Screening Ligand Databases for Potential Ligands to Protein Active Sites

Volker Schnecke and Leslie A. Kuhn, Dept. of Biochemistry, Michigan State University

Abstract

Screening of structural databases to identify potential docking candidates for active sites of proteins has become a competitive methodology for creating new leads in drug design. The efficiency of the screening process depends on sophisticated techniques to eliminate infeasible candidates during the search and to provide a ranking for the identified ligands. A screening and docking tool SPECITOPE is presented which uses distance geometry to rule out the majority of infeasible candidates, because they are unable to match the predefined template. This active-site template is only defined by four or five points, from which hydrogen bonds are required between the potential ligand and protein surface atoms. During the search, SPECITOPE combinatorially matches all hydrogen-bond donors and acceptors of the screened molecules to these template points. By skipping all molecules which are not able to form these hydrogen bonds based on distance constraints, the transformation of potential docking candidates into the active site and the check for shape complementarity have to be done only for a small subset of all screened molecules. SPECITOPE has been tested on a database of about 140,000 peptides to identify potential ligands to the target proteins anti-lysozyme F_{ab}, thrombin, subtilisin, β-lactamase, and collagenase. The runtime for screening all 140,000 peptides is between 10 minutes and 10 hours on a single processor workstation, depending on the template configuration. SPECITOPE has identified both known ligands and new potential ligands with good shape and chemical complementarity to the target active site and conformational similarity to known ligands.

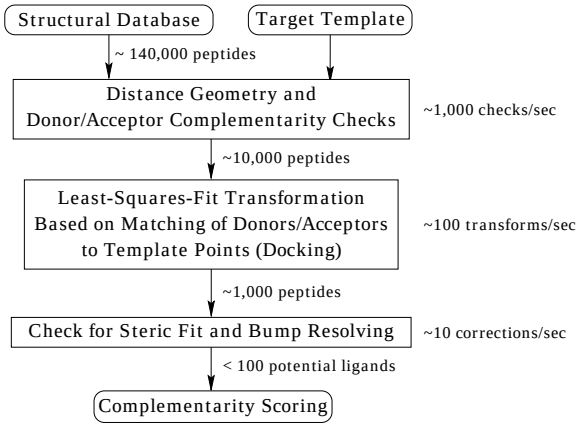
Introduction

With increase of computational power and availability of structural information for proteins and smaller molecules, rational drug design has become a competitive methodology to identify new inhibitors. While there is no way to replace the *in vitro* and *in vivo* tests during drug design, the *in computo* tests are able to propose new leads and thus provide a shortcut during the early design stages. Another advantage of the computational approach to analyzing protein-ligand interactions is the ability to test the relative importance of different complementarity features — shape, flexibility, hydrophobicity and the number of hydrogen bonds and salt bridges — for known complexes and thus derive a better scoring function for new potential ligands.

During the search for new inhibitors, large datasets of known compounds are screened for molecules with similar shape and chemical activity to known inhibitors, or for molecules with shape and chemical complementarity to a given active site. The main aim during this screening process is to find an efficient way to rule out infeasible molecules and provide a ranked list of, say, ten potential ligands for optimized docking.

Hydrogen bonds between the target active site and the ligand are assumed to give only moderate contributions to the overall stability of a protein-ligand complex, but are important for specifity. Taking the positions of water molecules or H-bonding atoms of known ligands bound to the active site of the target protein, a template consisting of interaction points can be designed. During the screening process, potential docking candidates can be identified by their ability to form hydrogen bonds to the target protein. This is done by matching donors or acceptors, as appropriate, to the template interaction points.

SPECITOPE



The figure above provides a schematic overview of the screening process in SPECITOPE. For all peptides in the database up to ten different checks are executed (see outline below). In each step those molecules are ruled out which

do not meet the particular threshold. Note that steps 1 to 4 do not require a one-to-one correspondence of peptide donor/acceptor atoms with template points. Step 5 combinatorially checks all possible matchings of n atoms in the peptide set to the n template points. Through step 7 only the donor/acceptor atoms of the peptides are considered, and the steric fit is only checked for those molecules which pass all steps up to this point. The checks become more complex by the end, thus the main goal is to rule out a maximal number of infeasible ligands in the early stages to optimize the effectiveness of the screening process.

For each peptide in database:

1. Compare shortest and longest distances for any donor/acceptor pair with shortest and longest distances between any pair of template points.

For each set of n donors/acceptors in peptide:

2. Compare H-bonding compatibility (number of donors and acceptors in set) to the template.

3. Compare shortest and longest distances for any donor/acceptor pair in set with shortest and longest distances in template.

4. Compute root-mean-square deviation for set and template distance lists.

For each matching of n donors/acceptors to the n template points:

5. Compute minimal distance matrix error for this set.

6. Check H-bonding complementarity of matched peptide atoms and template points.

7. Compute root-mean-square deviation for least-squares fit of peptide atoms to template points.

For the transformed peptide:

8. Check for overlaps of peptide main-chain atoms and target atoms and determine if they can be resolved by iterative translations.

9. Check for overlaps of peptide side-chain atoms and target atoms and resolve them, if possible, by minimally rotating side chains.

10. Evaluate chemical complementarity.

Ranking of Potential Ligands

It is important to provide a scoring for SPECITOPE's potential ligands to determine the best candidates to be further inspected and optimally docked into the active site. The scoring function which is currently used to rank the ligands is based on three components:

- The overall number of hydrogen bonds between peptide and target protein.
- The similarity of the main-chain conformation of the peptide to a known ligand, which binds to the active site (for validation of known complexes).
- The hydrophobic complementarity of the interface between the molecules.

The hydrophobicity measure is based on a statistical survey of atomic hyrophobicity in known protein structures [1]. For computation of the complementarity of the interface, all target atoms within 4.0Å of any ligand atom are considered. The contribution of a single ligand atom is based on its own hydrophobicity divided by the absolute difference of this value from the average hydrophobicity of the surrounding active-site atoms.

The Test Set

SPECITOPE has been used to screen a database of approx. 140,000 peptides to identify potential ligands to the active sites of the target proteins subtilisin, β-lactamase, thrombin, anti-lysozyme F_{ab}, and collagenase. The peptides have been taken as all overlapping linear sequences from the structures of 553 non-homologous chains [2] of the Brookhaven Protein Data Bank [4].

Target	PDB code	Template						peptide size
		interaction points		distances				
		acceptor	donor	shortest	longest	avg.	std. dev.	
Antibody	1mlb	2	2	5.8Å	25.3Å	14.6Å	49.1	7
Thrombin	1hah	4	1	2.3Å	5.3Å	5.2Å	3.5	3
β-Lactamase	rtem*	4	0	2.3Å	10.6Å	6.3Å	10.5	6
Subtilisin	1s01	1	4	2.4Å	11.5Å	7.1Å	11.2	6
Collagenase	1cge	3	2	2.2Å	8.0Å	4.3Å	3.9	6

*Coordinates provided by Natalie Strynadka and Michael James, University of Alberta, Edmonton

Different approaches have been used to design the templates for the five target proteins:

Anti-lysozyme F_{ab}: Superimposing antibody-lysozyme complex (1mlc) to free antibody (1mlb) and taking linear binding pattern of lysozyme (residues e41 to e47), oriented in such a way that it can form four hydrogen-bonds to the free antibody and taking the corresponding four donor/acceptor positions as template points.

Thrombin: Superimposing four thrombin-PPACK complexes (1abi, 1hai, 1ihs, 1ppb) onto the ligand-free thrombin (1hah) and taking the average positions of five donors and acceptors in PPACK, which form hydrogen bonds to thrombin atoms or waters in the active site.

Subtilisin: Superimposing subtilisin/eglin-c complex (2sec) to free subtilisin (1s01) and taking positions of five waters in 1s01 as template points.

β-lactamase: Superimposing rtem-BLIP (beta-lactamase inhibitory protein) complex to free β-lactamase and taking four displaced waters in the active site as template points.

Collagenase: Superimposing three complexes (1cgl, 1hfc, 1cgl) to free collagenase (1cge), and using the average positions of five donors or acceptors in the inhibitors as template points.

Results

SPECITOPE has identified between 11 and 180 potential docking candidates for the different test cases. For those cases where known peptidyl ligands are included in the database, these have been identified, plus many peptides with similar shapes to known ligands have received high rankings during the screening process. Figure 1 shows the surface of free thrombin with the three highest rated ligands. The black molecule is the native ligand PPACK, which has been superimposed from the complex.

The runtime of SPECITOPE varies between 10 minutes for anti-lysozyme F_{ab} to almost 10 hours for collagenase on a single processor Sun Ultra 1 workstation. The short runtimes of screening for potential ligands to anti-lysozyme F_{ab} is caused by the elongated template (largest distance 25.3 Å), for which distance geometry rules out most candidates. In case of thrombin (15 minutes), only tetrapeptides are screened, involving much fewer match comparisons. For β-lactamase (3 hours), subtilisin (7 hours), and collagenase (10 hours), 139,167 hexapeptides have been screened. The reason for the different runtimes are in these cases not as obvious as in the previous two. For β-lactamase and collagenase, quite uniform templates have been designed, so that in both cases the majority of the peptides pass the distance checks, and have to be transformed into the active site and checked for steric fit. A more detailed analysis showed that in both cases SPECITOPE spent most of its time in the combinatorial enumeration of all possible matchings of hydrogen donors/acceptors to template points when computing the distance matrix error. For collagenase, the cumulative runtime for this particular step is six hours, much larger than the computation time of one hour, which is the total amount spent for this check when screening for ligands matching the four-point β-lactamase template. The reason for this effect is that the template for collagenase contains an additional point.

Conclusions and Future Research

From the results obtained with the current version of SPECITOPE, already the following points can be stated:

- Criteria based on distances between hydrogen-bond donors and acceptors on the target active site and the ligand have been proven to be an efficient way to rule out infeasible docking candidates when screening a large structural database.
- A scoring function, including the number of hydrogen bonds and the hydrophobic complementarity of the interface, has been developed and shown to identify ligands from known complexes.
- A virtual peptide database consisting of peptide fragments from PDB entries has served as a set of potential ligands. These peptides are considered to have rigid (main-chain) and flexible (side-chain) regions, and the same approach can be applied to more general organic molecules.
- Although not explicitly required by the template, SPECITOPE's high-scoring peptides have similar shapes to known ligands.
- The runtime of SPECITOPE is highly dependent on the composition of the template: elongated templates like that for 1mlb favor the distance-geometry based checks to rule out infeasible ligands, whereas compact templates like those for 1cge and 1s01 let many peptides pass to the more expensive transformation and van der Waals overlap computations.

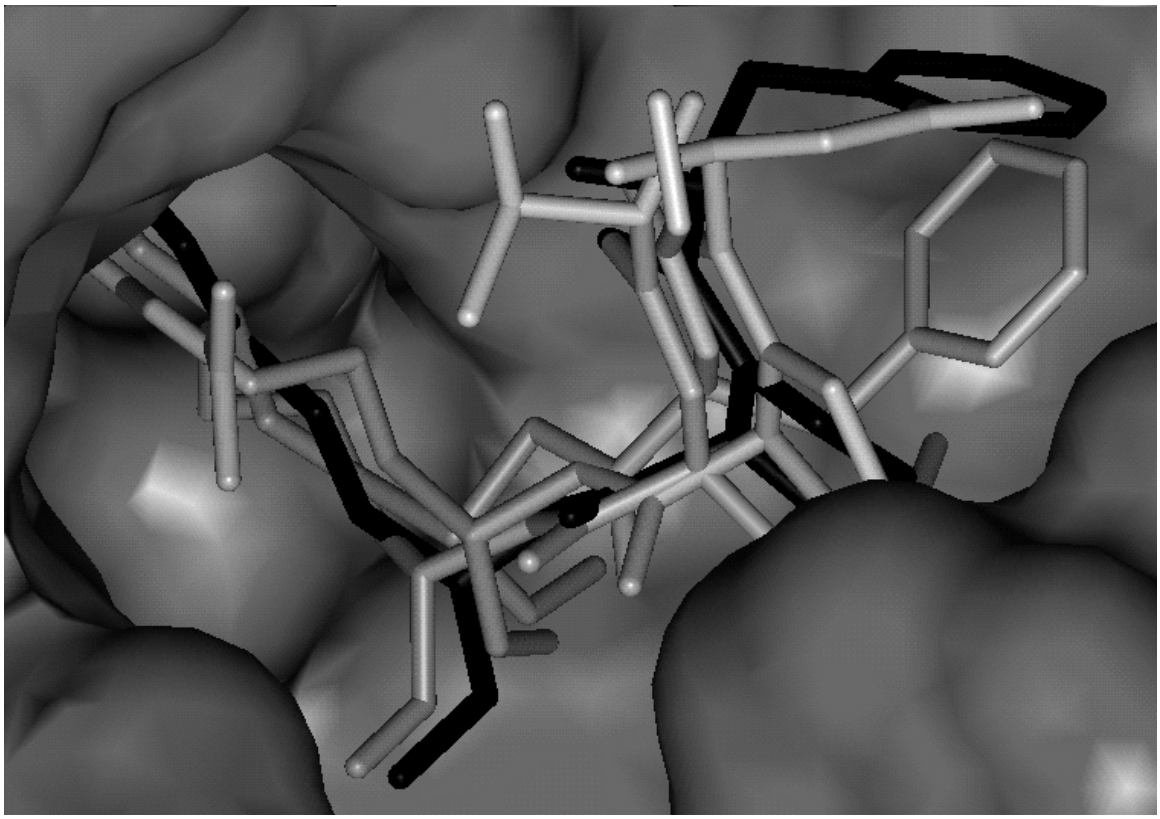


Figure 1: The top three ligands (grey) for thrombin (grey surface) are shown together with the native ligand PPACK (black)

To make SPECITOPE more scalable and realistic, it will be extended in the following directions during the next months:

- SPECITOPE will be extended by a genetic algorithm to more extensively sample subsets of a larger set of template points, thus lowering the bias of selecting a template.
- Entries from the Cambridge Crystallographic Database System (CSDS) will be screened to identify non-peptidyl organic ligands. A module already exists to derive hydrogen-bonding information from CSDS files.
- The complementarity measure and the ranking function will be expanded by considering van der Waals packing, electrostatic forces and orientation of the hydrogen bonds.
- A tool *Consolv* [3] developed in our lab will be integrated to evaluate water molecules in the active site conserved upon ligand binding.

References

[1] L. A. Kuhn, C. A. Swanson, M. E. Pique, J. A. Tainer, E. D. Getzoff, *Atomic and Residue Hydrophilicity in the Context of Folded Protein Structures*, Proteins: Structure, Function, and Genetics, Volume 23, Number 4, 1995, 536–547

[2] U. Hobohm, C. Sander, *Enlarged representative set of protein structures*, Protein Science, Volume 3, Number 3, 1994, 522–524

[3] M. L. Raymer, P. C. Sanschagrin, W. F. Punch, S. Venkataraman, E. D. Goodman, L. A. Kuhn, *Predicting Conserved Water-mediated and Polar Ligand Interactions in Proteins Using a K-nearest-neighbors Genetic Algorithm*, J. Mol. Biol., Volume 265, 1997, 445–464

[4] Brookhaven National Laboratory, *Protein Data Bank*, `http://www.pdb.bnl.gov`