Exploiting structural information for target assessment

Andrea Volkamer and Matthias Rarey
Center for Bioinformatics, University of Hamburg, Bundesstrae 43, 20146 Hamburg, Germany

volkamer@zbh.uni-hamburg.de

Abstract: The amount of solved protein structures is continuously growing. Pharmaceutical research recently recognized the potential of computationally extracting information from this large data pool and using them for homology-based knowledge transfer to new structures. This article focuses on computational approaches for structure-based target assessment. Highlighted are novel approaches for target classification, i.e., druggability or function prediction, and target comparison together with the underlying methods for active site detection and description. Protein function predictions, e.g., yielded accuracies between 54% and 81% for predicting the correct main class, subclass and substrate-specific subsubclass on a test set of 26632 pockets. Besides the presentation of successful retrospective application studies of these methods, challenging tasks in the individual computational steps are discussed in this article.

1 Introduction

Drug discovery is a cost and time consuming venture, thus, computational approaches have long entered the early drug development pipeline. While the classical computer-aided application is screening of large compound data sets for new lead compounds, recent advances in structure elucidation and structural genomic projects, enabled high-throughput approaches for target screening, e.g., target prioritization, characterization and comparison. Learning from what is known, extracting patterns and transferring them to novel targets is one major goal in modern structure-based computational approaches. In this article, we present our recent developments addressing target assessment. Comparing proteins on a functional level means comparing their centers of action. Starting from the protein structure, our in house software DoGSite [VGGR10] can be applied for automated active site detection and representation by numerical descriptors. For classification scenarios, the descriptors are incorporated

into a support vector machine (SVM) to learn from the features of known member of specific classes. In this context, SVMs have been trained for target prioritization [VKG+12] and function annotation [VKRR13]. Another important task is the direct comparison of proteins by specific active site features. In the novel approach TrixP [BVH+13], active sites are represented by pharmacophoric triangles and closest homologous can be identified with high-throughput by screening a pre-calculated index consisting of active sites with known classification.

2 Methods

DoGSite [VGGR10] detects potential pockets solely based on the 3D structure of a protein. For this purpose, a grid representation of the protein is used. Grid points are labeled as free or occupied, dependent on their coverage by a protein atom. Subsequently, a Difference of Gaussian (DoG) filter is applied to find positions on the protein surface where the location of small sphere-like objects is favorable. Grid points are clustered into subpockets based on the calculated DoG value and neighboring subpockets are assigned to pockets. The **DoGSiteScorer** [VKG+12, VKRR13] is a generic approach for structure-based classification scenarios. The program automatically calculates numerical descriptors for self-predicted pockets. These global descriptors include size, shape and physicochemical properties of the pockets, e.g. volume, surface, ellipsoidal shape, enclosure, hydrophilic surface fraction, functional groups, element and amino acid compositions. Given a training data set with annotated classes, a discriminant analysis is used to select those descriptors which separate best between the different input target classes. Eventually, a support vector machine is trained on a property of interest related to binding or function. **TrixP** [BVH+13] is a novel method enabling fast index-based binding site comparisons. Recognition features are encoded via a triangular descriptor, holding physicochemical and shape information of the binding site [SR09]. Triangles are spanned between all present hydrogen bond donor, acceptor, and apolar point triplets and the shape of the binding site is captured by an 80-ray bulk, placed at the respective triangle's center. For efficient screening applications, an index with known binding site descriptors can be built and unlimitedly queried. For a new query protein, descriptors are calculated, the database is screened, and targets with matching descriptors are returned, superimposed and ranked by their estimated similarity to the query.

3 Results and Discussion

In the following, the results of the presented methods and their contributions to overcome the difficulties within the processed tasks are summarized. The pocket prediction method, **DoGSite**, was evaluated on the PDBBind and scPDB data set, containing 828 and 6754 structures, respectively, and detected the true ligand binding site in over 92% of the test cases. The major challenge in correct pocket and boundary detections arises from the nature of pockets. In contrast to the crystallized representatives, protein structures are flexible which produces a magnitude of potential pocket shapes from being small to large, shallow to deep, and homogeneous to highly branched. In this context, DoGSite especially convinced by its novel granular subpocket detection and a globular pocket ceiling definition. **DoGSiteScorer** was trained and evaluated for two different classification scenarios, namely druggability [VKG$^+$12] and function prediction [VKRR13]. The one prerequisite, when working with machine learning techniques, is a large and reliable data set to train the method on. Unfortunately, most available data sets are either too small or suffer from wrong or miss-annotated structures. For target prioritization, the method was trained on a subset of the recently published DD data set containing 1069 druggable, difficult and undruggable protein pockets and yielded accuracies of 88% in the testing phase. Next, to allow for enzymatic function predictions on different granularity levels, we created a data set of over 26000 enzymatic pockets and classified them with respect to the enzymatic classification (EC) scheme. Subsequently, models for predicting enzyme class, subclass or substrate-specificity based on structural features were build. Cross-validation studies showed accuracies of 68% for correct main class prediction and accuracies between 63% and 81% for the six subclasses. Substrate-specific recall rates for a kinase subset were 54%. Finally, our active site comparison tool **TrixP** was evaluated on several screening scenarios based on the scPDB data set. Using a subset of 769 similar and dissimilar protein pairs, a similarity cut-off was introduced with which similar pairs could be recovered in 82% of the cases, while dissimilar pairs were discarded with 99.5%. Screening the complete data set with four query proteins, 84-100% of the index-contained family members could be identified. Even correct subfamilies could be assigned for a small kinase data set. Again, flexibility upon ligand binding challenges structure-based comparison methods. The consideration of partial similarities based on matching triangle descriptors in combination with introduced tolerance values, allows TrixP to recover similarities between partially different conformations.

4 Conclusion

Due to the continuous growth in elucidated structures, learning from known structures has become more and more important, increasing the demand of fast and reliable computational approaches for target assessment. A set of software approaches has been presented which can be used for active site detection, target druggability and function prediction as well as target comparison. The methods have been evaluated on large data sets and showed good results in retrospective applications. A major drawback of automated methods is generally the dependence on the quality of the structures, the size of the available training data, the reliability of the annotated classes of their members as well as the necessary homology to securely transfer a specific property. Nevertheless, such methods allow to perform high-throughput target screening and, thus, to think outside the box, and detect similarities and cross-links between structures that would not have been found by hand.

References

[BVH+13] M. v. Behren, A. Volkamer, A. M. Henzler, K. T. Schomburg, S. Urbaczek, and R. Rarey. Fast binding site comparison via an index-based screening technology. *Journal of Chemical Information and Modeling*, 53:411–422, 2013.

[SR09] J. Schlosser and M. Rarey. Beyond the Virtual Screening Paradigm: Structure-Based Searching for New Lead Compounds. *Journal of Chemical Information and Modeling*, 49(4):800–809, 2009.

[VGGR10] A. Volkamer, A. Griewel, T. Grombacher, and M. Rarey. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *Journal of Chemical Information and Modeling*, 50(11):2041–2052, 2010.

[VKG+12] A. Volkamer, D. Kuhn, T. Grombacher, F. Rippmann, and M. Rarey. Combining global and local measures for structure-based druggability predictions. *Journal of Chemical Information and Modeling*, 52(2):360–372, 2012.

[VKRR13] A. Volkamer, D. Kuhn, F. Rippmann, and M. Rarey. Predicting enzymatic function from global binsing site descriptors. *Proteins: Structure, Function and Bioinformatics*, 81(3):479–489, 2013.