Active site of lipase from XRD and SiteEngine: A Comparison

A

thesis Submitted

in partial fulfillments of requirements for the

Degree of

Master of Science in Chemistry

Submitted By

Nishu Jain

(Reg. no. 30702011)

Under the supervision of

Dr. Amjad Ali

Lecturer

School of Chemistry and Biochemistry

Thapar University

Patiala 147004

June 2009

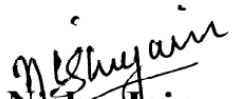
Acknowledgements

I take this opportunity to thank my guide Dr. Amjad Ali for his guidance and support in doing this thesis. He has instilled in me the knowledge and motivation to learn more about the subject.

I am grateful to Prof. Susheel Mittal for approving this thesis to me.

I am thankful to all the Ph.D. scholars for their timely help and support.

I thank all my friends who constantly motivated me and supported me throughout the thesis.



Nishu Jain

Candidate's Declaration

I hereby declare that the work being presented in the thesis entitled **"Active site of lipase from XRD and SiteEngine: A Comparison",** in partial fulfillment of the requirements for the award of the degree of Masters in Chemistry, School of Chemistry and Biochemistry (SCB), Thapar University, Patiala, is my own work during the period of Jan 2009 to May 2009, under the supervision of Dr. Amjad Ali, Lecturer , School of Chemistry and Biochemistry, Thapar University, Patiala. I have not submitted the matter embodied in this thesis for the award of any other degree.

Patiala

Date:

Nishu Jain

This is to certify that the above statement made by the candidate is correct and true to the best of our knowledge.

Dr. Amjad Ali
Project Supervisor,
Lecturer,
(SCB),
Thapar University

Dr. Susheel Mittal 29/6/09
Head, SCB
Thapar University

Certificate

This is to certify that the thesis entitled **"Active site of lipase from XRD and SiteEngine: A Comparison",** being submitted by Ms. Nishu Jain in partial fulfillment of the requirement for the award of degree of Master of Science in the School of Chemistry and Biochemistry, Thapar University, Patiala, is a bonifide work carried out under the supervision of Dr.Amjad Ali and that no part of this project has been submitted for the award of any other degree.

Supervisor,
(Dr. Amjad Ali)
Lecturer,
Thapar University

Head,
(Dr. Susheel Mittal)
School of Chemistry and Biochemistry,
Thapar University

Dean, Academic Affairs,
(Dr. R.K. Sharma)
Thapar University

List of Contents

Introduction

Molecular recognition is one of the central processes in molecular biology. Comparison and detection of binding sites is a key step in the prediction of potential interactions. Since proteins function by interacting with other molecules, similarity in the binding patterns of proteins is closely related to similarity in their biological functions. There are two potential ways to infer the function of a novel protein. The first is to recognize a sequence or fold similarity with a protein whose function is known. However, a similar fold does not necessarily imply a similar function. For example, proteins with the same fold, like TIM (triose phosphate isomerase) barrels i.e. conserved protein fold consisting of eight α-helices and eight parallel β-strands that alternate along the peptide backbone., can have multiple functions. On the other hand, proteins with different folds, like subtilisin and trypsin, can share the same function. The alternative approach, is to investigate the physico-chemical patterns and shape of the protein molecular surface.

Proteins are assumed to perform similar functions if they share similar binding patterns and recognize similar binding partners, even if they have different sequences and (overall) fold homology. Identification of regions on the surface of one protein that resemble a specific binding site of another is especially important for the following three applications

(1) Functional analysis and classification: Recognition of similarity in binding pattern to a well known protein may help in gaining a better understanding of its function and activation mechanism. These are crucial for the development of targeted drug leads like inhibitors. Functional annotation of newly determined structures can be a significant contribution to the Structural Genomics initiative.

(2) Potential ligands and ligand fragments: Analysis of ligands bound to proteins with similar binding sites may provide hints of chemical groups that can be used to develop a drug for the protein target. The method can be used for lead generation and optimization as well as for de novo drug design.

(3) Prediction of side-effects: Proteins with similar binding sites may bind the same drug and therefore may potentially cause side-effects. Thorough investigation of such proteins during the drug design process is important for the development of more specific drug lead.

Other methods that are commonly used for suggestion of new ligands or ligand fragments and for predictions of side-effects are alignment of small molecules and docking. These techniques model the interactions of the receptor with specific ligands and therefore do not analyze all potential interactions that a specific binding site may form. This is particularly important, since a single protein-binding site may have several binding patterns. Not only can the same binding site bind different ligands with different functional groups, but there is also evidence that at least in some enzymes a single compound can bind in different ways. A wide variety of methods have been developed for protein structural alignment. Most existing methods describe a protein structure by its C atoms and seek to maximize the overall similarity of the structures. However, when there is no fold similarity between the aligned structures, these methods usually do not provide a biologically significant alignment. Analysis of the similarities between binding sites can complement these techniques, ensuring full exploration of available structural data.

Introduction to SiteEngine

SiteEngine predicts regions that can potentially function as binding sites. The methods is based on recognition of geometrical and physico-chemical environments that are similar to known binding sites.i.e. Recognizes regions on the surface of one protein that resemble a specific binding site of another[1]**.**

Recognition of regions through which protein molecules function and interact is crucial for prediction of molecular interactions which govern most of the cellular processes. that recognizes regions on the surface of one protein that resemble a specific binding site of another. This may suggest the similarity of their binding partners and biological functions. Unlike methods that compare the locations of the backbone atoms or the

identity of the amino acids, the presented method takes into account the physico-chemical properties of both the backbone and the side-chains. Therefore it can recognize similar binding patterns shared by proteins that have no sequence or fold similarity. SiteEngine is highly efficient and suitable for large scale database searches of the entire PDB.

The biological significance of the SiteEngine method is validated on a set of biological applications. First, Introduce a benchmark dataset which is used to construct two databases: one of complete protein structures and the other of binding sites. These databases are used to perform three types of search applications:

(1) A given functional site is searched against a large set of complete protein structures
(2) A potential functional site of a protein of interest is compared with known binding sites

(3) A complete protein structure is searched for the presence of an a priori unknown functional site, similar to known sites.

While the second application compares between already known binding sites, the first and the third can recognize novel regions that can function as binding sites. From the biological standpoint, the first and the second applications may identify secondary binding sites of drugs that may lead to side effects. The third application finds new regions that may provide targets for drug design. Each of the three applications may aid in assigning a function and in classification of binding patterns.

In each application SiteEngine has successfully recognized specific types of protein binding sites such as estradiol binding, adenine and ATP binding sites that were used as queries. The same binding sites were further used to search the ASTRAL dataset constructed from the entire PDB. Since SiteEngine searches a complete structure of each protein in a matter of seconds, finding the first application to be the most reliable for such large scale applications. The method was also applied to classification and functional annotation of novel proteins determined as part of the Structural Genomics project.

Literature Review

Recognition and comparison of regions through which protein molecules function and interact are crucial for the prediction of molecular interactions, which govern practically all cellular processes. Consequently, a broad range of tools for sequence and overall structural alignment are routinely used by the scientific community in the analysis of biological processes and the prediction of function. However, the overall similarity of the sequences and folds does not necessarily imply similarity of biological function . It has been shown that proteins with the same fold can have different functions, and that proteins with different folds, such as serine proteases or zinc-binding proteins, can share the same function. Since proteins function by interacting with other molecules, similarity in their biological function is related to the similarity of their corresponding binding regions. These may be sequentially non-continuous regions with no common patterns of amino acids but sharing a set of physicochemical properties which create similar surface regions. Several approaches have been proposed for the recognition of such functional sites . These are important for drug design as well as functional annotation and biological classification.

SiteEngine is an efficient method for the recognition of functional sites in protein structures . It is motivated by several goals.[2] First, analysis of compounds bound to proteins with similar functional sites may suggest chemical groups and scaffolds that can be used in drug design and optimization. SiteEngine can also assist in the recognition of proteins with similar binding sites that can potentially cause side-effects. In addition, it can be applied to recognize regions on the surface of a novel protein that are similar to functional sites of known proteins. This may contribute to a better understanding of the novel proteins's function and activation mechanism. Furthermore, classification of proteins according to their functional site may facilitate the development of more efficient database organizations and search schemes.

Schmitt *et al.* [3], for each amino acid group atoms with similar physicochemical properties into functional groups, which are represented by three-dimensional points in space, denoted as pseudocenters. Each pseudocenter represents one of the following properties

important for protein– ligand interactions : hydrogen-bond donor (DON), hydrogen-bond acceptor (ACC), mixed donor/acceptor (DAC), hydrophobic aliphatic (ALI) and aromatic, pi interactions (PII). Construction of a smooth molecular surface as implemented by **Connolly *et al*[4].** and retain only pseudocenters that represent at least one surface exposed atom. When considering binding sites, consider only to the surface regions that are within 4s of the binding partner.

Calculate all possible transformations that superimpose the input binding site on a similar surface region of the other molecule. The algorithm is based on efficient hashing and matching of almost congruent triangles defined by triplets of pseudocenters. The hashing of the triangles is done with a key that consists of the three parameters of side lengths of a triangle and of an additional physicochemical index, which encodes the properties of its nodes. Each pair of matched triangles defines a candidate transformation which can superimpose the input binding site on a certain region of the complete protein. Similarity of the physicochemical properties and shapes aligned by each transformation is scored using a set of hierarchically applied scoring functions and a list of top ranking solutions is selected.

Several methods have been developed to identify specific three-dimensional patterns of amino acid side-chains. **Artymiuk *et al*[5].** represented each side-chain by pseudo-atoms and used a subgraph-isomorphism algorithm to identify the spatially conserved patterns. This algorithm (ASSAM) was recently enhanced to include additional constraints such as: the secondary structures, the solvent accessibility and the disulfide bridges. Wallace et al. have introduced “coordinate templates”. These allow recognition of the “catalytic triads” that are typical for some of the protein families, like serine proteases, triacylglycerol lipases, ribonucleases. and lysozymes. Using atomic representation, the geometric hashing technique was applied to efficiently compare a query protein to the template of the catalytic triad[6]. This algorithm (TESS) has been recently updated by **JESS**, which is flexible and unconstrained by the template syntax. **Binkowski *et al.*[7]**, have recently presented an elegant approach to assess the similarity of sequence patterns of surface pockets and voids, which are conveniently organized in CASTp[8]. However,

methods that recognize patterns of residues that are conserved in their 3D positions and in their amino acid identities are not always applicable. There are biological examples of proteins that can bind the same binding partners without sharing any conserved patterns of amino acid residues. **Rosen *et al*[9]**,searched for a site on the protein surface that resembles a specific, known active site. The molecular surface was represented using sparse critical points defined by **Lin *et al*[10-11]**.,The translation and rotation invariant characteristics of pairs of critical points were used as a key for the geometric hashing procedure. In addition, the reliability of surface comparisons in searches for active sites was examined. It was concluded that although pure geometric surface matching is capable of finding biologically correct solutions, utilizing additional chemical "labeling" information is required to correctly rank and analyze the obtained solutions.***Kinoshita et al***[12]**.**, performed clique detection on the vertices of the triangulated solvent- accessible surface. They constructed a database of binding sites, eF-site, and used a structure of a complete protein structure to search it. However, the number of vertices in their surface representation is too large and it is too sensitive to conformational flexibilities. One of their conclusions was that other representative surface points may be more effective for robust and accurate comparisons. An important contribution was recently published by **Schmitt *et al* .** Generic pseudo centers that efficiently encode the physico-chemical properties important for molecular interaction. Each amino acid residue of a protein is represented as a set of such centers. Assuming that small molecule binding sites are detected in cavities, they constructed a database of binding sites Cavebase, which is integrated with Relibase. The clique detection algorithm was used to retrieve cavities that are similar to a specific query cavity. The solutions were ranked according to the similarity of property-based surface patches. Here is a novel method i.e. capable of handeling large protein structure in a matter of seconds.

Objective:-

- To recognize the binding sites in lipase with the help of SiteEngine
- To dock the metal ion or any other ligand on the binding site of lipase and to see the effect of metal ion on the activity of lipase.

Methods And Materials

Working of SiteEngine

Stages of the SiteEngine server:

Stage1 - Input molecules definition

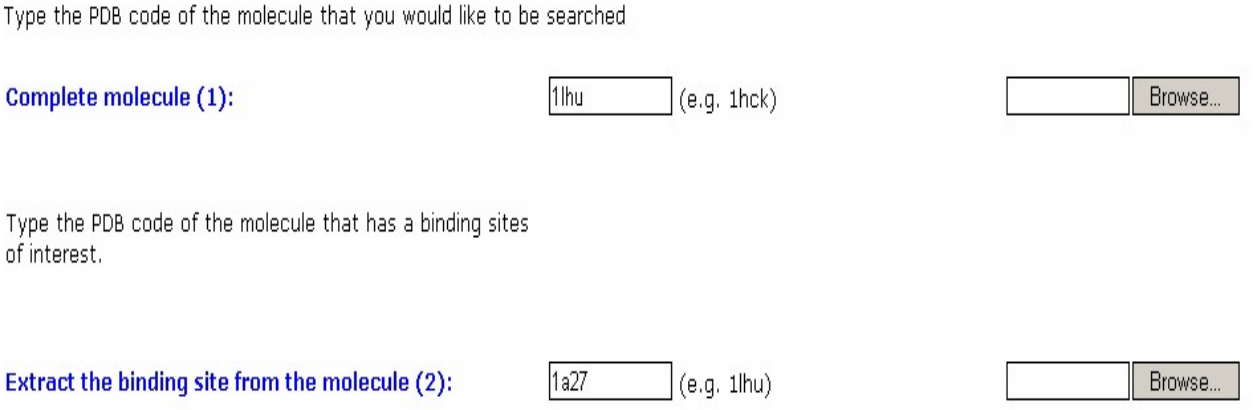
Stage2 - Selection of the chain and the binding site of interest

Stage3 - Process of SiteEngine

Stage4 - Output of SiteEngine

Stage1:- Input molecules definition:

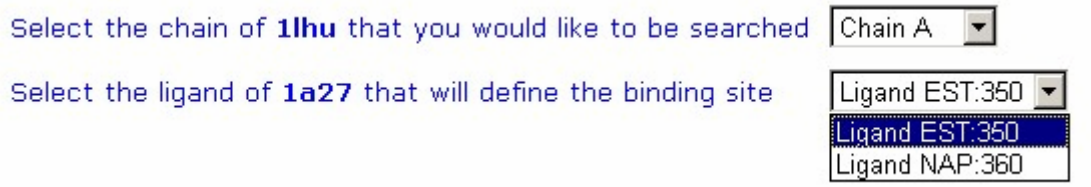
The first stage in activating the SiteEngine is the definition of the molecules of interest. The first field specifies the molecules that will be searched for the binding site of interest. The second field specifies the molecule from which the binding site will be extracted. If the molecules are available in the Protein Data Bank (PDB) the PDB codes are to be specified, otherwise the molecules of interest can be uploaded to our server. Using the PDB codes speeds up the process, since no file transfer is required.



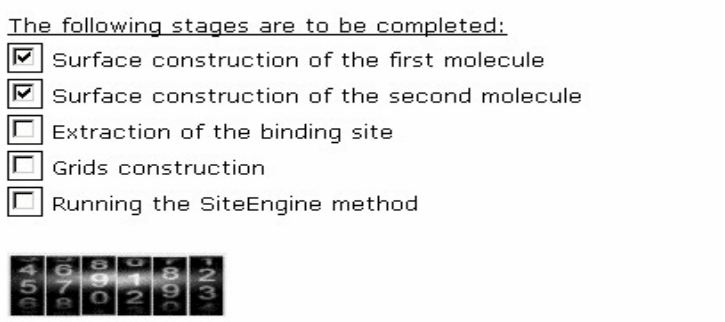
Stage2 - Selection of the chain and the binding site of interest:

First, the user can specify the specific chains of interest that are to be searched in the complete molecule. This restriction will significantly speed up the search process. Second the user must specify the ligand that is present in the binding site of interest. The region of radius 4.0A around the ligand will be extracted and used as the searched pattern. SiteEngine will recognize regions, similar to this pattern, on the surface of the complete molecule.

Below is an example of a molecule (pdb:1a27) that has two ligands EST and NAP. By selecting the ligand of interest the user specifies the regions of interest that his is interested to search in the other molecule.



Stage3 - Process of SiteEngine:
This window shows the process of activation of SiteEngine. The five main stages are presented and those that are complete are checked in the checkbox.
In most of the cases the most time consuming stages are the construction of the surfaces.



Stage 4:-Output of site engine

The 10 top ranking solutions are presented. These represent the 10 regions of the surface of the complete molecule that are recognized to be most similar to the binding site of interest. The file aligned.pdb is the superimposition of the two molecules by the transformation recognized by SiteEngine. It can be either downloaded of viewed directly from the browser. The file contains the superimposition of the two molecules as well as the functional groups that are recognized to be shared by the regions. These are also detailed in the output table. The output table of SiteEngine presents the details of the common functional groups.

Chain.ID

AminoAcid

Property

Source

Dist

Conserved AA

SiteEngine results for searching EST 350 binding site of **1a27** on the complete surface of **1lhu** (chain A).

Solution : 1 aligned.pdb

Rigid Transformation: -2.8971 0.550244 -1.57847 32.1748 18.7469 17.4889

Similarity Score: 3900.87 rmsd: 0.204804 Match Size: 13

Chain.ID	Amino Acid	Property	Source	Chain.ID	Amino Acid	Property	Source	Dist	Conserved AA
A.40	T	DON	b	.142	S	DAC	b	1.6	
A.40	T	DAC	s	.185	C	ACC	b	1.9	
A.40	T	ALI	s	.143	V	ALI	b	1	
A.41	S	DAC	s	.155	Y	DAC	b	0.53	
A.42	S	DAC	s	.152	N	DON	b	1.5	
A.56	F	PII	s	.218	Y	PII	b	2	
A.58	G	PII	b	.221	H	PII	b	1.9	
A.107	M	ACC	b	.186	G	ACC	b	0.7	
A.107	M	ALI	s	.187	P	ALI	b	1.5	
A.129	G	DON	b	.258	R	DON	b	0.85	
A.130	P	ACC	b	.258	R	ACC	b	1.5	
A.139	M	ALI	s	.279	M	ALI	s	2.3	*
A.171	L	ALI	s	.149	L	ALI	b	1.7	*

Chain.ID:

The protein chain, followed by the identity of the amino acid

AminoAcid:

The one letter amino acid code. However it must be noted that the method is based on the physico-chemical properties and does not consider the identity of the amino acids. These are only displayed for the convenience of analysis.

Property:

The physico-chemical property that is matched by the algorithm. The method is based on a representation of each amino acid of a protein as a set of features that are important for its interaction with other molecules. The abbreviations of these features are:

DON - Hydrogen bond donor

ACC - Hydrogen bond acceptor

DAC - Hydrogen bond donor and acceptor (e.g in histidine)

ALI - Aliphatic Hydrophobic property

PII - Aromatic property (pi contacts)

Source:

This field specifies whether the matched property is contributed by the backbone or the side-chain of the amino acid.

The abbreviations are:

b - feature contributed by the backbone

s - feature contributed by the backbone

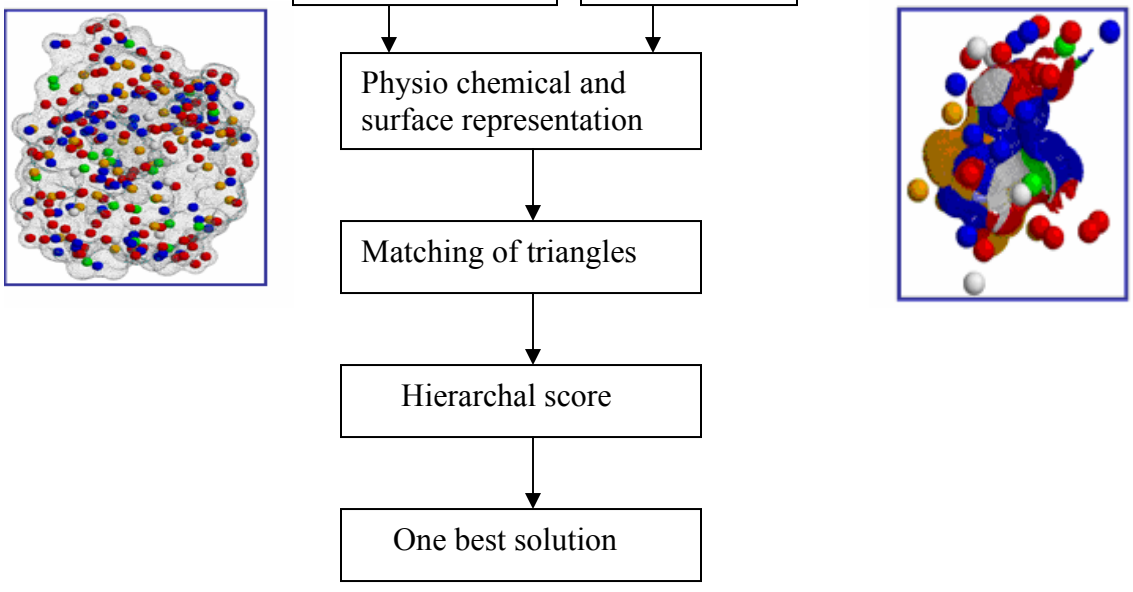
Dist:

The distance in space measured between the matched features.

Conserved AA:

Marks the features shared by the two molecules that are contributed by residues with the same identity of the amino acid.

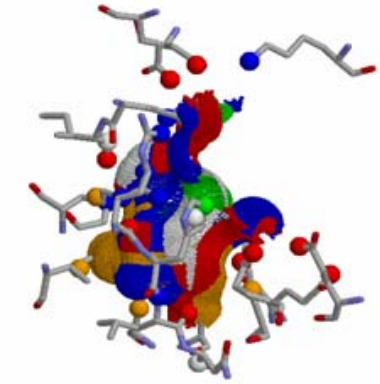
ALGORITHM OVERVIEW



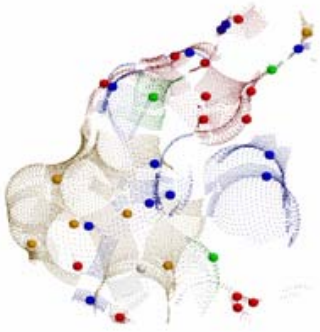
REPRESENTATION

I shows an example of a representation of cavity-flanking residues. In addition, consider the pseudocenters of H-bonding properties of the side-chains of Arg, Lys and His to be positively charged, and those of Asp and Glu to be negatively charged. These modifications lead to a slight improvement in experimental results. From the algorithmic standpoint, the similarity of charges is not a prerequisite for matching and is considered geometric hashing.

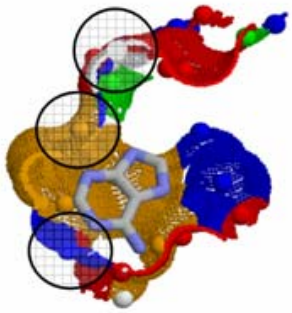
For each chemically labeled surface patch estimate the patch center by a surface point nearest to its center of gravity (Figure II and III). Each patch center is used to estimate the average curvature of its surface patch by calculation of the solid angle shape function. In this calculation, a sphere of a certain radius is placed at the patch center. The average curvature is approximated by the fraction of the sphere inside the solvent excluded volume of the protein. The radius of the sphere determines the region in which the curvature is approximated. The two calculations with different definitions of the radius of the sphere. In the first calculation,consider a minimum radius sphere bounding the surface patch represented by the patch center. In the second calculation, the radius is user defined (by default, 6A° for hydrophobic regions and 3A ° for others). An average of the two values is used to represent the shape of each surface patch.



I. Each amino acid is represented by a set of its physico-chemical properties (Schmitt 02').



II. Each property creates a physico-chemical patch on the protein surface.



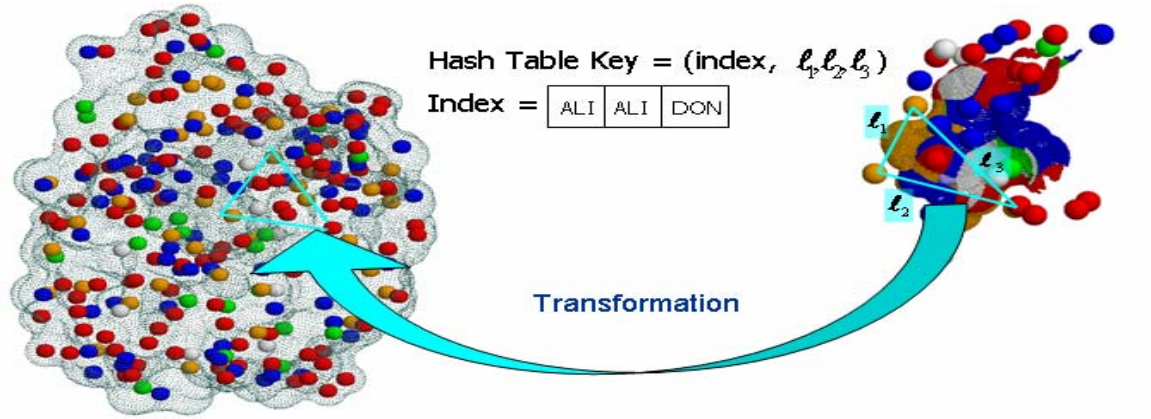
III. The solid angle shape function (Connolly 83') is calculated at the center of each physico-chemical patch.

- HYDROGEN BOND ACEPTOR
- HYDROGEN BOND DONOR
- DONOR/ACCEPTOR
- HYDROPHOBIL ALIPHATIC
- AROMATIC CONTACTS

Matching

At this stage, all possible transformations that will superimpose the input-binding site to a similar region of the surface of the other molecule. The algorithm is based on the matching of almost congruent triangles defined by triplets of pseudo centers

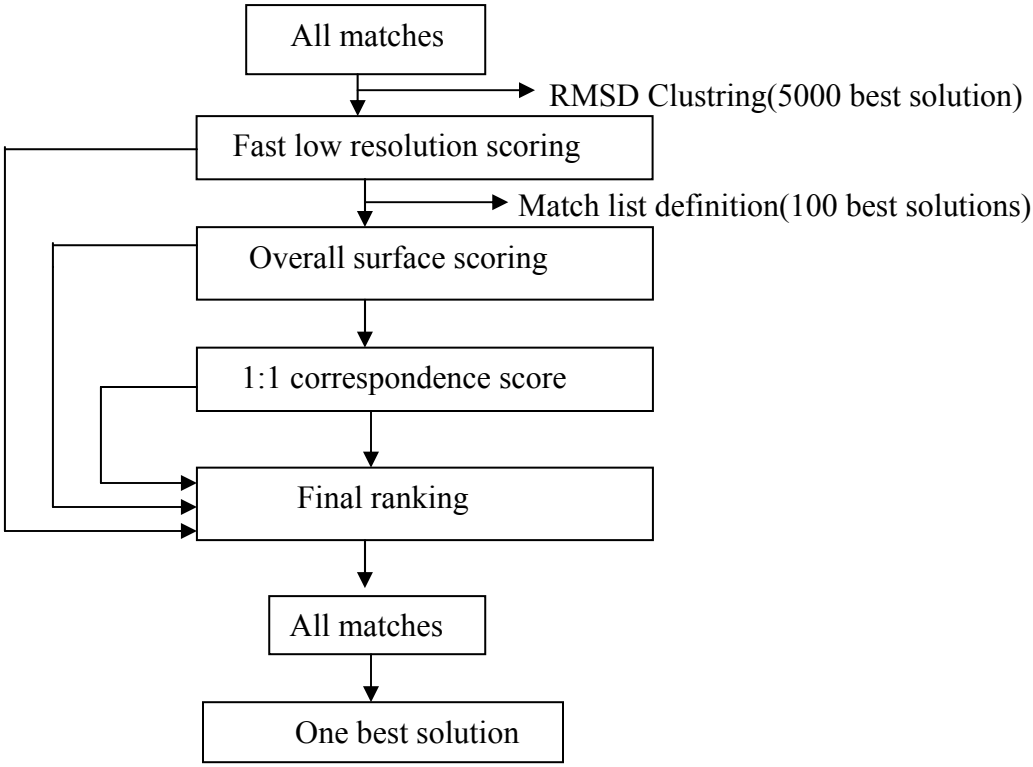
- Each pair of triangle with similar side lengths and similar nodes defines a candidate transformations
- Triangles are efficiently matched by the geometric hashing
- Each candidate transformation is scored



Scoring

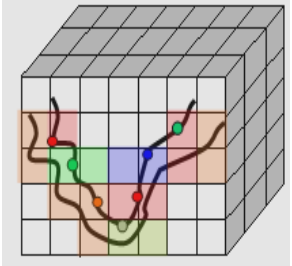
A hierarchical scoring scheme shown in fig.. The first scheme, which is applied to all potential solutions, is calculated based on a low-resolution representation of the

molecules and is therefore highly efficient. As the number of potential solutions is reduced to a smaller subset, the resolution of the molecular representation is increased leading to more precise calculations.



Fast low-resolution scoring:- The goal of this scoring scheme is to provide the initial ranking of candidate transformations and to filter out biologically unreasonable ones. The main idea is to select a small, chemically meaningful representative set of surface points and use them to efficiently estimate the potential surface similarity of the aligned surface patches. Select the points to be a set of patch centers, i.e. centers of physico-chemical surface patches of the input-binding site. The candidate transformation and consider the local environment to which each patch center is transformed.

First, To check whether the given patch centers are transformed to surface regions in the other molecule.



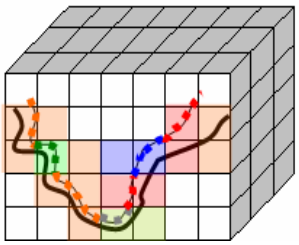
Second, To check whether the physico-chemical environment to which it is transformed is similar to the one in the original molecule.

Third, To compare the shape of the region to which it is transformed with the shape measured at the given patch center. Similarity in each of these attributes will increase the calculated score.

It is found that sufficient to consider only the 5000 highest-ranking solutions. Transformations which superimpose the pseudocenters of the input-binding site so that the root-mean-square deviation (RMSD) between them is lower than a predefined threshold (3A°), are considered to belong to the same cluster. For each cluster the best scoring transformation is selected.

Overall Surface Score Calculations:- This scoring scheme is applied to a smaller number of the retained candidate transformations. It can therefore examine them more thoroughly using a higher level of resolution of molecular representation. Each candidate transformation is applied to each surface point. Then, as in the lower solution score, compare the properties of each surface point with the properties of the environment in the other molecule to which this point is transformed. Here too, similarity of both chemical and geometrical properties is scored higher than the similarity of only one of these. Since the number of considered surface points is much higher, they are divided into different categories. The surface points of the input-binding site are divided into three categories according to their distance from the surface of the molecule on which it is superimposed. Each category counts the number of surface points within distance thresholds of 1A ° , 2A ° and 3A°,respectively. In addition, in each category we calculate the number of points with the same physico-chemical property and charge, and add them to the counter of that category. Then calculate the weighted sum of the counters of the three categories. The closer the category is to the surface the higher the weight that it receives

$$\text{Overall Surface Score}(T) = 1/\text{density} \sum_{i=1}^{i=2} (\in -i)\ (|S_i|+|P_i|) + |C_0|$$



Match List (1:1 Correspondence) Definition :- The 1:1 correspondence score as described in for each retained candidate transformation, determine a 1:1 correspondence (match list) between the sets of pseudocenters of the two molecules. The obtained 1:1 correspondence is used for two purposes, to improve each candidate transformation by the least-squares fitting method and to score the similarity of the environments of the corresponding pseudocentres.

The match list is defined by calculating the maximum weight matching in a bipartite graph. The bipartite graph is constructed in the following way.

(1) The nodes of the graph are the pseudocenters of the two molecules.

(2) An edge is added between each pair of pseudocenters that have similar (up to a threshold) spatial locations, physico-chemical properties and shape functions.

(3) Each edge is assigned a weight that represents the similarity between the corresponding pseudocenters together with their local environments. It measures the distance, the charge compatibility of the H-bonding properties and the similarity of the local shapes of hydrophobic aliphatic regions.

The maximum weight match in this graph provides a 1:1 correspondence between subsets of pseudocenters of the two molecules. The obtained match represents a set of pairs of pseudocenters of the two molecules, so that the points of each pair are the most similar in their geometrical and physicochemical properties. At the next stage, calculate the score of the obtained 1:1 correspondence. This score consists of two parts: first, To calculate a score, which estimates the goodness-of-fit between the corresponding pseudocenters of the two molecules. Second, for each pair of centers with hydrophobic aliphatic or aromatic properties perform a more thorough comparison of the corresponding surface patches. There are two factors that consider to be important in this context:

(1) the size of the overlap region between the patches superimposed by the candidate transformation;

(2) the shape of the common overlap region.

Final scoring and ranking :- For each potential solution the final score is the combination of all the scores calculated by the algorithm. When performing extensive database searches it is difficult to consider more than one solution for each comparison.

In these applications, select only one solution with the highest value of the final score that maximizes the similarity with the searched pattern. Ignore the other solutions obtained for the same comparison. However, in other applications the number of output solutions is user defined and can be much larger.

Complexity and running times:- The overall complexity of our algorithm is dominated by the complexity of the matching and low-resolution scoring stage. The worst case theoretical complexity of an algorithm is $O(n^3m^4)$. In practice, this bound is much lower, since there is a limited number of congruent triangles with similar physico-chemical properties. In addition, since only in triangles that represent potential binding patterns, it limit the side lengths of the considered triangles to be within a limited predefined range. Therefore, the practical running times of the method are proportional to $O(nm^2)$.

Experimental Work

A **lipase** is a water-soluble enzyme that catalyzes the hydrolysis of ester bonds in water–insoluble, lipid substrates. Lipases thus comprise a subclass of the esterases.

To find the binding sites, firstly compare the regioselectivity and catalytic sites of different lipases and which metal ion is present and its coordination.

Table 1. shows the coordination ,regioselectivity, name of metal ions present in various lipase and their catalytic sites from XRD and from SiteEngine.

S. No.	Name	Metal ion	Coordination	Regioselectivity	Catalytic sites (from XRD)	Catalytic sites from site engine
1.	*Pseudomonas glumae lipase*[14]	Ca^{2+}	$Ca^{2+}$is bonded with O-valine295, Asp241, O- Asp287, and O-Gln291and 2H_2Omolecules	Non-specific	Asp263, Ser87, His285	Arg94, Ala97, Ser181, Leu184
2.	*Burkholderia cepacia*[15]	Ca^{2+}	Ca^{2+} is bonded with 2 carboxylate group of Asp242 and Asp288, 2 carboxylate group of Gln292andVal296 and 2 H_2O molecules	Non-specific	Ser82, Asp264, His286	Leu17, Ser82, Ala217, Asp264, Val258, His286
3.	*Pseudomonas aeruginos*[16]	Ca^{2+}	Ca^{2+} bonded with 2 carboxylate grop of Asp209,andAsp253,carbonyl O- of Gln257and Leu261and 2 H_2O maolecules	Non-specific	Ser82, Asp229, His 251	Met16, Leu17, His81, Ser82, Leu231, His251, Leu252
4.	*Chromobacterium viscosium*[17]	Ca^{2+}	Ca^{2+} bonded with 4 oxygen of protein, 2 molecules of water.	Sn-1,3	Ser87, Asp263, His285	Thr129, Ser135, Asp157, Gln158, Leu161
5.	*Bacillus Stearothermophile*[18]	Zn^{2+}	Zn^{2+} is bonded by 2 Histidine and 2 Aspartate acid residues.	Sn-1,3	Ser, His, Asp	Val33, Ala186, Val196, Pro283, Glu284, Ala292, Ile293

6.	*Candida rugose*[19]	Glyce -rol	_____	Non-specific	Ser209, Glu341, His449	Asn16, Ala17, Ile19, His151, Tyr458
7.	*Candida antartica*[20]	____	_____	sn-1,3	Ser, His, Asp	Thr40, Leu140, Ala141, Ala276, Val285, Gly288
8.	*Human pancreatic lipase*[21]	Ca^{2+}	_____	sn-1,3	Phe196, Leu172, Asp195, His282	Leu42, Thr132, Glu197, Glu201, Asp265
9.	*Human gastric lipase*[22]	NAG	NAG is coordinated with Asn15,Asn80,gln252, gln308	Sn-3(acid stable)	Ser153, His353, Asp324	Lys189, Leu191, Gly226, Lys323, His353
11.	*Human Bile salt-stimulated lipase*[23]	—	_____	Non specific	Gly107, Ala108, Ser194, Ala195, Asp320, His435	Arg106, Gly107, Leu110, Ile323, His435
12.	*Dog gastric lipase*[24]	___	_____	Sn-1,3	Leu67, Ser153, Gln154, Asp333 His353.	Leu67, Leu68, His152, Ser153, His262, Asn263, Leu326, His353
13.	*Geotrichium candidium lipase*[25]	NAG NDG	_____	Non-specific	Ser217, Glu334, His463, Ala132, Ala218	Gly130, Ala132, Leu467, Leu306, Tyr135, Ser217
14.	*Bacillus subtilis* [26]	glyce rol	_____	Sn-1,3	Ile12, Ser77,	Ile12, His76,

	lipase				Met78, Asp133, His156	Met78, Ala105, Arg107, Leu108, Val136, Met137
15	*Rat pancreatic lipase*[27]	—	——	Non-specific	Phe77, Leu153, Asp176, His263	Phe77, Ile210, Leu213, Ile248, Phe258
16	*Humicola lanigunosa lipase*[28]	—	———	Sn-1,3	Ser, His, Asn	Ser83, Asn92, Leu93, Val203, Pro207
17	*RP2 lipase from guinea pig*[29]	Ca^{2+}	———	Sn-1,3	Phe77, Leu153, Asp176, His263	Phe77, Ser212, Leu153, Gly236, His263, Thr239
18	*Rhizopus niveus lipase*[30]	——	———	Non-specific	Ser, Asn, His	Tyr28, Thr83, Asn84, Phe86, Ile90, Phe95, Ala110
19	*Candida cylindracia lipase*[31]	——	———	Non-specific	Ser209, Glu341, His449	Met82, Gln83, Lys85, Pro92, Gln93, Ile127
20	*Rhizomucor miehei lipase*[32]	——	———	——	Ser, His, Asp	Arg178, Ile204, His207, Pro210, Val249, Thr252
21	*Rhizopus oryzae*	——	———	Sn-1,3	Ser, His, Gln	His152, Ile212, Pro235,

	lipase[33]					Leu308, Thr353
22	*Serratia marcescens lipase*[34]	Ca^{2+}	———	Non-specific	Ser207, His314, Glu321	Thr143, Leu150, Ile155, His206, Ser207, His314, Ile308
23	*Geobacillus thermoc-Atenulatus liapse*[35]	Ca^{2+} Zn^{2+} MPD EGC	———	Non-specific	Phe17, His113, Ser258	Phe17, Thr18, Trp20, His113, Ser114, Val188, Ile320, His359
24	*Geobacillus zalihae lipase*[36]	Ca^{2+} Cl^{-} Zn^{2+}	———	sn-1,3	Phe, Ser, His	Asp76, Lys138, His140, Val142, Leu144

Output of catalytic sites from SiteEngine of various lipase:-

Catalytic site of *pseudomonas glumae* lipase

SiteEngine results for searching **POT** 612 binding site of **2nw6** on the complete surface of **1qge** (chain D).
Save all results: all_results.zip

Solution : 1 aligned.pdb

Rigid Transformation: -2.26349 0.390893 -2.39528 42.6757 41.1845 60.3591

Similarity Score: 3179.45 rmsd: 0.0465488 Match Size: 7

Chain.ID	Amino Acid	Property	Source	Chain.ID	Amino Acid	Property	Source	Dist.	Conserved AA
D.94	Arg	ALI	s	A.266	Val	ALI	b	0.57	
D.97	Ala	ALI	s	A.267	Val	ALI	b	0.76	
D.181	Ser	ACC	b	A.17	Leu	ACC	s	1.4	
D.182	Ala	DON	b	A.17	Leu	DON	b	1.6	
D.184	Leu	ACC	b	A.18	Thr	DAC	b	1.8	
D.197	Glu	ACC	s	A.87	Ser	DAC	b	1.6	
D.206	Leu	ALI	s	A.167	Leu	ALI	s	1.1	*

Catalytic site of *Burkholderia Cepacia* from SiteEngine

SiteEngine results for searching **OCP** 400 binding site of **5lip** on the complete surface of **2nw6** (chain A).
Save all results: all_results.zip

Solution : 1 aligned.pdb

Rigid Transformation: 3.13837 0.00239938 3.13927 22.8029 24.8032 35.9591

Similarity Score: 15321.5 rmsd: 0.00209079 Match Size: 36

Chain.ID	Amino Acid	Property	Source	Chain.ID	Amino Acid	Property	Source	Dist.	Conserved AA
A.17	Leu	DON	b	A.17	Leu	DON	s	0.27	*
A.17	Leu	ACC	b	A.17	Leu	ACC	s	0.2	*
A.17	Leu	ALI	s	A.17	Leu	ALI	b	0.22	*
A.18	Thr	DAC	s	A.18	Thr	DAC	b	0.16	*
A.18	Thr	ALI	s	A.18	Thr	ALI	b	0.24	*
A.23	Tyr	DAC	s	A.23	Tyr	DAC	s	0.26	*
A.23	Tyr	PII	s	A.23	Tyr	PII	s	0.31	*
A.24	Ala	ALI	s	A.24	Ala	ALI	s	0.12	*
A.27	Leu	ALI	s	A.27	Leu	ALI	b	0.67	*
A.29	Tyr	DAC	s	A.29	Tyr	DAC	b	0.43	*
A.29	Tyr	PII	s	A.29	Tyr	PII	b	0.24	*
A.52	Phe	PII	s	A.52	Phe	PII	b	0.29	*
A.86	His	PII	s	A.86	His	PII	b	0.19	*
A.87	Ser	DAC	s	A.87	Ser	DAC	b	0.39	*
A.113	Pro	ALI	s	A.113	Pro	ALI	b	0.17	*
A.119	Phe	PII	s	A.119	Phe	PII	b	0.24	*
A.120	Ala	ALI	s	A.120	Ala	ALI	b	0.13	*
A.123	Val	ALI	s	A.123	Val	ALI	b	0.27	*
A.143	Val	ALI	s	A.143	Val	ALI	b	0.4	*
A.146	Phe	PII	s	A.146	Phe	PII	b	0.38	*
A.150	Thr	ALI	s	A.150	Thr	ALI	s	0.46	*
A.160	Ala	ACC	b	A.160	Ala	ACC	b	0.28	*
A.164	Leu	ALI	s	A.164	Leu	ALI	s	0.49	*
A.167	Leu	ALI	s	A.167	Leu	ALI	s	0.28	*
A.243	Pro	ACC	b	A.243	Pro	ACC	b	0.36	*
A.246	Leu	ACC	b	A.246	Leu	ACC	b	0.34	*
A.247	Ala	ACC	b	A.247	Ala	ACC	b	0.14	*
A.247	Ala	ALI	s	A.247	Ala	ALI	b	0.26	*
A.248	Leu	ALI	s	A.248	Leu	ALI	b	0.19	*
A.250	Gly	ACC	b	A.250	Gly	ACC	b	0.17	*
A.251	Thr	ALI	s	A.251	Thr	ALI	b	0.36	*
A.266	Val	ALI	s	A.266	Val	ALI	b	0.35	*
A.267	Val	ALI	s	A.267	Val	ALI	b	0.094	*
A.286	His	PII	s	A.286	His	PII	s	0.061	*
A.287	Leu	ALI	s	A.287	Leu	ALI	s	0.2	*
A.293	Leu	ALI	s	A.293	Leu	ALI	s	0.46	*

Catalytic site of *Porcine pancreatic procolipase* whose structure is determined by NMR

SiteEngine results for searching **OCP** 400 binding site of **5lip** on the complete surface of **1pco** (chain A).
Save all results: all_results.zip

Solution : 1 aligned.pdb

Rigid Transformation: 0.369485 0.00366248 -2.51068 1.80946 5.53818 -4.87394

Similarity Score: 4444.55 rmsd: 0.265534 Match Size: 9

Chain.ID	Amino Acid	Property	Source	Chain.ID	Amino Acid	Property	Source	Dist.	Conserved AA
A.8	Ile	ALI	s	A.164	Leu	ALI	b	1.5	
A.11	Leu	ALI	s	A.251	Thr	ALI	s	1.7	
A.14	Gly	ACC	b	A.243	Pro	ACC	s	2	
A.15	Glu	ACC	s	A.247	Ala	ACC	s	1	
A.16	Leu	ALI	s	A.287	Leu	ALI	s	0.63	*
A.18	Leu	ACC	b	A.17	Leu	ACC	b	0.96	*
A.31	Asp	ACC	b	A.23	Tyr	ACC	b	1.8	
A.36	Leu	ALI	s	A.27	Leu	ALI	b	1.9	*
A.37	Ser	DAC	s	A.23	Tyr	DAC	b	0.76	

Results an Discussion

In SiteEngine, the binding sites calculates by these are almost the same as that of given by XRD in pdb with few exceptions i.e. *Human pancreatic* lipase, *Candida antartica* lipase. The active site from XRD of *Human pancreatic lipase* are Phe, Leu, Asp, His but comes from SiteEngine are Ile, Arg, Asn, Glu, Ser etc.

The amino acids comes from SiteEngine in *Dog gastric* lipase, *Pseudomonas cepacia* lipase, and *Pseudomonas aeruginose* lipase are exactly the same as that of given in XRD. From SiteEngine we come to know about whether amino acids are conserved or not. From site engine, active sites of lipase whose structure is known with NMR can also be calculated. But to check whether that sites are accurate or not we have to do docking. but due to lack of time I was not able to perform.

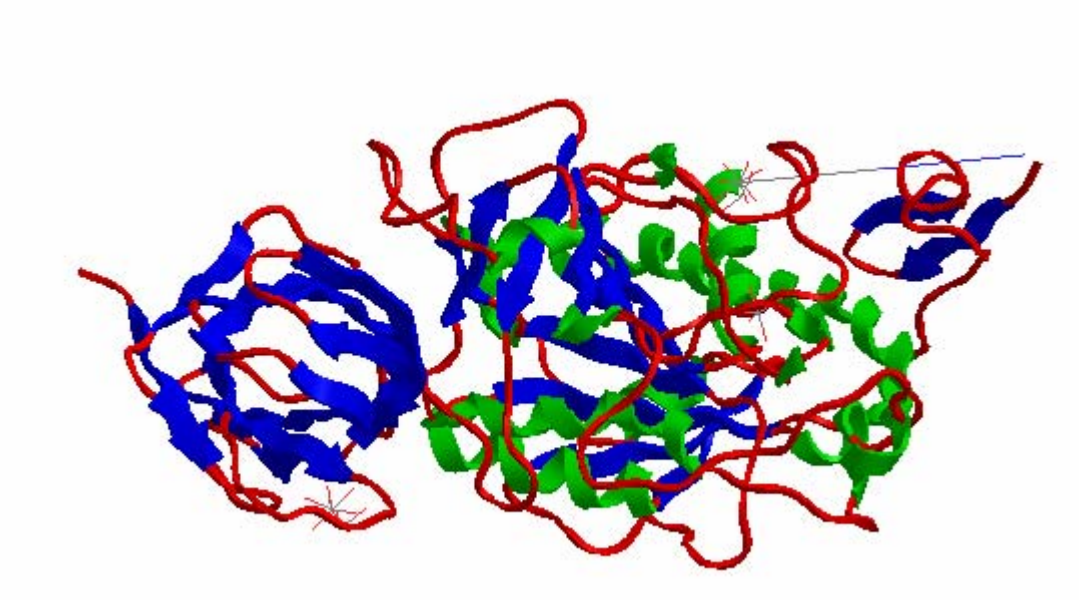


Fig III: Comparison of structure of lipase solved by XRD and given by SiteEngine

Conclusion

Recognition of functional sites in protein structures is extremely important for various biological applications, such as prediction of function and ligand binding. SiteEngine search large protein surfaces to recognize such sites and make predictions in seconds. One of the main advantages of the method is its speed, which is obtained due to the following factors:

(1) Introduction of a low-resolution surface representation via chemically important surface points;

(2) hashing and matching triangles of physico-chemical properties;

(3) application of hierarchical scoring schemes for a thorough exploration of global and local similarities.

However, SiteEngine is a software tool and therefore is limited in the quality of its biological predictions. It recognizes geometrically and chemically similar regions that belong to totally unrelated proteins. However, these similarities do not necessarily imply similarity in the binding partners and in the biological functions. SiteEngine can provide a list of proteins that are most likely to behave similarly to a binding site of interest.

Weaknesses of the method are the requirement of high-resolution protein structures and addressing protein molecules as rigid bodies. Protein flexibility is addressed only through a set of thresholds that allow a certain variability in the locations. These are definitely insufficient for efficient searches of binding sites that can bind large flexible molecules. Other limitations that influence the quality of the results are implied by the screening applications and are general to the problem. One is the absence of a clear definition of what exactly is a functional site and what are the features that define it. When the binding site is defined by its contacts with the smaller ligand, a significant amount of information may be missed. As a result, essential features might be ignored and the extracted pattern might be partially aligned to other functionally different binding sites. There is no simple automatic solution to this problem. One possibility is the construction of a database of consensus binding patterns, common to all proteins with the same function. Another problem is assessing the statistical significance of the obtained results. These are strongly influenced by the number of functional sites of the same type present in the searched database.

Refrences

1. S.P.Alexandra , R.Nussinov and H.J.Wolfson, *J. Mol. Biol.*, 2004, **339**, 607–633
2. S.P.Alexandra, R.Nussinov and H.J.Wolfson, *Nucleic Acids Research*, 2005, **33**, 337-341
3. S.Schmitt, D.Kuhn, and G.Klebe, *J. Mol. Biol.*,2002, **323**, 387–406
4. M.L.Connolly, *J. Appl. Crystallogr*, 1983 , **16**, 548–558
5. P.J.Artymiuk, A.R.Poirrette, H.M.Grindley, D.W.Rice, and P.Willett, *J. Mol. Biol,* 1994, **243**, 327–344
6. J.A.Barker, and J.M.Thornton, *Bioinformatics,* 2003, **19**, 1644–1649
7. T.Binkowski, L.Adamian, and J.Liang, *J. Mol.Biol.*, 2003, **232**, 505–526
8. T.Binkowski, S.Naghibzadeh, and J,Liang, *Nucl. Acids Res.,* 2003, **31**, 3352–3355.
9. M.Rosen, S. Lin, H.J. Wolfson, and R.Nussinov, *Protein Eng*., 1998, **11**, 263–277
10. S.Lin, R.Nussinov, D.Fischer, and H.J. Wolfson, *Proteins: Struct.Funct.Genet*.,1994, **18**,94-101
11. S. Lin, and R. Nussinov*, J. Mol. Graph.,* 1998, **14**, 78–90
12. K.Kinoshita, J.Furui, and H. Nakamura*, J. Struct. Funct. Genomics,* 2001, **2**, 9–22.
13. K. Kinoshita, and H.Nakamura, *ProteinSci.*, 2003, **12**, 1589–1595
14. M.E.Noble, A.Cleasby, L.N.Johnson, M.R.Egmond, L.G.Frenken, *FEBS Lett.,* 1993, **331**, 123
15. A.Mezzetti, J.D.Schrag, C.S.Cheong, R.J. Kazlauskas, *Chem.Biol*., 2005, **12**, 427-37
16. M.Nardini, D.A.Lang, K.Liebeton, K.E.Jaeger, B.W.Dijkstra, *J.Biol.Chem*., 2000, **275**, 31219-25
17. D.Lang, B.Hofmann, L.Haalck, H.J.Hecht, F.Spener, R,D.Schmid, D.Schomburg, *J.Mol.Biol*., 1996, **259**, 704-17
18. J.D.A.Tyndall, S.Sinchaikul, L.A.Fothergill-Gilmore, P.Taylor, M.D.Walkinshaw, *J.Mol.Biol.,* 2002*,* **323** ,859-69
19. P.Grochulski, F.Bouthillier, R.J.Kazlauskas, A.N. Serreqi, J.D.Schrag, E.Ziomek, M. Cygler, *Biochemistry*, 1994, **33,** 3494-500
20. J.Uppenberg, M.T.Hansen, S.Patkar, T.A.Jones, *Structure,* 1994, **2,** 293-308
21. H. Tilbeurgh, L. Sarda, R.Verger, C.Cambillau, *Nature*, 1992, 359, 159-62

22. A.Roussel, S.Canaan, M.P.Egloff, M.Riviere, L.Dupuis, R.Verger, C.Cambillau, *J.Biol.Chem.,* 1999**, 274**, 16995-17002

23. S.A.Moore, R.L.Kingston, K.M.Loomes, O.Hernell, L.Blackberg, H.M.Baker, *J.Mol.Biol*., 2001, **312**, 511-23

24. A.Roussel, N.Miled, L.Berti-Dupuis, M.Riviere, S.Spinelli, P.Berna, V.Gruber, R. Verger, C.Cambillau, *J.Biol.Chem., 2002***, 277**, 2266-74

25. J.D.Schrag, M.Cygler, *J.Mol.Biol*., 1993, **230**, 575-91

26. K.Kawasaki, H. Kondo, M.Suzuki, S.Ohgiya, S.Tsuda*, Acta Crystallogr.*, 2002, **58** 1168-74

27. A.Roussel, Y.Yang, F.Ferrato, R.Verger, C.Cambillau, M.Lowe, *J.Biol.Chem*.,1998, **273 ,** 32121-8

28. A.M.Brzozowski, H.Savage, C.S.Verma, J.P.Turkenburg, D.M.Lawson, A. Svendsen, S.Patkar, *Biochemistry*, 2000, **39**, 15071-82

29. C.Withers-Martinez, F.Carriere, R.Verger, D.Bourgeois, C.Cambillau, *Structure,* 1996*,* **4**, 1363-74

30. M.Kohno, J.Funatsu, B.Mikami, W.Kugimiya, T.Matsuo, Y.Morita, *J.Biochem*., 1996, 120, 505-10

31. V.Pletnev, A.Addlagatta, Z.Wawrzak, W.Duax, *Acta Crystallogr.*, 2003, **59**, 50-6

32. A.M.**B**rzozowski, Z.S.Derewenda, E.J.Dodson, G.G.Dodson, J.P.Turkenburg, *Acta Crystallogr.,* 1999, **48,** 307-319

33. U.Derewenda, L.Swenson, Y.Wei, R.Green, P.M.Kobos, R.Joerger, M.J.Haas, Z.S. Derewenda, *J.Lipid Res.* , 1994, **35**, 524-34

34. R.Meier, T.Drepper, V.Svensson, K.E.Jaeger, U.Baumann, *J.Biol.Chem.*, 2007, **282** 31477-83

35. U.Derewenda, L.Swenson, Y.Wei, R.Green, P.M.Kobos, R.Joerger, M.J.Haas, Z.S. Derewenda, *J.Lipid Res.*, 1994, **35**, 524-34,

36. H.Matsumura, T.Yamamoto, T.C.Leow, T.Mori, A.B.Salleh, M.Basri, T.Inoue, Y.Kai, Y., *Proteins* , 2008, **70**, 592-8,