Substrate Sequences Tell Similar Stories as Binding Cavities: Commentary

Similarities in binding cavities attract attention for the prediction and doptimization of ligand selectivity. Glinca and Klebe propose a clustering based on physicochemical properties of the binding site analyzed with Cavbase and conclude that their novel cavity-based method tells more than sequences.[5] We agree that protein structures are key in understanding of ligand recognition. Still, we think that sequences can tell a lot, if the focus is shifted away from protein sequences toward substrate sequences. We show that an analysis of protease substrates, inherently containing valuable information about binding site characteristics, can be directly utilized to predict potential off-targets.

Selectivity is a central issue in drug design, as drugs frequently hit more than a single target.[1] Therefore, molecular modeling aims at the prediction of polypharmacology with different approaches followed. Applied methods include ligand-based and structure-based methods as well as network analyses.[2−4]

Glinca and Klebe demonstrated recently that similarities in physicochemical characteristics of the binding cavity directly relate to overlapping substrate readout.[5] By application to protease test sets they show that their cavity-based approach yields similar results as analysis of ligand data from ChEMBL,[6] thereby outperforming a similarity analysis of protease sequences. Hence, they conclude, that "cavities tell more than sequences".

We definitely agree that structural information on the binding site is crucial in the rationalization of substrate recognition. Still, we think that sequence information can contribute significantly to an understanding of substrate specificity, when the focus is shifted from protease sequences toward substrate sequences.

A plethora of protease substrate sequences has been deposited in the MEROPS database in recent years.[7] They are frequently depicted as sequence logos[8] to visualize substrate preferences of proteases. Recently, we showed, how these sequence logos can be utilized to yield a quantitative metric for protease specificity.[9] Thereby, we also showed that information on protein sequences only is insufficient to predict protease specificity.

Furthermore, similarities in protease substrate recognition can be directly deduced via analysis of sequence logos.[10] We expect this approach to complement structure-based comparisons, as substrate sequences inherently contain information on binding site characteristics. Substrate peptides probe protease cavities via similar features as Cavbase[11] by binding of hydrophobic and hydrophilic, positively and negatively charged, and aromatic amino acids.

We performed a substrate sequence-based similarity analysis of the serine protease test set of Glinca and Klebe. Substrate data was downloaded from MEROPS, normalized to the respective natural abundance of amino acids,[12] and converted to vectors containing 20 amino acid probabilities at 8 substrate position P4 to P4′. After normalization, scalar products of these substrate vectors yield pairwise protease similarites ranging from 0 to 1.[10]

A comparison of all eleven serine proteases in the set yields a heat map depicting similarities in protease substrate recognition (see Figure 1). Furthermore, a hierarchical clustering based on complete-linkage yielding six clusters was performed as suggested by Glinca and Klebe.
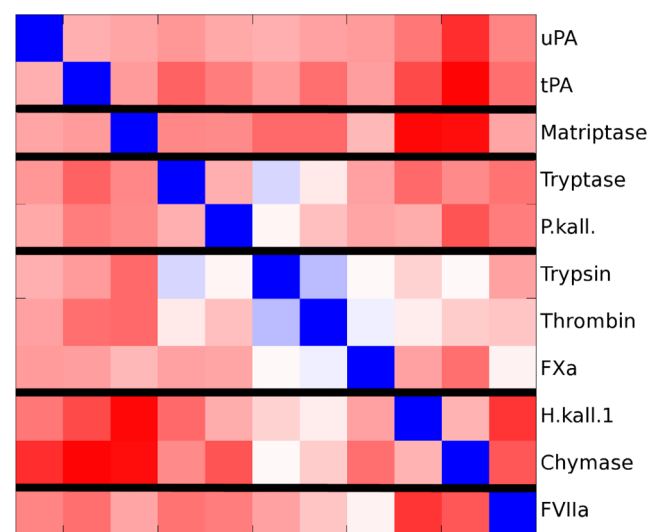
Figure 1. Heatmap obtained for clustering of proteases based on similarities in peptide substrates. Deep blue color depicts maximum similarity, whereas red regions show dissimilarity in substrate recognition. Six resulting protease clusters are separated with horizontal lines.

The resulting protease similarity map and clustering shows pronounced overlap with the cavity-based analysis of Glinca and Klebe. Thus, substrate sequence analysis shows similar discriminative power as an analysis of binding pockets. Urokinase-type (uPA) and tissue-type plasminogen activator (tPA) form a consistent cluster as in the study of Glinca and Klebe. Furhermore, our clustering nicely groups trypsin, thrombin, and factor Xa (FXa), known to show pronounced overlap in substrate recognition of small molecules.[13]

In conclusion we show that sequences can tell a lot on substrate recognition of proteases, if substrate sequences are considered. We are sure that peptide substrates comprise valuable information on protease recognition and propose their usage for the prediction of off-target effects, thereby complementing structure-based approaches.

Julian E. Fuchs
Klaus R. Liedl*

Institute of General, Inorganic and Theoretical Chemistry, and Center for Molecular Biosciences Innsbruck (CMBI),

University of Innsbruck, Innrain 80/82, A-6020 Innsbruck, Austria

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: klaus.liedl@uibk.ac.at.
Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ ACKNOWLEDGMENTS

■ ABBREVIATIONS

FVIIa, factor VIIa; FXa, factor Xa; H.kall.1, human kallikrein 1; P.kall, plasma kallikrein; tPA, tissue-type plasminogen activator; uPA, urokinase-type plasminogen activator

■ REFERENCES

(1) Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M. Drug-target network. *Nat. Biotechnol.* **2007**, *25*, 1119−1126.

(2) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197−206.

(3) Hert, J.; Keiser, M. J.; Irwin, J. J.; Oprea, T. I.; Shoichet, B. K. Quantifying the Relationship among Drug Classes. *J. Chem. Inf. Model.* **2008**, *48*, 755−765.

(4) Lounkine, E.; Keiser, M. J.; Whitebread, S.; Mikhailov, D.; Hamon, J.; Jenkins, J.; Lavan, P.; Weber, E.; Doak, A. K.; Cote, S.; Shoichet, B. K.; Urban, L. Large Scale Prediction and Testing of Drug Activity on Side-Effect Targets. *Nature* **2012**, *486*, 361−367.

(5) Glinca, S.; Klebe, G. Cavities Tell More than Sequences: Exploring Functional Relationships of Proteases via Binding Pockets. *J. Chem. Inf. Model.* **2013**, *53*, 2082−2092.

(6) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(7) Rawlings, N. D.; Barrett, A. J.; Bateman, A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* **2012**, *40*, D343−D350 version 9.8., accession 31.7.2013.

(8) Schneider, T. D.; Stephens, R. M. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.* **1990**, *18*, 6097−6100.

(9) Fuchs, J. E.; von Grafenstein, S.; Huber, R. G.; Margreiter, M. A.; Spitzer, G. M.; Wallnoefer, H. G.; Liedl, K. R. Cleavage Entropy as Quantitative Measure of Protease Specificity. *PLOS Comput. Biol.* **2013**, *9*, e1003007.

(10) Fuchs, J. E.; von Grafenstein, S.; Huber, R. G.; Kramer, C.; Liedl, K. R. Substrate-Driven Mapping of the Degradome by Comparison of Sequence Logos. *PLOS Comput. Biol.* **2013**, *9*, e1003353.

(11) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387−406.

(12) McCaldon, P.; Argos, P. Oligopeptide Biases in Protein Sequences and Their Use in Predicting Protein Coding Regions in Nucleotide-Sequences. *Proteins* **1988**, *2*, 99−122.

(13) Nar, H.; Bauer, M.; Schmid, A.; Stassen, J. M.; Wienen, W.; Priepke, H. W. M.; Kauffmann, I. K.; Ries, U. J.; Hauel, N. H. Structural Basis for Inhibition Promiscuity of Dual Specific Thrombin and Factor Xa Blood Coagulation Inhibitors. *Structure* **2001**, *9*, 29−37.