A lead discovery strategy driven by a comprehensive analysis of proteases in the peptide substrate space

Sai Chetan K. Sukuru,¹ Florian Nigsch,¹ Jean Quancard,² Martin Renatus,³ Rajiv Chopra,⁴ Natasja Brooijmans,¹ Dmitri Mikhailov,¹ Zhan Deng,¹ Allen Cornett,¹ Jeremy L. Jenkins,¹ Ulrich Hommel,⁵ John W. Davies,¹ and Meir Glick^{1*}

¹Lead Discovery Informatics, Lead Finding Platform, Novartis Institutes for BioMedical Research, Inc., 250 Massachusetts Avenue, Cambridge, Massachusetts 02139

²Global Discovery Chemistry, Novartis Institutes for BioMedical Research, 4002 Basel, Switzerland

³Expertise Platform Proteases, Novartis Institutes for BioMedical Research, 4002 Basel, Switzerland

⁴Protein Structure Unit, Center for Proteomic Chemistry, Novartis Institutes for BioMedical Research, Inc., 250 Massachusetts Avenue, Cambridge, Massachusetts 02139

⁵Structural Biology Platform, Center for Proteomic Chemistry, Novartis Institutes for BioMedical Research, 4002 Basel, Switzerland

Received 15 June 2010; Revised 6 August 2010; Accepted 17 August 2010

DOI: 10.1002/pro.490

Published online 26 August 2010 proteinscience.org

Abstract: We present here a comprehensive analysis of proteases in the peptide substrate space and demonstrate its applicability for lead discovery. Aligned octapeptide substrates of 498 proteases taken from the MEROPS peptidase database were used for the *in silico* analysis. A multiple-category naïve Bayes model, trained on the two-dimensional chemical features of the substrates, was able to classify the substrates of 365 (73%) proteases and elucidate statistically significant chemical features for each of their specific substrate positions. The positional awareness of the method allows us to identify the most similar substrate positions between proteases. Our analysis reveals that proteases from different families, based on the traditional classification (aspartic, cysteine, serine, and metallo), could have substrates that differ at the cleavage site (P1–P1') but are similar away from it. Caspase-3 (cysteine protease) and granzyme B (serine protease) are previously known examples of cross-family neighbors identified by this method. To assess whether peptide substrate similarity between unrelated proteases could reliably translate into the discovery of low molecular weight synthetic inhibitors, a lead discovery strategy was tested on two other cross-family neighbors—namely cathepsin L2 and matrix metallo proteinase 9, and calpain 1 and pepsin A. For both these pairs, a naïve Bayes classifier model trained on inhibitors of one protease could successfully enrich those of its neighbor from a different family and vice versa, indicating that this approach could be prospectively applied to lead discovery for a novel protease target with no known synthetic inhibitors.

Keywords: chemical space; *in silico*; lead discovery; proteases; substrate space

Additional Supporting Information may be found in the online version of this article.

Sai Chetan K. Sukuru's current address is: Pfizer Global Research & Development, 445 Eastern Point Road, Groton, Connecticut 06340.

*Correspondence to: Meir Glick, Lead Discovery Informatics, Lead Finding Platform, Novartis Institutes for BioMedical Research, Inc., 250 Massachusetts Avenue, Cambridge, MA 02139. E-mail: meir.glick@novartis.com

Introduction

The underlying principle of chemogenomics is that similar ligands bind to similar targets.^{1,2} Predictive ligand-centric chemogenomic strategies for early stage discovery have been tailored specifically around a single target class, often protein kinases or G-protein-coupled receptors.³ Conceptually, such strategies

will result in an alternate classification by linking the members of a given target class entirely in the chemical space of their ligands, compared with the conventional phylogenetic classification based on target protein sequence and/or tertiary structure.⁴ A phylochemical classification like this leads to improved medicinal chemistry understanding of the target family with impact on identifying novel drug targets, lead discovery and understanding selectivity. However, linking the targets in the synthetic chemical space limits the scope of such a classification by excluding targets for which no ligands are known yet and leaving out unexplored chemical space. Proteases offer a unique opportunity to address this challenge as their natural ligands are peptide substrates representing a chemically diverse combination of amino acids.

The human genome encodes at least 553 proteases,⁵ many of which are major drug targets.^{6,7} Based on the mechanism of catalysis and sequence and structural similarities, proteases are classified into four major families, aspartic, cysteine, serine, and metalloproteases. The MEROPS peptidase database⁸ is a comprehensive resource on proteases, their substrates, and inhibitors. The database reports the peptide substrate sequence in accordance with the Schechter and Berger nomenclature⁹ and highlights four residues each on both the amino nonprime (P4–P1) and carboxyl prime (P1'–P4') side of the scissile bond. The positional nomenclature allows a simple but unambiguous alignment of all the octapeptide substrate sequences of all the proteases in the database. The vast information contained in the database enables an unbiased comprehensive analysis of proteases in the chemical space of their natural ligands, referred to as the peptide substrate space in this article.

We have shown earlier that multiple-category naïve Bayes models trained on chemogenomic databases can effectively predict the most-likely biological targets of compounds.¹⁰ Here, we train a multiple-category naïve Bayes model on two-dimensional (2D) chemical features of peptide substrates from the MEROPS database. For each protease, the model can identify statistically significant chemical features for each of the eight positions (P4'–P4) that a peptide should have for it to be a probable substrate. The hierarchy implemented in the analysis scheme (Fig. 1) enables the identification of similarities between residues of different peptide substrates of a protease at each of the eight positions. The Pearson correlation matrix derived from the model quantifies the relationship between the proteases in the peptide substrate space.

In this article, we focus on the novel nonintuitive, cross-family relationships revealed by our approach which could shed light on how distant proteases might be linked together by ligands designed by nature for them. Caspase 3, member of the C14

Figure 1. The workflow for this analysis. The hierarchy implemented in the multiple-category naïve Bayes model is briefly shown in the bottom panel.

family of cysteine proteases and granzyme B, member of the S1 family of serine proteases, will be discussed in detail as one of the previously known examples of cross-family proteases, identified as neighbors in the peptide substrate space. The simple, yet informative positional awareness of the model maps back the similarity between the protease neighbors to specific substrate positions. The presence of compounds inhibiting both these proteases with nanomolar potency indicates a probable link between substrate and inhibitor similarity between cross-family proteases. We explored this link by developing an *in silico* strategy to harness the substrate relatedness for lead discovery against novel proteases. We tested this strategy on two other unrelated pairs with no known shared inhibitors—cathepsin L2, a cysteine protease, and matrix metallo proteinase 9 (MMP-9), a metallo protease, and calpain 1, a cysteine protease, and pepsin A, an aspartic protease.

Results and Discussion

Relating proteases in the peptide substrate space

Within each of the four major families (Table I), our approach identified protease pairs that were highly correlated in the peptide substrate space too. Analysis of intrafamily protease pairs revealed that their strong correlation in the substrate space originates from similarities in the P1 and/or P1' positions

Table I. Distribution of Proteases in the Multiple-category Naïve Bayes Model

Family	Initial dataset	ROC score ^a ≥0.7	Percentage of proteases with ROC score ≥0.7
Aspartic	65	50	77
Cysteine	104	78	75
Metallo	138	87	63
Serine	181	145	80
Other ^b	10	5	50

^a The receiver operating characteristic (ROC) score gives an assessment of the model performance. A rating of 0.7 or higher in this case would imply that the model is able to correctly classify the substrates of a given protease in the model.

^b Other proteases include threonine, glutamic, and unknown.

flanking the scissile bond. This corroborates the traditional classification of proteases based on the mechanism of catalysis. The protease pair highlighted in the zoomed-in section of the tree (upper right panel)

in Figure 2 exemplifies the cross-family protease pairs identified by our approach. A viral cysteine protease has a bacterial metalloprotease as a neighbor in the peptide substrate space with a Pearson correlation coefficient of 0.50. The vaccinia virus I7L processing peptidase (Merops ID: C57.001) is a cysteine protease that cleaves major structural and membrane proteins of the virus.¹¹ Vaccinia virus is a member of the poxvirus family and is closely related to variola virus, the causative agent of small pox. In fact I7L shares 99% sequence identity with the K7L protease of variola major virus, making it a therapeutically attractive antiviral target.¹² The *Bacteroides fragilis* enterotoxin fragilysin (Merops ID: M10.020) is a zinc-dependent metalloprotease that primarily cleaves E-cadherin.¹³ As part of the intestinal microbial flora, *B. fragilis* secretes fragilysin and has been linked to secretory diarrhea in children and may even be associated with inflammatory bowel syndrome and colon cancer.¹⁴ This particular example brings out the

Figure 2. The Pearson correlation matrix computed from the multiple-category naïve Bayes model to produce a novel classification is rendered as a tree/dendrogram in a circular layout. A viral cysteine protease I7L (Merops ID: C57.001) and a bacterial metalloprotease fragilysin (M10.020; highlighted in the zoomed-in section of the tree in the upper right panel) were identified as one of the cross-family neighbors with a Pearson correlation coefficient of 0.50.

strengths of our approach to the fore: two proteases, each from a different organism, belonging to a different family based on the traditional classification, with no apparent overlap in their biological functions, are neighbors in the peptide substrate space. These protease neighbors will be discussed further in the next section.

Mapping the substrate similarity back on to specific peptide positions

Coupling the 2D chemical features with the positional nomenclature of the peptide substrates allows effective interpretation of substrate similarity. Figure 3 illustrates how we can identify specific peptide positions that are predominantly responsible for the substrate similarity between two proteases, which would not be obvious by their multiple sequence alignment alone. The most significant (positional rank: 1) chemical features identified for each of the eight substrate positions of the viral cysteine protease and the bacterial metalloprotease are shown in Figure 3A and B, respectively. Four out of the eight peptide substrate positions – P4, P3, P2, and P2' – were found to have significant common chemical features contributing to the correlation between substrates of the cysteine protease I7L and metalloprotease fragilysin [Fig. 3(C)]. A chemical feature was considered significant only if it is ranked in the top three in that particular position. The positional ranking was determined by the statistically assigned normalized probabilities (reported as scores in Fig. 3) in the multiple-category naïve Bayes model. For example, the aromatic chemical feature, derived from Phe and Tyr residues, was found to be the most significant one (positional rank: 1) for both proteases at the P4 position. Similarly, the long alkyl chain chemical feature, derived from Arg and Lys residues, was found to be the most significant one (positional rank: 1) for both proteases at the P2' position. The common chemical features in the P3 and P2 positions, however, are more significant in the metalloprotease than the cysteine protease. For example, the branched side chain chemical feature at P2 is the second most significant one (positional rank: 2) for the cysteine protease and is derived from an Ala residue in that position. However, the same chemical feature, derived from Ala, Ile, Leu, and Val residues, is the most significant one (positional rank: 1) in the metalloprotease, even though it scores lower than that of the cysteine protease. Interestingly, the two proteases do not have any significant common chemical feature in the P1 and P1' positions that form the cleavage site. These positions might be designed by nature to be dissimilar between the cross-family proteases to preserve their catalytic specificity. Although the proteases might differ in the highly specific P1 and P1' positions, other more similar positions, away from the

cleavage site, identified by our approach open up interesting medicinal chemistry opportunities to be explored. It is important to note that two proteases could still have a common chemical feature in a particular position but it may not be significant enough for either of the two proteases to contribute meaningfully to their substrate similarity.

Does cross-family peptide substrate similarity imply inhibitor similarity as well?

Although the substrates of cysteine protease I7L and metalloprotease fragilysin are similar, to our knowledge, there are no known inhibitors, peptidic or otherwise, that inhibit both. Given the importance of proteases in physiological processes,⁷ it is pertinent to think about their substrate and inhibitor similarity in an early drug discovery setting. In this context, caspases and the granzyme B homologs (human and rodent) are a unique example in our analysis. As shown in a zoomed-in branch of the tree in Figure 4A, caspases (highlighted C14 family cysteine proteases) and granzyme B homologs (highlighted S01 serine proteases) are neighbors in the peptide substrate space. Three interesting factors make them unique in our analysis. First, the caspases are highly correlated within the family and preserve their evolutionary phylogenetic relationship in the substrate space too. Among the five caspases included in the table, caspase 3 (Merops ID: C14.003) and caspase 7 (Merops ID: C14.004) were found to be most strongly correlated in the peptide substrate space, with a high Pearson correlation coefficient of 0.80 (Supporting Information Table S1). This observation is consistent with published studies that report these two caspases to be the key mediators of apoptosis¹⁵ with virtually indistinguishable substrates,¹⁶ even though they have distinct functional roles within apoptosis.¹⁷ Second, multiple members of the caspase family of cysteine proteases are related in the peptide substrate space to granzyme B [Merops ID: S01.010 (human) and S01.136 (rodent)], a member of the chymotrypsin family of serine proteases. Thus, granzyme B displays an extended cross-familial relationship with caspases. Their grouping in the peptide substrate space is consistent with their established functional relationships based on the essential roles they play in mammalian apoptosis.¹⁸ The most strongly correlated caspase neighbor to granzyme B was found to be caspase 6 (Merops ID: C14.005) with a Pearson correlation coefficient of 0.68 (Supporting Information Table S1). This is again supported by Thornberry *et al.*'s observation that similar nonprime tetrapeptide recognition motifs of granzyme B and caspases 6, 8 and 9 resemble the cleavage sites of precursors or proenzymes of caspases 3 and 7.¹⁸ Third, members of this cross-family classification actually share a potent peptide inhibitor, indicating that their strong

Figure 3. Illustration of mapping the substrate similarity between two proteases back on to their specific peptide substrate positions (P4–P4'). The most significant chemical features identified in each position of (A) I7L (Merops ID: C57.001) and (B) fragilysin (M10.020). For each of these chemical features, three corresponding parameters have been reported—positional rank (determined by the score), score (statistically determined normalized probability in the multiple-category naïve Bayes model), and amino acid (AA) residue(s) from which the chemical feature is derived. For example, the chemical feature at P2 is most significant (positional rank: 1) in fragilysin where it is derived from four residues with branched side chains—Ala, Ile, Leu, and Val as shown in (B). Based on our analysis, P4, P3, P2, and P2' were found to be the most similar substrate positions between the two proteases as shown in (C). In the context of our analysis, a chemical feature is considered significant in a particular position, if it is ranked in the top three. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 4. (A) The caspases (C14 cysteine protease family) and granzyme B (S01 serine protease family member) are interesting examples of cross-family protease neighbors shown as a highlighted section of the tree. The Pearson correlation coefficient between human granzyme B (Merops ID: S01.010) and caspase 3 (C14.003) is 0.67. Based on our analysis, P3, P1, P1', and P3' were found to be the most similar substrate positions between the two proteases as shown in (B). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

substrate similarity may translate into inhibitor similarity. The myxoma virus serpin SERP2 has been shown to inhibit (K_i values in brackets) three caspases, namely caspase 1 (Merops ID: C14.001; $K_i = 0.1$ nM), caspase 8 (Merops ID: C14.009; $K_i = 0.5$ nM), and caspase 10 (Merops ID: C14.011; $K_i = 1.3$ nM) and granzyme B (Merops ID: S01.010; $K_i = 0.6$ nM).¹⁹ In addition to this shared peptide inhibitor, four synthetic compounds have been shown to inhibit both caspase 3 and human granzyme B (Table II),²⁰ further supporting the case that cross-family peptide substrate similarity could imply inhibitor similarity too. We further analyzed the substrate similarity between caspase 3 and granzyme B. The significant common chemical features between the substrates of caspase 3 and human granzyme B, identified by our approach, are shown in Figure 4B. The most similar positions between the two proteases were P3, P1, P1',

and P3. The chemical feature representing the Asp side chain at P1 position was indeed the most significant common chemical feature that scores and ranks highly in both the proteases.

The synthetic compounds in Table II explore the structure–activity relationship (SAR) only in the nonprime (P4–P1) substrate positions. Based on our analysis, there were no significant common chemical features between these two proteases in the P4 and P2 nonprime positions. The chemical feature representing Ile was the second most significant feature in P4 while that representing Pro was identified to be the most significant feature in P2 position of human granzyme B [Fig. 5(A)]. Similarly, the chemical features representing Asp and Val residues were identified to be the most significant feature in the P4 and P2 positions, respectively, of caspase 3 [Fig. 5(B)]. These significant chemical features, with the

Table II. Compounds Inhibiting both Cross-family Neighbors (Fig. 4), Caspase 3 (C14.003) and Granzyme B (S01.010) [Color table can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Compound	R	Granzyme B (S01.010), K_i (nM) ^a	Caspase 3 (C14.003), K_i (nM) ^a
3	CHO	8	71
4		16	202
5		85	143
6		7	26

^a Inhibition data from Ref. 20.

exception of Ile being the second most significant after Val in P4 of granzyme B, agree with the optimal tetrapeptide substrate sequence identified using positional scanning synthetic combinatorial library approach for these two proteases.¹⁸ The corresponding canonical substrate-based inhibitor Ac-IEPD-CHO [compound 1, Fig. 5(C)] of granzyme B has a K_i of 80 nM.²⁰ Similarly, corresponding canonical substrate-based inhibitor Ac-DEVD-CHO [compound 2, Fig. 5(D)] of caspase 3 has a K_i of 1.2 nM.²¹ The common chemical feature representing the Glu side chain at P3 position, although present in other nanomolar inhibitors of granzyme B [e.g., compound 1; Fig. 5(C)] and caspase 3 [e.g., compound 2; Fig. 5(D)], was replaced by a rigid tricyclic system (compounds in Table II), spanning both P2 and P3, to gain potency by Willoughby *et al.*²⁰ As there were no known inhibitors exploring the SAR in the prime substrate positions (P1'–P4') of these proteases, we could not assess the influence of significant common chemical features identified by our approach in P1' and P3' positions [Fig. 4(B)]. While the presence of compounds (Table II) that inhibit both caspase 3 and human granzyme B supports their proximity in the

Figure 5. The significant chemical features identified for (A) the nonprime positions (P4–P1) of human granzyme B (Merops ID: S01.010) and (B) for the nonprime positions of caspase 3 (Merops ID: C14.003). A corresponding canonical substrate-based inhibitor Ac-IEPD-CHO (compound 1, shown in C) that inhibits granzyme B with a K_i of 80 nM²⁰ has all the four chemical features as shown in (A). Similarly, a corresponding canonical substrate-based inhibitor Ac-DEVD-CHO (compound 2, shown in D) that inhibits caspase 3 with a K_i of 1.2 nM²¹ has all the four chemical features shown in (B). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 6. A lead discovery strategy for a novel protease target, identified by our model to have a nonintuitive (typically cross-family) neighbor with known ligands. The objective is to come up with a list of compounds that could be potentially validated leads for the novel protease target.

peptide substrate space, it does not preempt the design/discovery of selective inhibitors of either protease. Rotonda *et al.* have shown that the significantly larger S1 subsite (that accommodates the P1 residue) in the human granzyme B structure, compared with the corresponding, Asp-specific, subsite in caspases, enables the design of specific inhibitors.²² It is important to note that our approach incorporates only the substrate information and not the sequence and/or structural information of the protease catalytic pocket. However, the significant chemical features identified by our approach in each substrate position implicitly capture the relevant interactions with the neighboring residues in the catalytic pocket. The examples discussed so far, encouraged us that the results of our analysis could potentially be applied for early lead discovery, if not specific inhibitors, for a novel protease target. Results of a strategy devised and tested toward achieving this objective will be discussed in the next section.

A lead discovery strategy for proteases

Traditional, substrate-inspired protease lead discovery often involves development of noncleavable isosteres of the residues occupying key substrate positions.^{23,24} Chemical modifications away from the cleavage site have been effective in some cases to tackle selectivity and to optimize the typically unattractive pharmaceutical properties of peptide inhibitors.^{25,26} A lead discovery strategy for proteases, based on the results of our analysis could augment the existing approaches through, for example, subset design for focused screening. The caspase–granzyme example (Fig. 5 and Table II) has shown that the same synthetic compound(s) can inhibit the cross-family protease neighbors in substrate space too. An

exemplary schematic workflow is shown in Figure 6 wherein a novel protease target with known substrates, but having few or no known inhibitors, has a (typically nonintuitive cross-family) protease neighbor, with known ligands, in the peptide substrate space. The first step is to retrieve all the low molecular weight (LMW) inhibitors, characterizing the synthetic chemical space, of the protease neighbor from a biologically annotated database, for example, GVK target-inhibitor database.²⁷ Only those compounds from the database reported to inhibit a protease with a half maximal inhibitory concentration (IC_{50}) or K_i of 10 μM or better (lower) are retrieved. To learn what makes these compounds active inhibitors of the protease neighbor compared with a large pool of diverse lead-like decoys, a naïve Bayes classifier is trained from their 2D chemical features. This classifier model could then be used to discover hits for the novel protease target from public and/or proprietary compound libraries. Such a strategy would lead to a more cost-effective, hypothesis-based screening of compounds for the novel protease target. Also, this approach enables the identification of novel compounds outside of the existing screening libraries that could be purchased prior to the screen. The discovery of tool compounds or initial hits will be a positive outcome of the strategy. Even if the strategy yields no hits, it will still help to understand the relative substrate and thus inhibitor specificity of the novel protease target. Naïve Bayes classifier models have been shown to perform well in enrichment studies involving extremely noisy datasets.^{28,29} Here, they are assessed not only for enriching the active inhibitors of the corresponding protease but also for enriching those of its neighbors in the peptide substrate space.

Table III. Testing the Protease Lead Discovery Strategy (Fig. 6) by Assessing the Cross-performance of Naïve Bayes Classifier Models Built for Protease Neighbors Identified by our Model but Having no Known Shared Inhibitors

Merops ID	Protease	Substrates	Gene ID	Number of synthetic inhibitors	Protease neighbor	Merops ID of neighbor	Pearson correlation coefficient	ROC score for enriching inhibitors using neighbor's naïve Bayes classifier model
C01.009	Cathepsin L2	3	1515	1089	Matrix metallo proteinase 9	M10.004	0.24	0.87
M10.004	Matrix metallo proteinase 9	239	4318	10,236	Cathepsin L2	C01.009		0.71
A01.001	Pepsin A	88	643834	144	Calpain 1	C02.001	0.22	0.93
C02.001	Calpain 1	90	823	1746	Pepsin A	A01.001		0.93

We decided to test this strategy on cross-family protease pairs identified by our approach. The protease pairs were selected based on the following criteria. First, each of the selected proteases should have 100 or more inhibitors in the GVK database with an IC_{50} or K_i of 10 μM or lower. It should be noted here that we could find valid inhibitors only for 66 proteases (out of the 365 for which the model performs well) in the database. A reasonable minimum threshold of 100 (unique) inhibitors ensures that the naïve Bayes classifier model trained for the protease is reliable enough for virtual screening. Second, the protease partners should not have any common inhibitor(s). The rationale behind these criteria was to rigorously test our strategy in an unbiased fashion. The strategy would be considered successful when the naïve Bayes classifier model trained on inhibitors of one protease robustly classifies, from a large pool of diverse decoys, the inhibitors of the other and vice versa. This is particularly relevant to a prospective application of this strategy in a real-life scenario when one of the partners in such a pair is a novel protease target, that is, it lacks known inhibitors. The results pertaining to two such pairs selected to test our strategy have been tabulated in Table III and will be discussed in detail.

The first pair was cathepsin L2 (Merops ID: C01.009), a cysteine protease, and MMP-9 (Merops ID: M10.004) with a Pearson correlation coefficient of 0.24. Cathepsin L2 has been shown to be widely expressed in colorectal and breast carcinomas.³⁰ MMP-9 (gelatinase B) has been shown to promote tumor invasion and angiogenesis.³¹ While cathepsin L2 has three substrates, MMP-9 has 239 substrates in the dataset. The gene IDs of these proteases were used to retrieve their inhibitors from the GVK database (Table III). A naïve Bayes classifier model trained on the cathepsin L2 inhibitors performed well to enrich the MMP-9 inhibitors with an receiver operating characteristic (ROC) score (see Materials and Methods section) of 0.71. Similarly, a naïve Bayes classifier model trained on the MMP-9 inhibitors performed even well to enrich the cathepsin L2

inhibitors with an ROC score 0.87 (Table III). The cross-performance of the naïve Bayes classifiers for a cross-family pair of protease neighbors having no known shared inhibitors supports the application of our approach for protease lead discovery. To verify that the peptide substrate similarity between these two proteases indeed facilitates the cross-performance of their corresponding naïve Bayes classifier models, additional tests were carried out. Coagulation factor IXa (Merops ID: S01.214) is a serine protease that was found to have a weak Pearson correlation of 0.07 with cathepsin L2 in the peptide substrate space. As expected, the naïve Bayes classifier model trained on cathepsin L2 inhibitors failed to enrich the inhibitors of factor IXa. Similarly, memapsin 2, also known as beta-site APP-cleaving enzyme 1 (Merops ID: A01.004), is an aspartic protease that was found to have a low Pearson correlation of 0.11 with MMP-9 in the peptide substrate space. Again, as expected, the naïve Bayes classifier model trained on MMP-9 inhibitors failed to enrich the inhibitors of memapsin 2. These results (summarized in Supporting Information Table S2) indicate that the peptide substrate similarity between cathepsin L2 and MMP-9 is important for lead discovery strategy to work. In other words, a naïve Bayes classifier model trained on inhibitors of a protease with very low or even negative Pearson correlation with the protease of interest may not be able to enrich or identify its inhibitors.

The second pair was pepsin A (Merops ID: A01.001), an aspartic protease, and calpain 1 (Merops ID: C02.001), a cysteine protease, with a Pearson correlation coefficient of 0.22. Pepsins are one of the oldest and most studied digestive enzymes and have been recognized as suitable markers for gastric cell development and gastric disease.³² Calpain 1 has been shown to be functionally required for processing of Alzheimer's beta-amyloid precursor protein.³³ Both pepsin A and calpain 1 have approximately the same number of substrates, 88 and 90, respectively, in the dataset. The gene IDs of these proteases were used to retrieve their inhibitors from

Table IV. Comparing the Peptide Substrate Similarity of Representative Cross-family Protease Neighbors on the Prime and Nonprime Side

Partner 1 (Merops ID)	Partner 2 (Merops ID)	Peptide substrate, Pearson correlation coefficient			Most similar substrate positions
		Full length	Nonprime	Prime	
Caspase 3 (C14.003)	Granzyme B, human (S01.010)	0.67	0.79	0.36	P3, P1, P1', P3'
Vaccinia virus I7L processing peptidase (C57.001)	Fragilysin (M10.020)	0.50	0.56	0.45	P4, P3, P2, P2'
Cathepsin L2 (C01.009)	Matrix metallo proteinase 9 (M10.004)	0.24	0.35	0.07	P3, P1, P1'
Pepsin A (A01.001)	Calpain 1 (C02.001)	0.22	0.34	0.11	P2, P1

the GVK database (Table III). A naïve Bayes classifier model trained on the pepsin A inhibitors performed robustly to enrich the calpain 1 inhibitors with an ROC score of 0.93. Similarly, a naïve Bayes classifier model trained on the calpain 1 inhibitors performed equally well to enrich the pepsin A inhibitors with an ROC score 0.93 (Table III). The cross-performance of the naïve Bayes classifiers for the second cross-family pair of protease neighbors further supports the application of our approach for protease lead discovery. This is even more significant on considering that the relatively low number of pepsin A inhibitors used to train the corresponding naïve Bayes classifier model was still able to enrich calpain 1 inhibitors. Additional tests were carried out to verify that the peptide substrate similarity between these two proteases is important for the cross-performance of their corresponding naïve Bayes classifier models. Coagulation factor Xa (Merops ID: S01.216) is a serine protease that was found to have a negative Pearson correlation of -0.09 with pepsin A in the peptide substrate space. As expected, the naïve Bayes classifier model trained on pepsin A inhibitors failed to enrich the inhibitors of factor Xa. Similarly, caspase 14 (Merops ID: C14.018) is a cysteine protease that was found to have a negative Pearson correlation of -0.07 with calpain 1 in the peptide substrate space. The naïve Bayes classifier model trained on calpain 1 inhibitors failed to enrich the inhibitors of caspase 14, as one would expect. These results (summarized in Supporting Information Table S2) indicate that the peptide substrate similarity between pepsin A and calpain 1 enables the strong cross-performance of their corresponding naïve Bayes classifier models.

Two recurring themes have emerged from testing our lead discovery strategy on the two-example cross-family protease pairs with good Pearson correlation in the peptide substrate space. First, the strategy seems to work well, that is, a naïve Bayes classifier model trained on inhibitors of one protease can robustly enrich the inhibitors of the other, even though they do not share any common inhibitors. These results show that the method is able to extrap-

olate from the chemical matter inhibiting protease A to different chemical matter that inhibits protease B. This indicates that our strategy could be prospectively applied to identify potential hits of a novel protease target. Secondly, the chemical moieties in the P1 and P1' positions of the inhibitors, although required for inhibition of the respective proteases, are not similar. Instead, the inhibitors of the cross-family protease neighbors, like their peptide substrates, are similar away from the cleavage site. This may give further insight into unexplored medicinal chemistry opportunities in rational design of protease inhibitors.

Prime and nonprime substrate similarity between proteases

We further analyzed the protease neighbors in our analysis to assess whether there is a dichotomy in their substrate similarities on the prime and nonprime side of the scissile bond. The Pearson correlation coefficient quantifying the substrate similarity for the full length as well as the nonprime and prime sides is tabulated in Table IV for all the example protease pairs discussed in the article. For all the cross-family protease pairs presented in the table, the nonprime side substrate similarity is statistically significantly higher than the prime side ($P < 2.2\text{e-}16$ in both Student's t -test and Mann-Whitney U test, see Materials and Methods section for details on calculation of statistical significance). This general trend is observed for all the protease pairs, both intrafamily and cross-family, in our dataset. Caspase 3 and granzyme B are the only proteases among our example pairs to have equal number of similar positions on the nonprime and prime side of the peptide substrate. However, all the protease inhibitors discussed earlier in this article interact mostly with the nonprime site residues of their respective proteases. Structural and biochemical studies have understood and exploited the differences between the prime and nonprime side of the protease binding site for inhibitor design. A large number of matrix metallo-protease inhibitors have been designed to bind the nonprime side of the active site, primarily to chelate

the Zn required for the enzyme activity.³⁴ Binding of inhibitor in the prime side of the hepatitis C virus NS3 serine protease binding site was shown to produce conformational changes resulting in stabilization of the corresponding complex.³⁵ Variations at the nonprime side of chimeric peptides for the dengue virus NS3 protease had significant effect on substrate binding (K_m values), whereas variations at the prime side had a higher impact on catalytic efficiency (k_{cat} values).³⁶ Our results indicate that proteases are perhaps more closely related in their peptide substrate space corresponding to the amino nonprime side compared with the carboxyl prime side. Further, a pattern of cooperativity between the subsites that has previously been reported³⁷ is also evident. The majority of peptide substrates bound to a protease adopt an extended beta-strand conformation,³⁸ enabling alternate subsites to interact. Both caspase 3 and cathepsin L2 data (Table IV) reflect the importance of P3–P1 crosstalk. A more detailed analysis would be required to investigate why such a dichotomy exists and whether it comprehensively extends to all the synthetic protease inhibitors as well. Results from such an analysis could further improve our understanding of protease specificity and refine the lead discovery strategy presented in this article.

Conclusions

We have presented here a novel *in silico* approach to analyze proteases in their peptide substrate space. This approach identifies rational links between seemingly unrelated proteases from phylogenetically different families. The 2D chemical features, coupled with the positional nomenclature incorporated in our approach, enabled us to map the substrate similarity back on to specific peptide positions. Cross-family protease neighbors were found to differ at the cleavage site but were similar away from it in the peptide substrate space. Broad trends suggest that the proteases are more closely related in the nonprime side of the peptide substrate space than the prime side.

This analysis of proteases in the peptide substrate space could have a significant impact on protease drug discovery. As shown in the lead discovery strategy presented here, it could be directly applied to select or design a focused set of compounds for screening in a hit finding effort against a novel protease target that may not have any solved 3D structures but has known substrates. First, the novel protease could be mapped in the analysis through its known peptide substrates. Peptide libraries designed to identify more substrates of the novel protease could provide additional data, which could increase the knowledge around the chemical space in each of its peptide substrate position. This in turn could increase the signal to noise ratio while training the multiple-category naïve Bayes model, resulting in the identification of its protease neighbors in the peptide

substrate space with higher Pearson correlation values. Known inhibitors of the protease neighbors could be retrieved from biologically annotated databases like GVK to train a naïve Bayes classifier model. This model could then be deployed to select or design screening subsets for the novel protease target.

In addition to its application in protease lead discovery, the comprehensive analysis of proteases can also guide lead optimization to improve selectivity against other proteases. Specific peptide substrate positions identified by our approach as not similar between the protease neighbors might offer opportunities for optimization. As part of the essential pre-clinical safety profiling process, pharmaceutical lead compounds are routinely assayed against a panel of other targets, so-called off-targets. While selecting targets on the safety panel is a subject of intensive research by itself, our approach could provide valuable information for protease inhibitor discovery projects. Including neighbors of the protease of interest, especially the nonintuitive ones, identified by our approach, in the safety panel could mitigate potential risks associated with the lead compound.

The protease lead discovery strategy presented here depends on LMW inhibitors retrieved from biologically annotated databases. These databases are a bottleneck as they cover relatively few (~20%) protease targets. The foremost challenge is to further streamline this strategy so as to hop from the peptide substrate space to the synthetic chemical space of inhibitors. The aim would be to circumvent the need for retrieving inhibitors from the database, ideally, by directly screening compound libraries using specific peptide position similarities identified by the approach. The biggest opportunity from the results of this approach is to bridge the medicinal chemistry gap created by synthetic efforts focused on few chemotypes targeting even fewer proteases. Applying the proposed lead discovery strategy for identifying inhibitors of novel protease targets would in turn expand the chemical space relevant to proteases.

Materials and Methods

Peptide substrates dataset

The peptide substrates were extracted from the MEROPS peptidase database, version 8.5, in September 2009.⁸ Only proteases that met the following criteria were included in the dataset: (a) substrates should be eight residues long, spanning P4–P4', (b) substrates should consist only of the 20 natural amino acids, and (c) at least two or more substrates that meet the first two criteria should be present for the protease. The rationale behind these criteria was to have a dataset that would enable analysis of proteases in the chemical space of their aligned, equal-length natural ligands. A total of 498 proteases from the database met these criteria and were included in the final dataset.

Multiple-category naïve Bayes model

The multiple-category naïve Bayes model was trained on 2D chemical features, namely extended connectivity fingerprints with a diameter of 6 (ECFP-6),³⁹ of the aligned peptide substrates of the proteases. The Laplacian-modified multiple-category naïve Bayes models have been explained in detail elsewhere.¹⁰ The “Learn Molecular Categories” component in Pipeline Pilot, version 7.5.2 (Accelrys Software), was used to build the multiple-category naïve Bayes model for the proteases in this study. ECFP-6 chemical features, annotated with the peptide position, were used to train the model. The objective of training the model was to generate a ligand-centric hypothesis for each protease which in turn would enable us to compare the key features among their substrates.

Assessing model performance

Before using the model to develop a comprehensive analysis of proteases in their peptide substrate space, we assessed its robustness by asking the following question. For how many proteases does the model correctly identify their corresponding substrates? In other words, for how many proteases can the model correctly distinguish the corresponding substrates from substrates belonging to other proteases? The ROC score is a common metric to measure model performance for binary classification,⁴⁰ and it was used here to assess whether our model was robust enough for further analysis. The ROC plot for each protease in the model is generated during the leave-one-out cross validation performed while building the model in Pipeline Pilot. Each substrate was left out one at a time, and a model built using the remaining substrates is used to predict the left-out substrate. Once all the substrates have predictions, an ROC plot is generated, and the area under the curve gives the ROC score. An ROC score of one for a protease in the model would imply perfect classification, that is, all its substrates were correctly distinguished from the rest of the substrates with no false positives or false negatives, while an ROC score of 0.5 implies a random model. A classification model with an ROC score of 0.7 or higher is considered fairly accurate according to rules implemented in Pipeline Pilot. The multiple-category naïve Bayes model performed well for 80% of the serine, 77% of the aspartic, 75% of the cysteine, and 63% of the metalloproteases (Table I). Overall, the model performed well for 365 (73%) of the 498 proteases in the dataset. The robustness of our multiple-category naïve Bayes model was further reinforced by the fact that the accuracy of performance was excellent (ROC score ≥ 0.9) for 302 (61%) and good (ROC score ≥ 0.8) for 341 (68%) proteases in the dataset. Serine proteases have the highest representation, approximately one-third in our dataset,

followed by metalloproteases, cysteine, and aspartic proteases. It was observed that the accuracy of the model did not depend on the number of substrates of a protease. The model performance across the major classes (Table I) gives us confidence that it could also be used to predict the most probable protease(s) for new peptide substrates.

Pearson correlation matrix

A Pearson correlation matrix was computed using the “Correlation Matrix” component in Pipeline Pilot to quantify the interrelatedness between the proteases in the model. The pairwise Pearson correlation is computed using the statistical weights placed on up to 10,000 chemical features (ECFP-6) by the Bayesian classifier for each protease model. The pairwise Pearson correlation coefficient between proteases provides a measure of similarity between the 2D chemical features derived from their corresponding peptide substrates. The correlation matrix, rendered as a circular tree/dendrogram in Figure 2, includes the 365 proteases for which the model performs well. Only 10% of the pairs had a Pearson correlation of 0.2 or higher, while 48% of the pairs had a Pearson correlation between 0 and 0.2. The entire list of Pearson correlation coefficients for each of the 365 proteases in our analysis has been shared with the readers through the Supporting Information Table S3 (a compressed csv file).

Visualizing the correlation matrix

To visualize the Pearson correlation matrix as a dendrogram, it was converted into a distance matrix using the KITSCH program in the PHYLIP package.⁴¹ The resulting distance matrix was then visualized as a tree using the PhyloWidget tool.⁴²

Calculation of statistical significance

Statistical tests were performed using the freely available software package R. The *z*-scores of the prime and nonprime correlation differences were tested for normality using a Kolmogorov–Smirnov (test statistic $D = 0.1156$) and a Darling–Anderson test (test statistic $A = 15.1$). In both cases, the null hypothesis that the data were sampled from a normal distribution had to be rejected ($P < 2.2\text{e-}16$).

Given the large sample size ($n = 66430$ pairwise combinations of 365 proteases), a paired Student's *t*-test can be applied to test if the difference in Pearson correlation resulting from the prime and nonprime sites is significantly different from zero. For small sample sizes or when the data (i.e., the differences of the prime and nonprime sites) is far from being normally distributed, the nonparametric paired Wilcoxon signed rank test (also referred to as Mann–Whitney *U* test) can be applied. The result was the same in both cases, the null hypothesis that the *z*-scores were randomly distributed around zero had

to be rejected (Student's t test: $t = 12.8825$, $P < 2.2 \times 10^{-16}$; Mann–Whitney U test: $V > 1e9$, $P < 2.2 \times 10^{-16}$).

Naïve Bayes classifier models for lead discovery

The low molecular inhibitors of proteases selected for testing the lead discovery strategy were retrieved from the GVK target-inhibitor database²⁷ using the protease gene IDs. Only inhibitors with an IC₅₀ or K_i of 10 μ M or lower were included in the final dataset. Laplacian-modified naïve Bayesian classifiers^{28,29} were built for each of the selected protease using the “Learn Good Molecules” component in Pipeline Pilot. ECFP-6 chemical features were again used to train these classifiers.

Acknowledgements

The authors thank colleagues at the Center for Proteomic Chemistry and Global Discovery Chemistry in Novartis Institutes for BioMedical Research (NIBR) for valuable feedback on the work presented here. S.C.K.S and F.N. thank the NIBR Education Office for supporting postdoctoral fellowships.

References

- Johnson MA, Maggiora GM (1990) Concepts and applications of molecular similarity. New York: Wiley.
- Jacoby E, Schuffenhauer A, Floersheim P (2003) Chemogenomics knowledge-based strategies in drug discovery. *Drug News Perspect* 16:93–102.
- Harris CJ, Stevens AP (2006) Chemogenomics: structuring the drug discovery process to gene families. *Drug Discov Today* 11:880–888.
- Vieth M, Higgs RE, Robertson DH, Shapiro M, Gragg EA, Hemmerle H (2004) Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim Biophys Acta* 1697:243–257.
- Puente XS, Sanchez LM, Overall CM, Lopez-Otin C (2003) Human and mouse proteases: a comparative genomic approach. *Nat Rev Genet* 4:544–558.
- Turk B (2006) Targeting proteases: successes, failures and future prospects. *Nat Rev Drug Discov* 5:785–799.
- Lopez-Otin C, Bond JS (2008) Proteases: multifunctional enzymes in life and disease. *J Biol Chem* 283:30433–30437.
- Rawlings ND, Morton FR, Kok CY, Kong J, Barrett AJ (2008) MEROPS: the peptidase database. *Nucleic Acids Res* 36:D320–D325.
- Schechter I, Berger A (1967) On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun* 27:157–162.
- Nidhi Glick M, Davies JW, Jenkins JL (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* 46:1124–1133.
- Ansarah-Sobrinho C, Moss B (2004) Role of the I7 protein in proteolytic processing of vaccinia virus membrane and core components. *J Virol* 78:6335–6343.
- Byrd CM, Bolken TC, Hruba DE (2003) Molecular dissection of the vaccinia virus I7L core protein proteinase. *J Virol* 77:11279–11283.
- Wu S, Lim KC, Huang J, Saidi RF, Sears CL (1998) *Bacteroides fragilis* enterotoxin cleaves the zonula adherens protein, E-cadherin. *Proc Natl Acad Sci USA* 95:14979–14984.
- Holton J (2008) Enterotoxigenic *Bacteroides fragilis*. *Curr Infect Dis Rep* 10:99–104.
- Lakhani SA, Masud A, Kuida K, Porter GA, Jr, Booth CJ, Mehal WZ, Inayat I, Flavell RA (2006) Caspases 3 and 7: key mediators of mitochondrial events of apoptosis. *Science* 311:847–851.
- Demon D, Van Damme P, Vanden Berghe T, Deceuninck A, Van Durme J, Verspurten J, Helsens K, Impens F, Wejda M, Schymkowitz J, Rousseau F, Madder A, Vandekerckhove J, Declercq W, Gevaert K, Vandenabeele P (2009) Proteome-wide substrate analysis indicates substrate exclusion as a mechanism to generate caspase-7 versus caspase-3 specificity. *Mol Cell Proteomics* 8:2700–2714.
- Walsh JG, Cullen SP, Sheridan C, Luthi AU, Gerner C, Martin SJ (2008) Executioner caspase-3 and caspase-7 are functionally distinct proteases. *Proc Natl Acad Sci USA* 105:12815–12819.
- Thornberry NA, Rano TA, Peterson EP, Rasper DM, Timkey T, Garcia-Calvo M, Houtzager VM, Nordstrom PA, Roy S, Vaillancourt JP, Chapman KT, Nicholson DW (1997) A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. *J Biol Chem* 272:17907–17911.
- MacNeill AL, Turner PC, Moyer RW (2006) Mutation of the Myxoma virus SERP2 P1-site to prevent proteinase inhibition causes apoptosis in cultured RK-13 cells and attenuates disease in rabbits, but mutation to alter specificity causes apoptosis without reducing virulence. *Virology* 356:12–22.
- Willoughby CA, Bull HG, Garcia-Calvo M, Jiang J, Chapman KT, Thornberry NA (2002) Discovery of potent, selective human granzyme B inhibitors that inhibit CTL mediated apoptosis. *Bioorg Med Chem Lett* 12:2197–2200.
- Fang B, Fu G, Agniswamy J, Harrison RW, Weber IT (2009) Caspase-3 binds diverse P4 residues in peptides as revealed by crystallography and structural modeling. *Apoptosis* 14:741–752.
- Rotonda J, Garcia-Calvo M, Bull HG, Geissler WM, McKeever BM, Willoughby CA, Thornberry NA, Becker JW (2001) The three-dimensional structure of human granzyme B compared to caspase-3, key mediators of cell death with cleavage specificity for aspartic acid in P1. *Chem Biol* 8:357–368.
- Leung D, Abbenante G, Fairlie DP (2000) Protease inhibitors: current status and future prospects. *J Med Chem* 43:305–341.
- Abbenante G, Fairlie DP (2005) Protease inhibitors in the clinic. *Med Chem* 1:71–104.
- Becker JW, Rotonda J, Soisson SM, Aspiotis R, Bayly C, Francoeur S, Gallant M, Garcia-Calvo M, Giroux A, Grimm E, Han Y, McKay D, Nicholson DW, Peterson E, Renaud J, Roy S, Thornberry N, Zamboni R (2004) Reducing the peptidyl features of caspase-3 inhibitors: a structural analysis. *J Med Chem* 47:2466–2474.
- Komoriya S, Kobayashi S, Osanai K, Yoshino T, Nagata T, Haginoya N, Nakamoto Y, Mochizuki A, Nagahara T, Suzuki M, Shimada T, Watanabe K, Isobe Y, Furugoori T (2006) Design, synthesis, and biological activity of novel factor Xa inhibitors: improving metabolic stability by S1 and S4 ligand modification. *Bioorg Med Chem* 14:1309–1330.
- Jagarlapudi SA, Kishan KV (2009) Database systems for knowledge-based discovery. *Methods Mol Biol* 575:159–172.

28. Glick M, Klon AE, Acklin P, Davies JW (2004) Enrichment of extremely noisy high-throughput screening data using a naive Bayes classifier. *J Biomol Screen* 9:32–36.
29. Xia X, Maliski EG, Gallant P, Rogers D (2004) Classification of kinase inhibitors using a Bayesian model. *J Med Chem* 47:4463–4470.
30. Santamaria I, Velasco G, Cazorla M, Fueyo A, Campo E, Lopez-Otin C (1998) Cathepsin L2, a novel human cysteine proteinase produced by breast and colorectal carcinomas. *Cancer Res* 58:1624–1630.
31. Yu Q, Stamenkovic I (2000) Cell surface-localized matrix metalloproteinase-9 proteolytically activates TGF-beta and promotes tumor invasion and angiogenesis. *Genes Dev* 14:163–176.
32. Kageyama T (2002) Pepsinogens, progastricsins, and prochymosins: structure, function, evolution, and development. *Cell Mol Life Sci* 59:288–306.
33. Scheiber J, Chen B, Milik M, Sukuru SC, Bender A, Mikhailov D, Whitebread S, Hamon J, Azzaoui K, Urban L, Glick M, Davies JW, Jenkins JL (2009) Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *J Chem Inf Model* 49:308–317.
34. Rao BG (2005) Recent developments in the design of specific matrix metalloproteinase inhibitors aided by structural and computational studies. *Curr Pharm Des* 11:295–322.
35. Casbarra A, Piaz FD, Ingallinella P, Orru S, Pucci P, Pessi A, Bianchi E (2002) The effect of prime-site occupancy on the hepatitis C virus NS3 protease structure. *Protein Sci* 11:2102–2112.
36. Niyomrattanakit P, Yahorava S, Mutule I, Mutulis F, Petrovska R, Prusis P, Katzenmeier G, Wikberg JE (2006) Probing the substrate specificity of the dengue virus type 2 NS3 serine protease by using internally quenched fluorescent peptides. *Biochem J* 397:203–211.
37. Ng NM, Pike RN, Boyd SE (2009) Subsite cooperativity in protease specificity. *Biol Chem* 390:401–407.
38. Tyndall JD, Nall T, Fairlie DP (2005) Proteases universally recognize beta strands in their active sites. *Chem Rev* 105:973–999.
39. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50:742–754.
40. Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO (2005) Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J Med Chem* 48:2534–2547.
41. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.68. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
42. Jordan GE, Piel WH (2008) PhyloWidget: web-based visualizations for the tree of life. *Bioinformatics* 24: 1641–1642.