OpenEye Scientific Software

SiteHopper

A Unique Tool for Binding Site Comparison

P. C. D. Hawkins, J. Batista and G. L. Warren

Abstract

By comparing binding sites within and across protein families, relevant details about the functionality and selectivity of a target protein can be extracted, leading to useful insights for the development of new ligands. Typical approaches focus on the analysis of the protein sequence, which lead to uncertain predictions in cases of low sequence similarity. SiteHopper provides a powerful alternate method to the traditional use of sequence alignment for the purpose of drug discovery.

Introduction

The central idea of ligand-based virtual screening is that similar ligands often bind to similar protein binding sites. A corollary of this idea is that similar binding pockets might bind the same ligand, which makes binding site comparison a useful tool, e.g. for selectivity studies or the analysis of off-target effects. We have applied a shape-based approach to the realm of this problem in a new tool, SiteHopper, which defines binding sites by 3D shape and surface chemical features. These pocket definitions, called *patches*, can be collected from public or internal sources into a database. Using a query of the specific pocket of interest, this database can be searched, producing a hit list of aligned similar pockets.

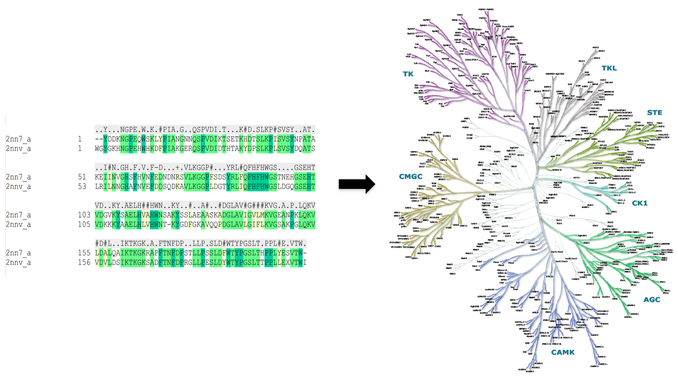


Figure 1: Clustering of the protein kinase family based on sequence. This example demonstrates that similar proteins can show high diversity in their sequence.

Method

SiteHopper relies on a definition of the binding site surface (obtained from a variety of methods). This surface is decorated with carbon atoms to provide a 3D surface representation, then color points (chemical features derived from active site residues beneath the surface) are added. This fully three-dimensional representation of the binding site is then used for the pairwise alignment of binding sites and scoring their similarity.

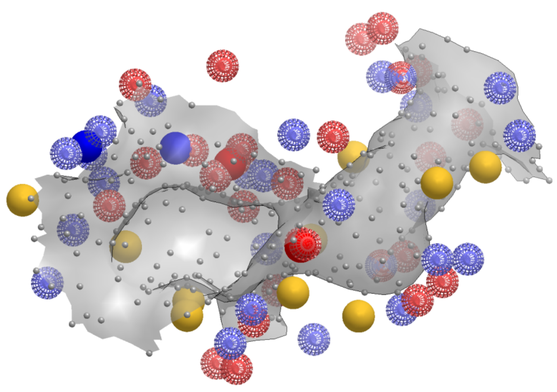


Figure 2: A SiteHopper binding site model or "patch".

Algorithm

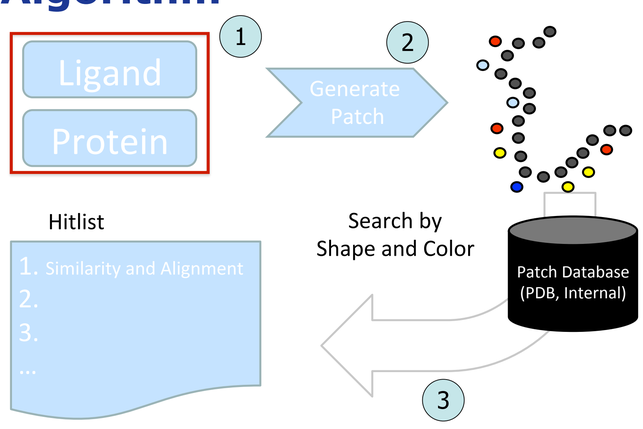


Figure 3: Overview of the general algorithm.

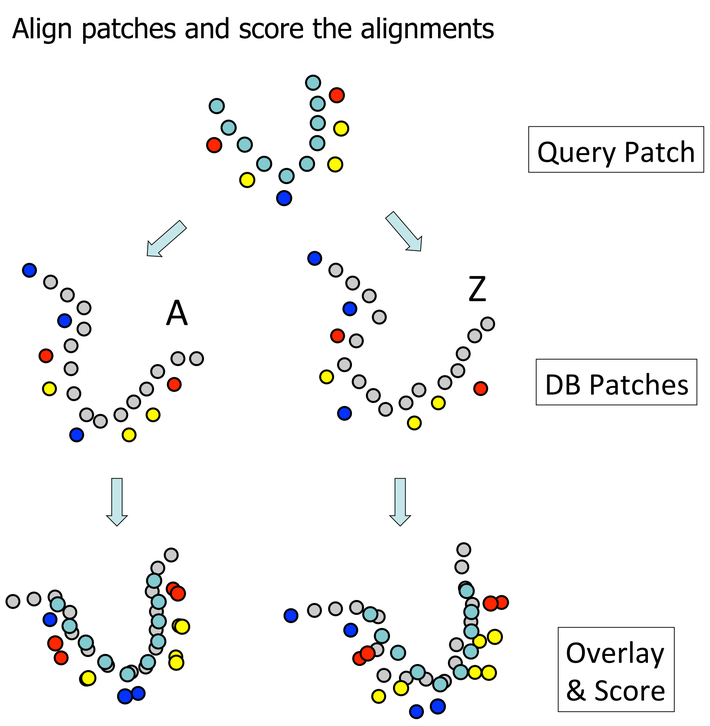
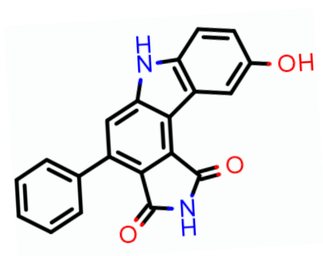


Figure 4: Schematic depiction of the alignment/scoring process (step 3 in Figure 3). Individual DB patches are aligned to the query patch based on matching the overall shape of the patch and also matching the position of color features like acceptors ●, donors ●, hydrophobes ●. The quality of the alignment is quantified by a score that represents the summation of two metrics: shape overlap and color overlap. This score is used to rank DB Patches.

Results

SiteHopper analysis of a kinase database enabled the detection of similar binding sites in proteins of low sequence similarity. An example is shown in Figure 4; the same inhibitor is bound by two kinases (Wee-1 and Traf-2) with similar affinity, implying a similar binding site. The binding site sequence similarity is low (< 45%), while the SiteHopper similarity is highly significant (occurring < 3 times in 1000 comparisons).



Methods	Score
SiteHopper	1.65
Seq Identity	23.3%
Seq Similarity	42.5%

Figure 5: Binding site similarity for Wee-1 and Traf-2 kinases detected by SiteHopper but not sequence.

Hierarchical clustering of a serine protease database (extracted from the PDB) by SiteHopper scores reproduced, almost exactly, the relationships between the sub-families defined by sequence. However, for the factor Xa branch a single trypsin structure was clustered in the factor Xa family, seemingly a mistake. A literature search revealed that this trypsin structure was the subject of a rational mutation program to reconstruct the factor Xa binding site [1]. As intended the resulting mutated trypsin (PDB code 1J17) has a binding site very similar to Factor Xa while its sequence is almost entirely that of trypsin.

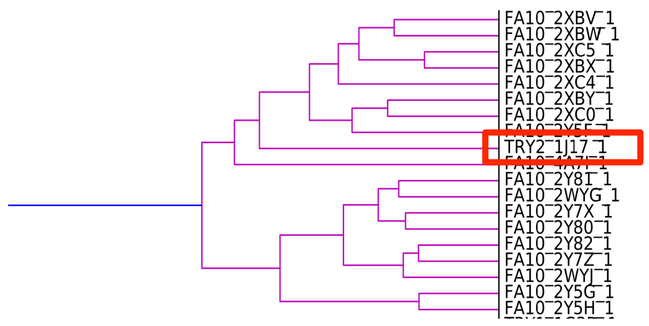


Figure 6: Hierarchical clustering of Factor Xa by SiteHopper. The trypsin structure representing a mutated Factor Xa binding site is highlighted.

Very recent work has compared sequence and SiteHopper similarities for predicting ligand selectivity across a large panel of kinases [2]. Statistical models for predicting cross-target ligand binding were generated by Gaussian Process modeling [3] of all-by-all similarity matrices for around 120 of these kinases. The models were then used to predict the binding of ligands for another 180 kinases to the original 120 kinases. SiteHopper similarities were found to be statistically significantly better than sequence similarities in predicting the binding of these new ligands [4].

Conclusions

SiteHopper is a highly flexible tool for binding site comparison. It uses a 3D shape and color feature representation of binding sites, and is therefore independent of protein sequence. Early studies have demonstrated its applicability to polypharmacology, especially in situations of low sequence similarity between proteins.

References

[1] Reyda *et al.*, *J. Mol. Biol.* **325**, 963 (2003).
[2] Dranchak *et al., PLOS One*, **8**, e57888 (2013).
[3] Gaussian Processes for Machine Learning. Rasmussen & Williams. MIT Press, 2006.
[4] Warren *et al.*, manuscript in preparation.

Acknowledgments

Dr. Robert Tolbert and Dr. Matthew Geballe for the implementation of SiteHopper. Dr. Brian Kelley for his initial studies and data.

OpenEye Scientific Software

9 Bisbee Court
Suite D
Santa Fe, NM 87508

505.473.7385
info@eyesopen.com
www.eyesopen.com