Looking for the needle in the haystack

An algorithm for the rapid recognition of local structures in proteins

by Andreas Hoppe and Cornelius Frömmel

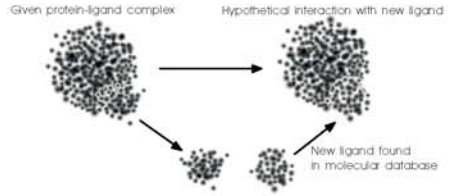
Abstract

We present a new algorithm for superposition of two sets of vectors which represent atom coordinates of a protein or another molecule. It is developed to find the best superpositions of the majority of a set of few atoms (the needle) with a subset of a set of many (the haystack). The algorithm is not dependent on either chemical information like covalent bonds or protein sequence information. It can handle the many thousands of atoms of a large protein. Compared to other algorithms the concentration to the needle-haystack-problem greatly reduces the problem complexity and therefore allows a considerable increase of the efficiency. Our algorithm utilizes a brute force strategy which guarantees to find a certain superposition in a given tolerance if it exists. The use of a 7-dimensional numerical method for the determination of the optimal translation and rotation represented by quaternions results in a large likelihood to find the superposition with globally minimized root-mean-square deviation. Pruning increases the efficiency such that the algorithm can be used to screen a large database in an acceptable time.

We give an example for a successful database search for a potentially alternative ligand. As another application we present the automatic localization of active sites of proteins of the subtilisin protein family.

Application that inspired this work:

Automated search for alternative ligands using the 3D structure of a protein-ligand complex and a 3D structural database



There are two different approaches:

- Docking
  - More accurate in binding prediction
  - Does not require the structure of the original ligand
- Superposition
  - Suited to suggest a lead substance for further chemical modifications
  - Uses the usually near-optimal solution given by reality

Course of alternative ligand search by superposition:

- Selection of relevant atoms, usually the superficial atoms of the ligand in contact to the protein. This set will be called model.
- Superposition of the model with the targets, entries of a 3D structure database, selection of the best candidates using a suitable measure function
- Validation of the results (e.g. a modified docking algorithm)

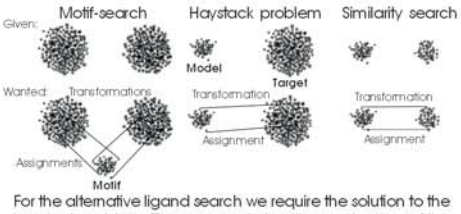
This work concentrates on superposition.

Superposition

Let two atom sets *A* and *B* be given. A superposition is an assignment, a bijective map from a subset *S* of *A* to a subset of *B*, and a Cartesian transformation *T*. A superposition is measured using the RMSD (root mean square deviation), the quadratic mean of the distances of atoms in *S* (transformed by *T*) to their assigned counterparts in *B*. The superposition problem is the search for the superposition with minimal RMSD or for all superpositions with an RMSD below a given threshold.

Different tasks of superposition

Based on the cardinalities of *S* in comparison to *A* and *B* we can classify:

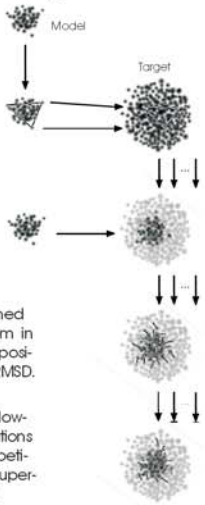


For the alternative ligand search we require the solution to the haystack problem. The proposed algorithm makes use of the fact that for an acceptable superposition *S* consists of the majority of *A*. This guarantees a superior efficiency compared to algorithms concentrating on the motif search problem.

The superposition measure pRMSD

- Modification of RMSD
- Any atom in *A* not assigned counts penalty c
- c is the upper cut-off for a distance of assigned atoms
- Quadratic mean of distances and penalties

The needle-haystack-algorithm:



1. Fix an anchor (three non-colinear points) in the model.
2. Find a set of atoms in the target which approximately superpose with the anchor - the anchor match. The following steps are repeated for every anchor match.
3. Compute the Cartesian transformation determined by the anchor match and apply the inverse of the transformation to the model.
4. For each atom in the transformed model try to find one near atom in the target. The result is a superposition which is valuated by its pRMSD.
5. For the superpositions with the lowest pRMSDs use small modifications of the transformation and a repetition of step 5 to improve the superposition in terms of the pRMSD.

Proof of completeness

For a subproblem of the haystack problem the proposed algorithm is guaranteed to find the optimal superposition.

If the model is a subset of the target (transformed with a Cartesian transformation) the needle-haystack-algorithm finds it. The proof uses the fact that the possible anchor matches are searched "brute-force". Therefore one anchor match leads to a superposition with pRMSD 0. This superposition will pass the rest of the steps of the algorithm since it is always recognized as the best.

The proof remains valid if the atoms of the model are not only transformed but also displaced up to a certain distance.

Features of the algorithm

- Full-atom representation
- Sequence independency
- Independence on chemical properties
- Efficiency in time and computer resources
- Feasibility even if the target molecule is large

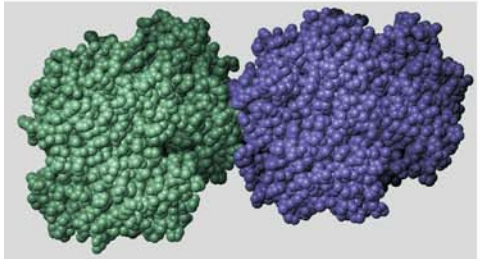
The algorithm is suited for:

- Rapid recognition of already identified local 3D structures
- Classification of protein on the basis of crucial local structures
- Identification of the position of sterical atom configurations
- Database search for substances containing a given atom configuration

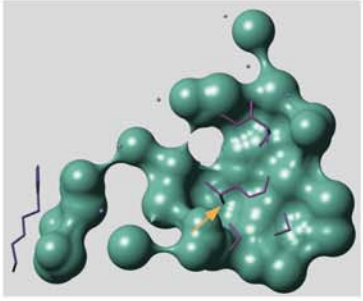
Applications:

Search of a ligand to prevent HbS polymerization

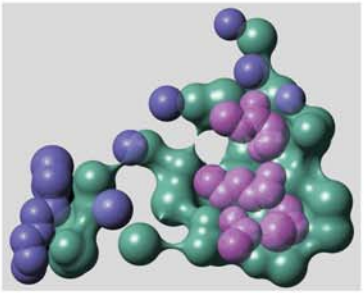
The pathogenesis of sickle-cell anemia involves the formation of hemoglobin macrofibers from mutant hemoglobin HbS. The genetic substitution of glutamic acid by valine at the sixth position in the β chain creates a binding region to the δ chain of another hemoglobin molecule:



A closer look at the contact region reveals a double mould in the β chain (green, Connolly surface) which is filled by 21 atoms of the δ chain (pink and violet, sticks and small balls):

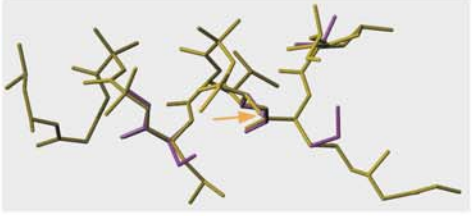


For this purpose we select the atoms necessary to fill the double mould - they are coloured pink (as opposed to the other contact atoms which are coloured violet). The next picture shows the atoms of the δ chain as van-der-Waal spheres:



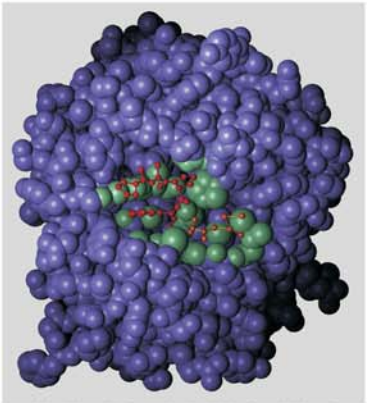
The question is if there is a substance which can also fill this binding pocket?

The search in the Cambridge Structure Database revealed that a part of Leucinostatin A can be superposed with 18 of the 21 atoms with a RMSD of 0.28Å. It is a likely binding partner since the atomic structure of Leucinostatin A is similar to the local structure of the contact atoms of the δ chain H of HbS. In particular the atom 7993, Cγ of Val6 of the δ chain, which fills a narrow pocket of chain β corresponds to an oxygen atom in Leucinostatin A (marked with an orange arrow in both figures).



Similarity of subtilisin binding sites

Similar functionality of enzymes is caused by a similar topology of their substrate binding region. The algorithm presented allows to localize the binding site of a protein only by the geometric similarity even if no ligand-complex is structurally known. Starting point is the binding region (green, van-der-Waal spheres) of Proteinase K in complex with an inhibitor (small red and yellow balls and sticks).



We were able to locate the substrate binding site in other members of the subtilisin protease family: Subtilisin (PDB codes: 1be8, 1bh6, 1cse, 1gci, 1jea, 1s01, 1s02, 1sbh, 1sbi, 1sbn, 1sbt, 2sbt, 1sca, 1scb, 1scd, 2sec, 1sel, 1sib, 2sic, 5sic, 2sni, 2st1, 1st2, 1st3, 1sub, 1suc, 1sud, 1sue, 1sup, 1yja), Proteinase K (1cnm, 2prk, 3prk), Thermitase (1tec, 2tec, 3tec, 1thm), Alcalase (1a10), Peptidase (1mee, 1sbc), M-protease (1mpt), Savinase (1svn).

We were able to reject other proteins: Myeloblastin (1fuj), Chymotrypsin (2gch, 4gch), Trypsin (5ptp, 1try), Subtilisin prosegment (1spb), Plasminogen activator (1tpk), Thrombin (1vr1).

Our findings correlate with known substrate binding affinities also if sequence similarity does not, in particular the sequence similarity of Proteinase K (1pek) to Thermitase (1thm) is far lower than of Proteinase K to Trypsin (1try) and Chymotrypsin (2gch).