-hydroxycinnamic acid matrix: mass calc'd for $C_{39}H_{39}N_3O_9Br$ (MH$^+$) 772.2, found 772.5.

FHO. ^{1}H NMR (400 MHz, $CDCl_3$) δ 2.65 (t, *J* = 7.2 Hz, 2H), 2.82 (t, *J* = 7.6 Hz, 2H), 2.91 (m, 2H), 3.07 (dd, *J* = 2.2, 14.3 Hz, 1H), 3.34–3.49 (m, 2H), 3.63 (dd, *J* = 8.9, 14.3 Hz, 1H), 3.83 (d, *J* = 8.4 Hz, 1H), 3.94 (m, 2H), 4.20 (apparent q, *J* = 8.0 Hz, 1H), 4.38 (d, *J* = 14.8 Hz, 1H), 4.46 (d, *J* = 14.8 Hz, 1H), 6.76 (dd, *J* 3.0, 8.9 Hz, 1H), 6.92 (d, *J* = 9.4 Hz, 1H), 7.00 (d, *J* = 8.2 Hz, 1H), 7.02 (d, *J* = 3.0 Hz, 1H), 7.15 (dd, *J* = 2.1, 8.2 Hz, 1H), 7.18–7.27 (m, 5H), 7.34 (d, *J* = 8.9 Hz, 1H), 7.73 (dd, *J* = 3.0, 5.4 Hz, 2H), 7.84 (dd, *J* = 3.0, 5.4 Hz, 2H). FABHRMS: *m/e* 756.1204 (M$^+$ + H, $C_{37}H_{34}N_3O_6Cl_4$ requires 756.1202).

UHD. ^{1}H NMR (300 MHz, $CDCl_3$) δ 2.29 (s, 3H), 2.38–2.57 (m, 2H), 2.68 (t, *J* = 7.2 Hz, 2H), 2.93 (m, 2H), 3.03 (dd, *J* = 1.6, 14.2 Hz, 1H), 3.34–3.48 (m, 2H), 3.71 (dd, J = 9.0, 14.2 Hz, 1H), 3.81–3.91 (m, 3H), 4.19 (apparent q, J = 8.1, 1H), 4.41 (d, J = 14.3, 1H), 4.50 (d, J = 14.3 Hz, 1H), 6.91–6.98 (m, 2H), 7.05 (d, J = 8.3 Hz, 2H), 7.17–7.28 (m, 5H), 7.50 (s, 1H), 7.67 (s, 1H), 7.71 (dd, J = 3.0, 5.5 Hz, 2H), 7.83 (dd, J = 3.0, 5.5 Hz, 2H). LRMS (MALDI-TOF) α-cyano-4-hydroxycinnamic acid matrix: mass calc'd for $C_{38}H_{37}N_3O_6Cl_3$ (MH^+) 736.2, found 736.3.

Cathepsin D assay

The cathepsin D assay for the compounds that had been fully characterized was performed in a quartz cuvette with a Perkin-Elmer LS-50B spectrometer. The assay was performed in 0.1 M formic acid buffer (pH = 3.7) with 2.5 μM peptide substrate and 10 nM cathepsin D. Inhibition constants (K_i) were determined from IC_{50} values taken from plots of V_i/V_o versus inhibitor concentration, where V_o is the velocity in absence of the inhibitor and V_i is the velocity with inhibitor. Because the substrate concentration is significantly below K_m, the IC_{50} values were converted to K_i by the equation $K_i \approx (IC_{50} - E_t/2)$, where E_t = enzyme concentration [38].

Acknowledgements

Support for this work was provided by the National Institutes of Health to I.D.K. and by the National Science Foundation, the A.P. Sloan Foundation and the Burroughs Wellcome Fund to J.A.E. We gratefully acknowledge predoctoral fellowship support from The American Chemical Society Division of Medicinal Chemistry and Parke-Davis Research to E.K.K., Pharmaceutical Training Grant, GM07075, to D.C.R. and A.G.S. and Pharmacia-Upjohn to G.L. We thank Jack Kirsch for advice on kinetic analysis and use of a fluorescence spectrophotometer, Barbara Chapman for advice and assistance in the fluorescence assay, Christina E. Lee and Connie Oshiro for helpful comments and discussions. Structure searches were performed using the ACD provided by MDL Information Systems, Inc., 14600 Catalina St., San Leandro, CA 94577, USA.

References

1. Kuntz, I.D. (1992). Structure-based strategies for drug design and discovery. *Science* **257**, 1078–1082.
2. Kuntz, I.D., Meng, E.C. & Shoichet, B.K. (1994). Structure-based molecular design. *Accts. Chem. Res.* **27**, 117–123.
3. Thompson, L.A. & Ellman, J.A. (1996). Synthesis and applications of small molecule libraries. *Chem. Rev.* **96**, 555–600.
4. Gordon, E.M., Barrett, R.W., Dower, W.J., Fodor, S.P.A. & Gallop, M.A. (1994). Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions. *J. Med. Chem.* **37**, 1385–1401.
5. Cohen, N.C., Blaney, J.M., Humblet, C., Gund, P. & Barry, D.C. (1990). Molecular modeling software and methods for medicinal chemistry. *J. Med. Chem.* **33**, 883–894.
6. Gallop, M.A., Barrett, R.W., Dower, W.J., Fodor, S.P.A. & Gordon, E.M. (1994). Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J. Med. Chem.* **37**, 1233–1251.
7. Terrett, N.K., *et al.*, & Steele, J. (1995). The combinatorial synthesis of a 30,752-compound library: discovery of SAR around the endothelin antagonist, FR-139,317. *Bioorg. Med. Chem. Lett.* **5**, 917–922.
8. Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K. & Moos, W.H. (1995). Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **38**, 1431–1436.
9. Takahashki, K. (1995). *Aspartic Proteinases Structure, Function, Biology, and Biomedical Implications.* Plenum Press, New York, USA.
10. Adams, J. & Stein, R. (1996). Novel inhibitors of the proteasome and their therapeutic use in inflammation. *Ann. Rep. Med. Chem.* **31**, 279–288.
11. Edmunds, J.J., Rapundalo, S.T. & Siddiqui, M.A. (1996). Thrombin and factor Xa inhibition. *Ann. Rep. Med. Chem.* **31**, 51–60.
12. Miller, D.K. (1996). Regulation of apoptosis by members of the ICE family and the Bcl-2 family. *Ann. Rep. Med. Chem.* **31**, 249–268.
13. Lam, P.Y.S., *et al.*, & Erickson-Viitanen, S. (1994). Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science* **263**, 380–384.
14. DeCamp, D., Ogden, R., Kuntz, I. & Craik, C. (1996). Site-Directed Drug Design. In *Site-Directed Drug Design*, (Cleland, J.L. & Craik, C.S., eds), pp. 467–505, Wiley-Liss, New York, USA.
15. Westley, B.R. & May, F.E.B. (1996). Cathepsin D and breast cancer. *Eur. J. Cancer* **32**, 15–24.
16. Ladror, U.S., Snyder, S.W., Wang, G.T., Holzman, T.F. & Krafft, G.A. (1994). Cleavage at the amino and carboxyl termini of Alzheimers amyloid-beta by cathepsin D. *J. Biol. Chem.* **269**, 18422–18428.
17. Cataldo, A.M., Hamilton, D.J., Barnett, J.L., Paskevich, P.A. & Nixon, R.A. (1996). Properties of the endosomal-lysosomal system in the human central nervous system: disturbances mark most neurons in populations at risk to degenerate in Alzheimer's disease. *J. Neurosci.* **16**, 186–199.
18. Whitesitt, C.A., *et al.*, & Panetta, J.A. (1996). Synthesis and structure-activity relationships of benzophenone as inhibitors of cathepsin D. *Bioorg. Med. Chem. Lett.* **6**, 2157–2162.
19. Wiley, R.A. & Rich, D.H. (1993). Peptidomimetics derived from natural products. *Med. Res. Rev.* **13**, 327–384.
20. Baldwin, E.T., *et al.*, & Erickson, J.W. (1993). Crystal structures of native and inhibited forms of human cathepsin D: implications for lysosomal targeting and drug design. *Proc. Natl Acad. Sci. U.S.A.* **90**, 6796–6800.
21. Jupp, R.A., *et al.*, & J. Kay (1990). The selectivity of statine-based inhibitors against various human aspartic proteinases. *Biochem. J.* **265**, 871–878.
22. Agarwal, N.S. & Rich, D.H. (1986). Inhibition of cathepsin D by substrate analogues containing statine and by analogues of pepstatin. *J. Med. Chem.* **29**, 2519–2524.
23. Kick, E.K. & Ellman, J.A. (1995). Expedient method for the solid-phase synthesis of aspartic acid protease inhibitors directed toward the generation of libraries. *J. Med. Chem.* **38**, 1427–1430.
24. Lewis, R.A., *et al.*, & Kuntz, I.D. (1992). Automated site-directed drug design using molecular lattices. *J. Mol. Graphics* **10**, 66–78.
25. Roe, D.C. & Kuntz, I.D. (1995). BUILDER v2: improving the chemistry of a *de novo* design strategy. *J. Comput. Aided Mol. Des.* **9**, 269–282.
26. Roe, D.C. (1995). *Application and Development of Tools for Structure-based Drug Design.* University of California, San Francisco, USA.
27. Jarvis, R.A. & Patrick, E.A. (1973). Clustering using a similarity measure based on shared near neighbors. *IEEE Comput.* **C22**, 1025–1034.
28. Willett, P., Winterman, V. & Bawden, D. (1986). Implementation of non-hierarchical cluster-analysis methods in chemical information systems: selection of compounds for biological testing and clustering of substructure search output. *J. Chem. Inf. Comput. Sci.* **26**, 109–118.
29. Willett, P. (1987). *Similarity and Clustering in Chemical Information Systems.* John Wiley & Sons, New York, USA.
30. Krafft, G.A. & Wang, G.T. (1994). Synthetic approaches to continuous assays of retroviral proteases. *Methods Enzymol.* **241**, 70–86.
31. Schecter, I. & Berger, A. (1968). On the size of the active site in proteases. I. Papain. *Biochem. Biophys. Res. Commun.* **27**, 157–162.
32. Weiner, S.J., *et al.*, & Weiner, P.A. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765–784.
33. Gasteiger, J. & Marsili, M. (1980). Iterative partial equalization of orbital electronegativity: rapid access to atomic charges. *Tetrahedron Lett.* **36**, 3219.
34. Gasteiger, J. & Marsili, M. (1981). Prediction of proton magnetic resonance shifts: the dependence on hydrogen charges obtained by iterative partial equalization of orbital electronegativity. *Organ. Magn. Reson.* **15**, 353–360.
35. Gschwend, D.A. & Kuntz, I.D. (1996). Orientational sampling and rigid-body minimization in molecular docking, revisited: on-the-fly optimization and degeneracy removal. *J. Comput-Aided Drug Des.* **10**, 123–132.
36. Meng, E.C., Shoichet, B.K. & Kuntz, I.D. (1992). Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **13**, 505–524.
37. Romesburg, H.C. (1984). *Cluster Analysis For Researchers.* Lifetime Learning Publications, Belmont, CA, USA.
38. Cha, S., Agarwal, R.P. & Parks, J. (1975). Tight-binding inhibitors-II: non-steady state nature of inhibition of mild xanthine oxidase by allopurinol and alloxanthine and of human erythrocytic adenosine deaminase by coformycin. *Biochem. Pharmacol.* **24**, 2187–2197.