



NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

**SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS**

**POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"**

MASTER THESIS

**Current state-of-the-art of the research conducted in
mapping protein cavities – binding sites of bioactive
compounds, peptides or other proteins**

Katerina Evangelos Dalamara

Supervisor: **Evangelia D. Chrysina**, Senior Researcher at NHRF
Associate Professor, Örebro University Sweden

ATHENS

SEPTEMBER 2017



ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**

**ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**Αποτύπωση των ερευνητικών μελετών που αφορούν στον
χαρακτηρισμό μιας πρωτεϊνικής κοιλότητας – κέντρου
πρόσδεσης βιοδραστικών ενώσεων, πεπτιδίων ή άλλων
πρωτεϊνών**

Κατερίνα Ευάγγελος Δαλαμάρα

Επιβλέπουσα: **Ευαγγελία Δ. Χρυσίνα**, Κύρια Ερευνήτρια, ΕΙΕ
Αναπληρώτρια Καθηγήτρια, Örebro University, Sweden

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2017

MASTER THESIS

Current state-of-the-art of the research conducted in mapping protein cavities - binding sites of bioactive compounds, peptides or other proteins

Katerina E. Dalamara

S.N.: PIV025

SUPERVISOR: **Evangelia D. Chrysina**, Senior Researcher at NHRF
Associate Professor, Örebro University Sweden

**EXAMINATION
COMMITTEE:**

Elias S. Manolakos, Professor,
Dept. of Informatics, Univ. of Athens
Ioannis Z. Emiris, Professor,
Dept. of Informatics, Univ. of Athens
Evangelia D. Chrysina, Senior Researcher NHRF,
Assoc. Prof. Örebro University

September 2017

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αποτύπωση των ερευνητικών μελετών που αφορούν στον χαρακτηρισμό μιας πρωτεϊνικής κοιλότητας - κέντρου πρόσδεσης βιοδραστικών ενώσεων, πεπτιδίων ή άλλων πρωτεϊνών

Κατερίνα Ε. Δαλαμάρα

A.M.: ΠΙΒ025

ΕΠΙΒΛΕΠΟΥΣΑ: Ευαγγελία Δ. Χρυσίνα, Αναπληρώτρια Καθηγήτρια

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Ηλίας Σ. Μανωλάκος**, Καθηγητής,
Τμ. Πληροφορικής & Τηλ/νιών, ΕΚΠΑ
Ιωάννης Ζ. Εμίρης, Καθηγητής,
Τμ. Πληροφορικής & Τηλ/νιών, ΕΚΠΑ
Ευαγγελία Δ. Χρυσίνα, Κύρια Ερευνήτρια, ΕΙΕ,
Αν. Καθ. Örebro University

Σεπτέμβριος 2017

ABSTRACT

The aim of this thesis was to report on the current state-of-the-art of the research conducted concerning mapping of protein cavities with a potential function role as binding sites of bioactive compounds, peptides or other proteins.

A literature review was performed with emphasis on the relevant tools developed during the last decade. In addition, the main research findings regarding drug design and druggable targets based on binding sites are presented.

Processes performed in protein cavity detection and analysis, of previous research articles, are compared with the approach described by Anaxagoras Fotopoulos and Athanasios Papathanasiou (2015). The results showed that a competitive advantage of their approach is the multidimensional k-means algorithm for clustering.

For the bibliographic review the scientific knowledgebase has been used, which includes international articles and journals, book chapters, as well as online articles regarding drug design and protein cavity.

Search keywords such as protein cavity dynamics, catalytic sites of enzymes, protein pocket etc. were used to identify bioinformatics tools with text mining. A catalogue of the most recently developed tools is presented followed by a brief description of selected tools. The selection criteria imposed for preparing the catalogue and the detailed description included the publication date, as well as the algorithms and the methods they use. The tools were then classified according to the search keywords.

The findings of this research are discussed, and the algorithms and methods they use are compared, highlighting the advantages of protein cavity detection.

SUBJECT AREA: Protein Structure

KEYWORDS: protein cavity dynamics, catalytic sites of enzymes, ligand binding, in silico docking, molecular dynamic simulations, active sites, protein pocket, protein conformer, rotamers

ΠΕΡΙΛΗΨΗ

Ο σκοπός της διπλωματικής εργασίας είναι η διερεύνηση και αποτύπωση των ερευνητικών μελετών που αφορούν στον χαρακτηρισμό μιας πρωτεϊνικής κοιλότητας – κέντρου πρόσδεσης βιοδραστικών ενώσεων, πεπτιδίων ή άλλων πρωτεϊνών.

Στην παρούσα εργασία χρησιμοποιήθηκε η μέθοδος της βιβλιογραφικής επισκόπησης.

Παρουσιάζονται τα κυριότερα ευρήματα προηγούμενων ερευνών που σχετίζονται με τη διαδικασία σχεδιασμού φαρμάκων και τον εντοπισμό φαρμακοφόρων με βάση ένα σύνολο προσδετών.

Στη συνέχεια συγκρίνονται διαδικασίες επεξεργασίας και ανάλυσης της πρωτεϊνικής κοιλότητας προγενέστερων ερευνών με τη προσέγγιση που προτάθηκε από τους Παπαθανασίου και Φωτόπουλου το 2015. Αναδεικνύονται βασικά πλεονεκτήματα της προσέγγισης αυτής, όπως η εφαρμογή του αλγορίθμου πολυδιάστατη k-means ομαδοποίηση (multidimensional k-means clustering).

Η εύρεση βιβλιογραφίας βασίστηκε σε αναζήτηση επιστημονικών άρθρων σε ξενόγλωσσα επιστημονικά περιοδικά, σε κεφάλαια βιβλίων και σε διάφορα άρθρα σε ηλεκτρονικούς ιστότοπους σχετικά με τον σχεδιασμό φαρμάκων και τις κοιλότητες που απαντώνται στις πρωτεΐνες.

Στην παρούσα εργασία παρουσιάζονται εν συντομία εργαλεία που εντοπίστηκαν χρησιμοποιώντας λέξεις κλειδιά όπως για παράδειγμα δυναμική πρωτεϊνικής κοιλότητας, καταλυτικό κέντρο ενός ενζύμου, πρόσδεση, πρωτεϊνική θήκη κλπ. Στη συνέχεια συγκροτήθηκε κατάλογος με τα εργαλεία βιοπληροφορικής ανάλυσης που βρέθηκαν και ακολούθησε εκτενής αναφορά επιλεκτικά σε κάποια από αυτά. Κριτήριο επιλογής αυτών των εργαλείων αποτέλεσε η ημερομηνία δημοσίευσής τους, οι αλγόριθμοι και η μεθοδολογία που χρησιμοποιούν. Τα εργαλεία αυτά κατηγοριοποιήθηκαν με βάση τις λέξεις κλειδιά που χρησιμοποιήθηκαν για την εξόρυξη των δεδομένων από την βιβλιογραφία. Τέλος πραγματοποιήθηκε συγκριτική μελέτη αυτών αναδεικνύοντας τα πλεονεκτήματα και εστιάζοντας στην περαιτέρω αξιοποίησή τους.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Πρωτεϊνική Δομή

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: δυναμική πρωτεϊνικής κοιλότητας, καταλυτικό κέντρο ενός ενζύμου, πρόσδεση προσδέτη, προσομοίωση πρόσδεσης, προσομοίωση μοριακής δυναμικής, κατηγοριοποίηση ενεργών περιοχών ενζυμικής κοιλότητας, πρωτεϊνική θήκη, πρωτεϊνικές διαμορφώσεις, διαμορφωμερή πλευρικής αλυσίδας

ACKNOWLEDGMENTS

First and foremost I would like to thank Professor Elias Manolakos for giving me a second chance to work on my Master. His believe in my capabilities and his continuous support kept me going. I am grateful for the opportunity given to me to study and learn.

I would like to express my gratitude to Professor Evangelia Chrysina, who was my thesis advisor, for her trust and continuous support while I was writing my thesis. She has been a valuable asset.

Furthermore I would like to thank Mr. Fotopoulos and Mr. Papathanasiou for guiding me through my first steps in this thesis.

Lastly, I would like to thank my family for providing me with unfailing support and encouragement.

CONTENTS

PROLOGUE	11
1. INTRODUCTION	12
1.1.1 Drug Design	13
2. CURRENT STATE OF THE ART	15
3. PREVIOUS RESEARCH	20
4. BIOINFORMATICS TOOLS	21
4.1.1 Protein cavity dynamics	24
4.1.1.1 TRAPP Transient Pockets in Proteins	27
4.1.1.2 PockDrug-Predict Pocket Druggability	29
4.1.1.3 EDTSurf	31
4.1.2 Catalytic sites of enzyme / Classify active sites	32
4.1.2.1 HotSpot Wizard	35
4.1.2.2 RRDistMaps	36
4.1.3 Protein pocket	37
4.1.3.1 PRNAClass	41
4.1.3.2 FlexPred	41
4.1.3.3 locPREFMD	42
4.1.3.4 MSMBuild	43
4.1.4 In silico docking	45
4.1.4.1 VirtualToxLab	47
4.1.5 Molecular dynamics simulations	48
4.1.6 Protein conformers / Side chain angle	50
5. CONCLUSIONS	53
APPENDIX I	54
DRUG DEVELOPMENT & PROTEIN STRUCTURE	54
REFERENCES	58

LIST OF FIGURES

Figure 1: TRAPP Workflow.....	28
Figure 2: PockDrug Workflow	30
Figure 3: HotSpot Workflow.....	35
Figure 4: PRNAclass Process	41
Figure 5: locPREFMD Flow	42
Figure 6: MSMBuilder process steps.....	43
Figure 7: molecule bonds	55
Figure 8: different types of isomers	55
Figure 9: protein structure	56
Figure 10: amino acid.....	56
Figure 11: protein rotation.....	57
Figure 12: secondary structure	57
Figure 13: tertiary structure.....	57

LIST OF TABLES

Table 1: Bioinformatics Tools	22
Table 2: Protein cavity dynamics	24
Table 3: Catalytic sites of enzyme	32
Table 4: Protein pocket.....	37
Table 5: In silico docking	45
Table 6: Molecular dynamics simulations	48
Table 7: Protein conformers	50

PROLOGUE

A drug is a substance or a product with attributes that help to relief from pain, heal diseases or improve health. Through its action, it may have an impact on how the function of an organism or its individual components.

Drugs affect biochemical interactions in living species. More specifically, they affect the molecular interactions at protein's binding sites or active sites if it is for enzymes. Since each living organism is different, drugs may not influence the target functionality serving the purpose that they are designed for and cause side effects. Therefore it is necessary to evaluate a target's druggability by finding regions that can be binding sites with increased specificity.

The scope of Master Thesis of Anaxagoras Fotopoulos and Athanasios Papathanasiou was to find structural determinants that compose binding by mapping cavity topology and designing ligand. Attempts were made for better docking methods based on active site conformations analysis and a comprehensive range of bioinformatics tools were used. [1]

A validation study is recommended to verify the results, code review to fix bugs and improve program performance.

The current thesis focuses on finding the current state of the art of the research conducted in mapping protein cavities, specifically of catalytic sites of enzymes. Emphasis will be given on the various tools and methods available for drug design described in the literature and also in A. Fotopoulos & A. Papathanasiou MSc thesis. Suggestions and recommendations are provided.

1. INTRODUCTION

The process to design and develop effective drugs performing basic research can be costly and time consuming, especially because it involves a number of steps. To improve this process, a broad range of computer applications have been developed concerning data analysis, storage and management.

Different databases keep information such as genome and protein sequences, peptides, protein and peptide interactions, ligand associations, biomolecular structures etc. that can be used as input in drug design analysis. Besides genetics data, examples and results are available as well.

Mathematical methods and algorithms are applied for molecular geometry and topology analysis. Molecular geometry is focused on distances and angles whereas molecular topology is focused on connectedness. Distances, bonds and angles are quantities that are changing with time and determine the molecule flexibility. Topology refers to changes of a topological object and mapping on other topological objects. Tunnel and cavity preserve transformations in an object. Different objects can be identified from geometrical properties. [2]


Computer methods based on statistics compare and classify the above mentioned structures. [3]

In the current thesis, **Chapter 2** includes the current state of the art of the research conducted in mapping protein cavities, specifically of catalytic sites of enzymes.

Chapter 3 is a short description of the tools and algorithms flow used previously with emphasis on Anaxagoras Fotopoulos and Athanasios Papathanasiou Master thesis. [1]

Chapter 4 refers to tools that were identified to combine protein cavity dynamics, catalytic sites of an enzyme, *in silico docking*, molecular dynamics simulations and protein conformers. Emphasis is given on the tools available in the literature as mentioned in recent published papers that indicate the current state of the art, are compared to the algorithms and methods previously used Fotopoulos and Papathanasiou. [1]

Notes for Chapter 4:

1. The symbol  that is used next to some tools, represents the fact that more information follows.
2. The color **Green** symbolizes the comparison made and points worth to be mentioned.

Based on the analysis we can determine the novelty of the approach followed and its relationship to previously published tools.

Finally in **Appendix** section basic information regarding Drug Design and Protein Structure is described.

1.1.1 Drug Design

The goal of drug discovery is to improve health and prevent diseases and contribute towards health management challenges. Scientists are looking for molecules that modify the disease features find at least treatments of a disease if not a cure. For this purpose, a molecule editor, a graphical definition of molecules, is used in all process systems. [4]

Scientists are searching for compounds that bind to protein targets involved in a disease and alter their function. Those bioactive molecules combinations must be non toxic, must have increased specificity for the target and be effective. Potential drug targets are RNA molecules that block or enable a function, small peptides or small molecules/organic compounds. [5]

Note that natural products have also pharmacological or biological effects that can be of therapeutic value in treating diseases. They are promising for finding applications in medicinal chemistry, molecular biology and pharmaceutical sciences. [6]

Since drugs may cause unpleasant or risky situations, known as side effects, tests are required before any drug is released to the market. The tests are usually performed on animals, however for safety reasons drug testing on human volunteers is performed as well. Overall, there is pressure to develop other solutions that will unveil any possible drug side effect. Computational testing and *in silico* models meet this need. [7]

Importance of Active Site

In the drug discovery process, employing enzymes as protein targets, the identification of active sites is essential. An active site is a region of an enzyme where substrate molecules bind and chemical reaction occurs. This means that residues, named binding site, interact with the substrate, and catalyze a reaction. The region where the reaction takes place is called catalytic site. [8] Ligand transport from and to protein active sites is very important at the beginning and at the end of enzyme catalysis. [9]

The prediction of ligand binding and catalytic sites can be used to design small molecules. The pocket identification process finds cavities in protein structures. [10]

Protein x-ray crystallography is a technique that determines the three dimensional structure of a protein structure and reveals cavities that could serve as binding sites. Molecular Dynamics (MD) simulations of ligand affinity provides shape-related models that reveal its binding mode and the interactions in forms. Lennard-Jones potential is a mathematically model that describes these kind of interactions. [9]

Automated Active Site Detection, Docking and Scoring (AADS) is an online tool that computes protein cavities. The tool is based on physicochemical properties. Each pocket is represented by a single cavity point. The cavities are ordered descending by their volumes. The top ten molecules that dock successfully are candidate drug molecules. For this type of simulations Monte Carlo methodology is used. [11]

Mapping of the active site could be achieved, for example, through an enzyme inhibitor. An inhibitor slows down the reaction rate and can affect enzyme functions in pharmacological drug targets upon binding to the active site or other binding sites of the enzyme. [8]

Analysis of the active site prediction methods can be divided in three groups: the sequence homology methods to find conserved residues, the structural homology

methods to find functional regions, and the methods based on geometry and physico-chemical attributes of the sequence and protein structure. [12]

During the last decade, the redesign of active sites is another approach followed for studying the active sites of enzymes. It is specifically based on teaching the “old” enzymes “new” tricks that can lead to new functions. Changes on chemical transformations can affect enzyme activity. Searching for chemical spaces could help increase the drug effectiveness. Defined changes in the atomic structure of an enzyme is a way to access alternative catalytic activities. [13]

The following examples confirm the importance of the aforementioned facts.

Scientists focus on proteins that might have interesting characteristics, for which though nothing much is really known. By analyzing their structure, they want to understand what their function and what they are related to. This is how they find out that a metabolic pathway is involved in stuttering. A computer model predicts the structure of the same region in a human UCE (UnCovering Enzyme). It shows a cavity that appears to be an active site. With this model they created various mutations in UCE to see what effect they had on the enzyme's function and verify that they had identified the enzyme's active site correctly. [14]

Lipocalin prostaglandin D synthase L-PGDS is an enzyme related to the central nervous system. It regulates pain and can be effective in brain tumor patients. By studying its structure and catalysis mechanism researches came to the conclusion that it is released to the environment by interacting with membranes. [15]

Kynurenine formamidase kynB is an enzyme related to brain disorders. Experiments and kynB complex structure showed an ensemble of ZINC catalytic sites. [16]

Carbonic Anhydrase (CA) inhibitors are used for treatment or prevention of diseases. The problem by studying their structure is that sequence and structures of different CA isoforms are very similar. The structure of the enzyme inhibitor complex was analyzed by different Fourier techniques using 1CAZ structure. Inhibitor molecule was observed in a cavity located on the protein surface and stabilized by several polar interactions. [17]

An important role in human diseases plays the protein kinase, hence it has been extensively studies as a drug target. More specifically different mutations of the enzyme affect its regulatory role and it is of great interest to understand. All approved protein kinase antagonists are steady-state competitive enzyme inhibitors and they interact with the ATP-binding pocket. Prototype for all protein kinases used in EGFR protein kinase contain several conserved α -helices and β -strands. (C-spine for catalytic, R-spine for regulatory structural skeleton analysis). Approved EGFR antagonists on the duration of their inhibitory effects on human cells. The duration of drug effectiveness after removing the drug was also checked. [18]

2. CURRENT STATE OF THE ART

In drug design both chemical and biochemical analysis are required. Efficient binding affinity of macromolecular interactions depend on geometry and physicochemical properties. [5]

Importance of cavities

The importance of cavities becomes clear with the variety of examples and case studies listed below:

Researchers are focusing on finding protein templates with correct ligand-binding pockets geometry in order to design molecules that specifically bind to proteins and enzymes. A problem that might occur can be the mutations, which will change the pocket structure. Active sites and ligand-binding cavities are often formed by curved β -sheets. By studying the geometry of β -sheets folding simulations we can control the atomic level accuracy and proceed with designing proteins with cavities formed by curved β -sheets. [19]

Protein-protein and protein-ligand docking are important for understanding disease mechanisms. It is considered a key computational method in the design of starting points for the drug discovery process. Annotation of data and examples of protein-protein interactions (PPI) are kept in online databases. [20]

Inhibition of protein-protein interaction (PPI) is promising to improve the specificity with less side-effects. Computational methods are desired because the drug target space is wide and drug target discovery is difficult. [21]

In the drug discovery analysis and decision-making, an extensive number of bioactivity databases is important. Actions that can be performed are for example to find target pathways, explain side effects or analyze the structure activity relationships (SAR). [22]

Peptide-protein interactions are involved in a wide range of biological processes. They are important for signal transmission and to regulate mechanisms. They do have small interfaces. In the designing process of experimental interactions what is used are candidate sites of interaction at the protein surface. Protein-peptide docking techniques help to understand protein actions in an atomic-level. The increasing number of protein-peptide structures in protein data bank can increase the prediction accuracy of protein-peptide docking. [23] [24]

The prediction of side-chain conformations of residues that form a binding site of a protein can be helpful in protein docking, drug design and virtual screening. Side-chain conformations can be flexible and their flexibility depends on the type of residues but also their environment. They also depend on their backbone torsion angles. By combining side-chain conformations, atom coordinates can be predicted and score calculations performed.

By checking the side chain conformations of important catalytic residues, observed in the crystal structures of various mutants, the scientists came to the conclusion that the distance and orientation conformation is correlated to the backbone amide dynamics. A better understanding of these conformations can help design inhibitors against bacterial Pth proteins. [25]

A maximum-likelihood method has been approached to parameterize the side chain model using input data from high resolution protein crystal structures. The maximum

likelihood parameters assign high probability to the observed rotamer states and no laborious discrete sampling of the angles are required. The method called “Upside” performs dynamics simulations of the backbone trace and structural details are included. This reduces the steric rattling. [26]

Recently studies have suggested that β -lactams may have anti-TB (mycobacterium tuberculosis) activity. How types of β -lactam anti-bacterial act against each other can lead to structural improvements, therefore become treatment drugs for TB. The superposition of these β -lactams are spotted in all conformations at the active site and show that the R1 side chain prove additional interactions with residues inside and outside cavities. [27]

G protein-coupled receptors (GPCRs) are top drug target membrane proteins. They are important in signal transduction. New compounds can be identified by using specific databases that contain annotated GPCR-ligand associations. [28] [29]

In contrast to ligand-based approaches that need an initial set of bioactive compounds, structure-based docking requires only the 3D structure of the protein target. The structure-based molecular docking shows binding sites and affinity of the receptor protein. [5]

Multitask learning between drugs establish the best treatment strategy and drug sensitivity prediction. [30] One drug can increase or decrease the effect of another drug, a drug–drug interaction (DDI). [31] The drug-drug interaction is a major cause of adverse drug effects and a key role in patients' safety. [32] A typical experiment exposes cells to doses of a drug and evaluates the response. Understanding this relationship of therapeutic compounds is very important. [33] Pharmacology-based prediction of multi-targeted drug combinations is a promising approach to improve anticancer efficacy and safety. [34]

Identification of drug-target interactions is extremely costly, time-consuming and challenging. Therefore, computational models have been developed [35] to visualize drug combination effects and analyze the results. [36] Recent studies have generated interest in identifying synergistic combinations for therapy. [37] For binding affinity predictions approaches such as MM/PBSA and LIE, have been used. They do provide prediction accuracy and reduce computational cost. [38]

Scientific research has been focused on developing peptide based therapies to treat various diseases. Peptides have been stored in databases and computation tools have been developed to predict and design cell penetrating peptides.

A sufficient interpretation of the data requires understanding in which are the molecular families, structural motifs properties and structural changes. Sufficient chemical intelligence is required but uncommon. [39]

Interaction of a drug or chemical with a biological system can result in a gene expression connectivity mapping. Connectivity mapping is a process to recognize novel pharmacological and toxicological properties in small molecules by comparing their gene expression signatures with others in a database. [40]

Interesting recent work describes how amino acid positions conserved within a group of orthologues can be distinguished from those conserved in broader family of proteins. [41]

The pharmacokinetic profile of a compound defines its absorption, distribution, metabolism and excretion (ADME) properties, and is crucial for effective drugs. The ADMET properties are considered in early stage drug development. [42]

Cytochrome P450 (CYPs) are the major enzymes involved in drug metabolism and bioactivation. [43] They belong to a family of heme proteins. The sites of metabolism (SOMs), specific atoms of a molecule that are oxidized by specific CYP isozymes, are influential for early-stage lead design and optimization. [44]

European medicines agency (EMA) and US Food and Drug Administration (FDA) main focus is on CYP enzymes. Their analysis is based on QSAR Quantitative structure activity relationships and SOM sites of metabolism. [45] 3D-QSAR analysis and clustering of binding cavities is a computational approach to understand binding affinity and interaction. [27] Size and shape of the binding cavity are critical for selective ligand binding. In docking to find active conformations for substrates two basic assumptions are made: SOM should be close to heme iron and binding energy of substrate should be low. Docking is form-specific, 3D structure of enzyme cavity. Cytochromes are proteins that play important role in energy transfer in cells. They are found in mitochondria. 57 CYP forms are found. Ligand-based, target-based and combined methods give precise information of key features like ligands, binding cavities for metabolizing CYP enzymes.

To trace the effectiveness or side effects of drugs we need to understand metabolism. In drug development, metabolic stability can be predicted very early in the process but *in silico* can predict drug-drug interaction due inhibition of CYP enzyme activity. Enzymes types CYP450 that belong to protein family that include heme are being part of oxidant metabolism of *in vivo* drugs and influence drug effectiveness. If the heme has iron Fe² then cytochrome creates a complex with CO. This plays an important role in drug tension and duration.

New algorithms analysis methods give us information about features affecting CYP enzyme-ligand interactions. OECD QSAR Toolbox program is used to identify relevant structural characteristics and potential toxic mechanism of chemicals attempts to develop general CYP inhibitor docking protocols. It provides accuracy and speed. [45]

Computational prediction of distribution, metabolism, excretion and toxicity (ADMET) properties is an effective method to minimize attrition. Suitable chemical space has been estimated. [46]

Pharmacometric approaches, such as pharmacokinetic, pharmacodynamics, pharmacology, physiology modeling and simulation, are being more and more applied in the drug development process. For this purpose, mathematical models are used. [47]

The octanol-water partition coefficient (logP) is one of the most important physico-chemical parameters for metal-based anticancer drug discovery compounds with improved pharmacokinetic properties. [48]

The molecular lipophilicity potential (MLP) is a method to calculate and visualize lipophilicity on molecules. [49]

Determining the toxicity of chemicals is necessary to identify their harmful effects. [50] StARD3 a StARD related lipid transfer protein, is a carotenoid-binding protein in the primate retina, and adopts the helix-grip fold. An essential characteristic of StARD3 is the tunnel-like cavity. One of the ionone rings must stick out of the cavity. The cavity communicates with bulk solvent through two openings. The binding-pocket asymmetry may identify lutein, thus b-ionone and e-ionone rings. [51]

The ZINC dependent MMPs Matrix Metalloproteinases are linked to pathological conditions such arthritis, cancer, skin ulceration. Studying structural properties of MMP collagen using a combined molecular dynamics umbrella sampling MDUS approach, allows including an additional single polypeptide chain into the active site. GROMACS is

used for molecular simulations and LINCS algorithm to homologically constraint bound interactions. MMPs are promising drug targets. [52]

Open Drug Discovery Toolkit (ODDT) is an open source tool for computer aided drug discovery (CADD). It combines machine learning scoring functions. [53]

Ontology-based enrichment analysis for small molecules uses ontology annotations. They link ontology classes to biological entities. [54]

The γ -secretase enzyme cuts proteins in small pieces. Amyloid precursor protein is targeted by γ -secretase and is linked to brain related diseases such as Alzheimer's. Cryo-electron microscopy (cryo-EM) technique creates 3D models of γ -secretase and this means that the active site is flexible. 3D classifications help analyzing conformational landscape of γ -secretase. For complexes larger than several hundred thousand Daltons, dynamic changes in tertiary and quaternary structure have been studied using by cryo-EM image classification with improved computer algorithms. β -secretase 1 (BACE1) inhibitors are also a new promising target for Alzheimer's disease since they reduce dendritic spine dynamics. [55]

Available tools and algorithms for cavity extraction

Cavity analysis requires to trace Molecular Dynamics (MD) trajectories. Molecular Dynamics (MD) calculations and docking algorithms are compared, classified and combined for more practical and user friendly tools. They do locate binding sites of already known active sites and try to detect other ones unknown. Depending on the context of molecular processes the available tools for cavity analysis, represented below, are classified based on their functionality.

Algorithms to detect cavities are based on Grids and Voronoi diagrams. [56]

The Molecular Dynamics (MD) simulation is a technique which is used to collect transitional conformations, meta-stable states, transition states, helical structures, and stochastic dynamics, of biomolecules such as proteins, and with this sample determine if the clustering algorithms are actually extracting useful information. This can be done by validating the performance of clustering algorithms.

Spectral clustering is an algorithm which can applied to cluster polymer models and MD simulations. Root Mean Square Distance (RMSD) calculates differences between structures and is used to find the top k eigenvectors. [57]

Conformational entropy is a protein property that varies in sequence, secondary structure, and tertiary fold.

The backbone entropy and side chain flexibility can be independent. α -helices have lower entropy than β -sheets. [51]

There are two categories of clustering algorithms, metrics and clusters.

Features like global folds, active sites, and ligand poses contribute in developing a clustering algorithm. The best cluster has the highest probability, which is calculated by a scoring function. A deterministic geometric clustering algorithm clusters all structures until a satisfying classification criterion is met. Analysis and visualization follows in order to discover and compare common structural features. The number of clusters generated by the top-down algorithm are defined by the user. Pairwise Root Mean Square Distance (RMSD) between structures is performed. The procedures terminates when the given criterion is true. [58]

Fast_protein_cluster is a toolkit that uses k-means and hierarchical clustering methods and calculates Root Mean Square Deviation (RMSD) after optimal superposition and Template Modeling score (TM-score). [18] It is faster than Clusco. [59]

MolAxis is used to find pockets and cavities and their geometric characteristics.

EXPOSITE (EXPOsure of active SITes through normal modEs) is an improved structure based technique in order to find active sites based on normal modes where surface changes are accessible. Low-frequency motions are analyzed by using techniques such as Elnemo and GROMACS. In addition EXPOSITE can be used to rank enzyme pockets according to their degree of exposure in normal mode dynamics.

MDpocket (Molecular Dynamic pocket) is an open-source tool to track small molecule binding sites and gas pathways. The method is based on the fpocket cavity detection algorithm. [60] CAVER tool searches paths by the same way.

MOLE algorithm uses Voronoi diagrams to detect channels and tunnels. The set of points in those diagrams are evaluated by their atom distance. The Dijkstra algorithm calculates the “shortest” path. In a similar way the CAST algorithm can detect channels and tunnels by Voronoi diagrams and beta complexes.

HOLE software calculates a possible molecular path by measuring ion conduction.

PocketFinder is a Lennard Jones based tool.

SITEHOOD finds potential ligand binding sites.

LIGSITE maps Solvent Accessible Surface (SAS) in a 3D grid to define cavities, similar to VOIDOO.

CHUNNEL uses Solvent Excluded Surface (SES).

SURFNET places a sphere between each pair of atoms whereas HOLLOW places a fixed sphere size grid. [12]

3. PREVIOUS RESEARCH

Drugs do have side effects because they lack specificity. Therefore, drug design efforts focus on receptor analysis. Part of structure-based drug design are the structural determinants that dictate binding, meaning a detailed mapping of cavity topology and designing a ligand accordingly. Finding conformations of active sites leads to better docking and accurate prediction.

Previous work performed by Fotopoulos & Papathanasiou combines algorithms and statistical analysis techniques to develop accurate binding affinity predictions.

The following steps indicate a short overview of their approach flow and methodology:

1. Degrees of freedom, restraints of backbone define complexity
2. Creation of conformations is based on simulation of chi angles rotation
3. Conformers are clusters based biological validation or superposition, RMSD, k-means
4. Polygonal shapes and their size help ligand analysis and predict potential ligand

Amino acids of the active site are selected from online databases [124]. The active site is expanded with geometrical techniques or analysis of secondary structure. After the selection of the active site aminoacids and their surrounding regions, different conformers are produced with divide and conquer approach.

Conformers are clustered in a 2-level hierarchical analysis: clustering of the proteins, clustering of the conformers for each cluster. Clustering with the multidimensional k-means algorithm is performed. Iterative closest point was used for the alignment. For each cluster of conformers α -shapes and rotation regions of the active site are visualized. Parallel programming techniques are used for faster execution. Extensive analysis of phi and psi angles with the Ramachandran diagram comes next. Regions with k-nearest neighbors are classified as α -helix or β -strand. Conformation space patterns can be displayed. Structure mining tool is based on user-defined keyword combinations with the ability to download locally only the structures that have a corresponding paper.

Bioinformatics tools were developed or extension of function for statistical visualization tools.

4. BIOINFORMATICS TOOLS

Mining bioinformatics data is an emerging field. Data are analyzed and classified or predicted. Classification from data, predict categorical class labels and prediction models use statistical methodology such as regression analysis to predict functionality cases.

Major issues in classification and prediction is preparing data: removing noise (**Data Cleaning**), correlation attribute analysis (**Relevance Analysis**), perform transformation or classification methods to reduce data (**Data Transformation and Reduction**).

Conditions and evaluation follows. Classify or predict data correctly and efficiently, minimize computational cost, namely achieving *Accuracy*, *Speed*, *Robustness*, *Scalability*, *Interpretability* are critical. [61]

Bioinformatics tools assist to retrieve and analyze biological data using biochemical, mathematical and statistical methods.

Computational tools involve following activities [62]:

- [1] *Sequence databases*, a vast collection of biological molecule information
- [2] *Genome sequence databases*, a vast collection of genome sequences species
- [3] *Protein sequence databases*, it is worth mentioning the world wide Protein Data Bank (wwPDB) that has been exclusively designed to achieve and become freely available 3D protein structures
- [4] *Miscellaneous databases*
- [5] *Gene identification and sequence analysis*, understand different features of molecules or proteins
- [6] *Phylogenetic analysis*, track gene flow and predict certain features of molecules with unknown function
- [7] *Predict protein structure and functionality*, from 3D protein structures, mainly determined by X-ray crystallography and NMR techniques
- [8] *Molecular interactions*, predict protein structures and spot protein interfaces by docking protein in ligand
- [9] *Drug designing*, discover new drug molecules to cure diseases
- [10] *Molecular dynamics simulations*, molecular interactions occur in a time dependent manner

Clearly, 3D structures and architecture provide essential and additional information about biochemical mechanisms.

The prediction of ligand binding and catalytic sites can refer to design small molecules. The pocket identification process finds cavities in protein structures. [10]

In contrast to ligand-based approaches that need an initial set of bioactive compounds, structure-based docking requires only the 3D structure of the protein target. The structure-based molecular docking shows binding sites and affinity of the receptor protein. [5]

Over the years, many different approaches have been developed. The methods and tools are categorized by the searching keyword and combined approaches. Since a great number of tools have been published, emphasis has been given only on the recent advances in the field.

Table 1: Selected Bioinformatics Tools (2016-2017)

Tools classified by search keyword – name & publication date			
Protein cavity dynamics			
TRAPP TRAnsient Pockets in Proteins	2017	PrinCCes Protein internal Channel & Cavity estimation	2015
WATCLUST	2015	BetaCavity	2015
PockDrug	2015	trj-cavity	2014
EDTSurf	2013	PROPORES	2012
Voroprot	2011	fpocket	2010
ConCavity	2009	PoreWalker	2009
Screen2 Surface Cavity REcognition and EvaluationN	2006		
Catalytic sites enzyme / Classify active sites			
GASS-WEB Genetic Active Site Search-WEB	2017	HotSpot Wizard	2016
ACFIS Auto Core Fragment In silico Screening	2016	CleavePred	2016
iCataly-PseAAC	2015	PatternQuery	2015
LIBRA LIgand Binding site Recognition Application	2015	RRDistMaps	2015
Protein pockets			
PRNAclass	2017	JED Java Essential Dynamics	2017


SeekR	2017	FlexPred	2017
BALL-SNPgp	2016	locPREFMD local Protein structure REFinement via Molecular Dynamics	2016
HTMD High-Throughput Molecular Dynamics	2016	MSMBuilder Markov State Models Builder	2016
UNRES	2016	iGNM	2016
In silico docking			
Octopus	2017	DUck Dynamic Undocking	2016
bSiteFinder	2016	SCAR Steric-Clashes Alleviating Receptor	2016
VirtualToxLab	2016	WATsite	2014
ChemBioServer	2012		
Molecular dynamics simulation			
PrimaDORAC	2017	ReFOLD	2017
Bio3D-web	2016		
Protein conformers / Side chain angle			
GOAP Generalized Orientation-dependent All-atom Potential	2011	IBS Illustrator of Biological Sequences	2015
GASV Geometric Analysis of Structural Variants	2009	Jzy3d	2014
ArchPRED	2006	VTK Visualization Toolkit	2008


Methods and algorithms of some of those tools have been compared to the ones used by Fotopoulos and Papathasiou [1]. Since a great number of bioinformatics tools do exist some of them were excluded and placed in the tools table as “Other tools”. Selection criteria for the covered tools in this thesis was the publication date, as well as the algorithms and the methods they do use.


4.1.1 Protein cavity dynamics

A cavity is a void or channel inside a molecule that is not accessible to bulk solvent. [63]
A catalogue of commonly used tools for performing protein cavity dynamics is presented below.

Table 2: Protein cavity dynamics

Tools	
Search for: Protein cavity dynamics	
Name (publication date)	Description
TRAPP TRAnsient Pockets in Proteins (2017) 	<p>TRAPP (TRAnsient Pockets in Proteins) is an online automated software platform to track, analyze and visualize protein dynamics cavity using protein structure or motion trajectory. It performs grid-based calculations to detect the conserved and temporary regions. TRAPP workflow consists of three modules: Structure module produces structures using molecular simulation methods such as tCONCOORD, L-RIP, RIPlig, MD, Analysis module retrieves structures from Structure module or existing ones from PDB and categorizes them. Methods used for clustering of the binding site conformations are RMSD, single-linkage clustering, k-means clustering, Pocket detects, analyses, and visualizes binding pocket dynamics and characteristics. Lennard-Jones function is used to describe the interaction energy. Sequence conservation along with the Pocket module results are evaluated. Conservation scores, calculated with Jensen-Shannon divergence, are placed at the b-factor position in the structure. To map the conservation score on the 3D structure MUSCLE tool. [64] [65]</p>
PrinCCes Protein internal Channel & Cavity estimation (2015)	<p>PrinCCes (Protein internal Channel & Cavity estimation) is a computer program with easy approaches to visualize protein voids. Visualization of chambers, channels and other kind of cavities in proteins or protein complexes. Structures can be retrieved from PDB. Structure components can be shown and B and C chains removed. Set parameters for calculations. 3D grid is applied for the calculation. The algorithm is based on a continuously applied void decomposition. Large voids and their combinations are examined. A probe can be moved between any voids. 3D structure can be shown by VMD representation any time. The output result can be exported to VMD (Visual Molecular Dynamics) or Chimera software. PrinCCes is available for MS Windows and Linux. [66]</p>
WATCLUST	WATCLUST is a user-friendly tool, on the VMD platform, to determine WS and

(2015)	<p>their properties and used to analyze them. Water molecules are important in protein folding and function. Additionally WS can be structurally and thermodynamically characterized. Other tools for WS detection are WaterMap, a grid based method, WATsite that is restricted to perform MD simulations only with GROMACS. The following parameters are calculated: water finding probability (WFP), R90, WS-protein mean interaction energy, WS-water mean interaction energy, mean total interaction energy with respect to bulk, and excess rotational (Sr) and translational (St) entropies. WS information can be used in Autodock program to perform biased docking (WSBD). WATCLUST executes four steps: Load Trajectory water MD simulation data set is loaded in the VMD program, Determine WS parameters such as how and where the WS are to be determined are defined. It is important not to include residues with high mobility, Analyze WS colored WS along with parameters can be shown in VMD. All structural and thermodynamic parameters are kept in log files for each WS, Transfer WS data to Autodock after a conventional oxygen grid map is loaded into VMD and required WS are selected docking calculations can be performed. [67]</p>
BetaCavity (2015)	<p>BetaCavityWeb is an online tool that computes, for a given molecular structure, molecular cavities, voids and channels. Molecular structures can be selected from PDB or the user can uploaded his own ones. Calculations are based on geometrical properties such as volume, boundary area. The algorithm uses beta-complex based on the Voronoi diagram of atoms. It produces text and graphic (JSmol) results, and channel spines and their bottleneck can also be reported. BetaCavityWeb has four elements: geometric kernel computes the Voronoi diagram and extracts the beta-complex, trimmer of the Voronoi diagram that uses van der Waals molecule or a Lee–Richards solvent accessible surface model, classifier recognizes voids and channels in the previously generated Voronoi graph, evaluator of geometric properties. [63]</p>
PockDrug (2015) 	<p>PockDrug is a robust tool that provides reliable druggability by using different estimation pocket methods. It can predict docking (holo) and homology form (apo) pocket druggability. A dataset is retrieved from NRDLD (NonRedundant dataset of Druggable and Less Druggable binding sites) that is spliced into two test sets: a training and an independent set. Another dataset is retrieved from DCD (Druggable Cavity Directory) database. Pocket estimation methods are prox and fpocket. Three statistical steps are executed: linear analysis models combined with a set of pocket descriptors are optimized by dataset cross-validation, seven of the most stable and efficient models using NRDLD independent test set are estimated, the construction of one unique consensus druggability model. Matthew's correlation coefficient (MCC) is applied to validate the model. The results can be visualized and saved in different formats: sortable table, pocket visualization using Jmol, compressed file. The tool has been implemented using</p>

	Python, HTML, CSS and JavaScript. [68]
trj-cavity (2014)	A tool that finds protein cavities throughout Molecular Dynamics (MD) simulation trajectories and analyzes their occupancy. GROMACS framework is used. Provides computational efficiency and can be time-dependent. Protein channels are characterized. Lennard-Jones cavity detection algorithm is used. Parameters for the simulation of MD are taken from an input file or GROMACS directory. To find the " burial state " linear search is performed. Possible option is the cutoff, which sets a maximum distance around the currently inspected voxel. [69]
EDTSurf (2013) 	EDTSurf Michel Sanner's Molecular Surface (MSMS) is an efficient tool for protein structure analysis and structure-based function annotation. It identifies cavities. The algorithm converts protein structure in a 3D gray-scale image. It uses Euclidean distance transform (EDT). It generates three macromolecular surfaces: van der Waals, solvent-accessible, and molecular. To build triangulated surface Marching Cube algorithm is used. To align sequences Needleman–Wunsch algorithm is used and a two-layer back-propagation neural network for prediction. The method is faster than the one proposed by Chakravarty and Varadarajan (CV method). Pearson's correlation coefficient executes an exhaustive search. Protein fold recognition uses residue depth and is calculated by Macromolecular Visualization and Processing (MVP) program. The calculation is performed on a 2.27 GHZ Intel E5520 Xeon processor and 24 GB memory. [70]
PROPORES (2012)	PROPORES is a toolkit that identifies pockets, cavities and channels of protein structures. The program checks if ligand binding pockets or blocked channels can be achieved. Side chains are systematically rotated and more flexible view is available. Structure data can be selected from PDB. The program is implemented in PERL language. PoreID stands for pore identification, PoreTrace is the pore axes determination, and GateOpen represents opening gate between neighboring pores. [71]
Voroprot (2011)	Voroprot is an interactive platform tool and gives the opportunity to explore geometric features of protein structure. Uses Voronoi diagram. [72]
fpocket (2010)	fpocket is online tool small-molecule pocket detection. Actions that can be performed are: fpocket to find candidate pockets, mdpocket for pocket tracking during molecular dynamics, hpocket a transposition of mdpocket to the combined analysis of homologous structures. The protein pocket detection algorithm is

	based on Voronoi diagrams. [10]
ConCavity (2009)	ConCavity is an online tool that identifies protein surface cavities. The prediction algorithm combines sequence conservation and structure-based methods. Data, source code, and prediction visualizations are available on the ConCavity web site. [73]
PoreWalker (2009)	A tool that uses protein 3D structure to detect and characterize transmembrane protein channels. The user is looking for big and long cavities. Features such as size, shape and regularity of the pore are calculated and visualized. [74]
Screen2 Surface Cavity REcognition and EvaluationN (2006)	SCREEN (Surface Cavity REcognition and EvaluationN) is an application for identification of protein cavities. Cavity attributes are calculated and used in classification analysis. It defines surface cavities geometrically, meaning the empty space between the protein's molecular surface and an envelope surface. It supports precise detection, characterization, and classification of protein surface cavities. By applying modern machine learning techniques (Random Forests) drug-binding cavities with a balanced error rate are identified. [75]
Other tools: IsoCleft Finder (2013), AADS automated active site detection, docking, and scoring (2011), RosettaHoles (2010), pdBFun (2005)	

Detailed description of selected tools is presented in the following paragraphs.

4.1.1.1 TRAPP Transient Pockets in Proteins

TRAPP is an automated workflow for tracking, analysis and visualization of protein dynamics cavity using protein structure or motion trajectory. It consists of three modules as someone can see in Figure [1] TRAPP Workflow.

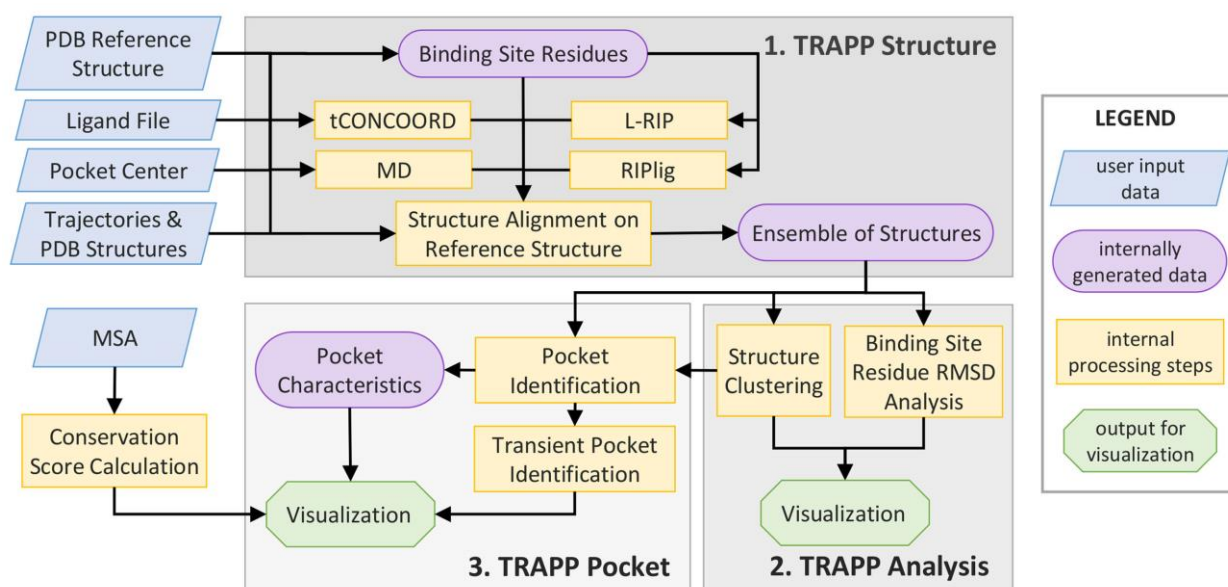


Figure 1: TRAPP Workflow [64]

TRAPP Structure

tCONCOORD method predicts protein flexibility based on geometrical considerations. RIP (Rotamerically Induced Perturbation) generates deformation maps that characterize the mobility in a protein structure. RIP procedure is modified by using a Langevin equation thermostat in the simulations.

TRAPP Analysis

RMSD (Root Mean Square Deviation of atomic positions). Single-linkage clustering is compared with k-means centers. If the distance is less than the cluster threshold then merge them. Residue vs Snapshot of a particular trajectory.

TRAPP Pocket

Lennard-Jones atomic radii. A pocket cavity detection algorithm. A function that describes the energy of interaction between two rare gas atoms as a function of the distance between their centers.

The sequence conservation per residue together with the above pocket results is analyzed. The user can choose between an average conservation score and difference with and without off-target sequence. Multiple Sequence Alignment in FASTA format is uploaded. For the calculation of MSA, Jensen-Shannon divergence is used with some exceptions. Jensen-Shannon divergence is a method of measuring the similarity between two probability distributions based on Kullback-Leibler divergence.

Scores are placed at the B-factor and visualized by color (JSMOL). MUSCLE tool is used to map conservation score and 3D reference structure. MUSCLE Tool is a computer program that creates multiple alignments of protein sequences.

MUSCLE algorithm steps are:

Distance estimation

Uses k-mer counting and Kimura distance. Related sequences tend to have more k-mers. Hierarchical clustering method UPGMA is used to classify distance matrices and producing a binary tree.

Profile alignment

A modified version of log-average function (maximum likelihood approach).

Refinement

The score of the produced multiple alignments is checked. If it is improved than the new alignment is kept. If not than it is discarded.

MUSCLE Tool uses k-mer clustering. To identify the relevant positions of an entry can be very time consuming. Whereas Master Thesis of Fotopoulos and Papathasiou [1] uses multidimensional k-means clustering (RMSD), specifically distance all-vs-k. This algorithm uses superposition and does not compare pairwise differences, leading to a faster computation.

Further, a different way to categorize gene information is the **J-Express**. It is a Java application to analyze gene expression (microarray) data in a flexible way giving access to multidimensional scaling, clustering, and visualization methods in an integrated manner. J-Express includes implementations of hierarchical clustering, k-means, principal component analysis, and self-organizing maps. [76]

An alternative approach of clustering is **GrammR**, which is a tool for graphical representation and clustering of metagenomic count data. Given the matrix of metagenomic counts for samples, this package quantifies dissimilarity between samples using Kendall's tau-distance, constructs multidimensional models of different dimension, and plots the models for visualization and comparison. [77]

4.1.1.2 PockDrug-Predict Pocket Druggability

PockDrug-Server uses different pocket estimation methods to predict pocket druggability. Ligand proximity helps to estimates pocket that is based on protein structure information.

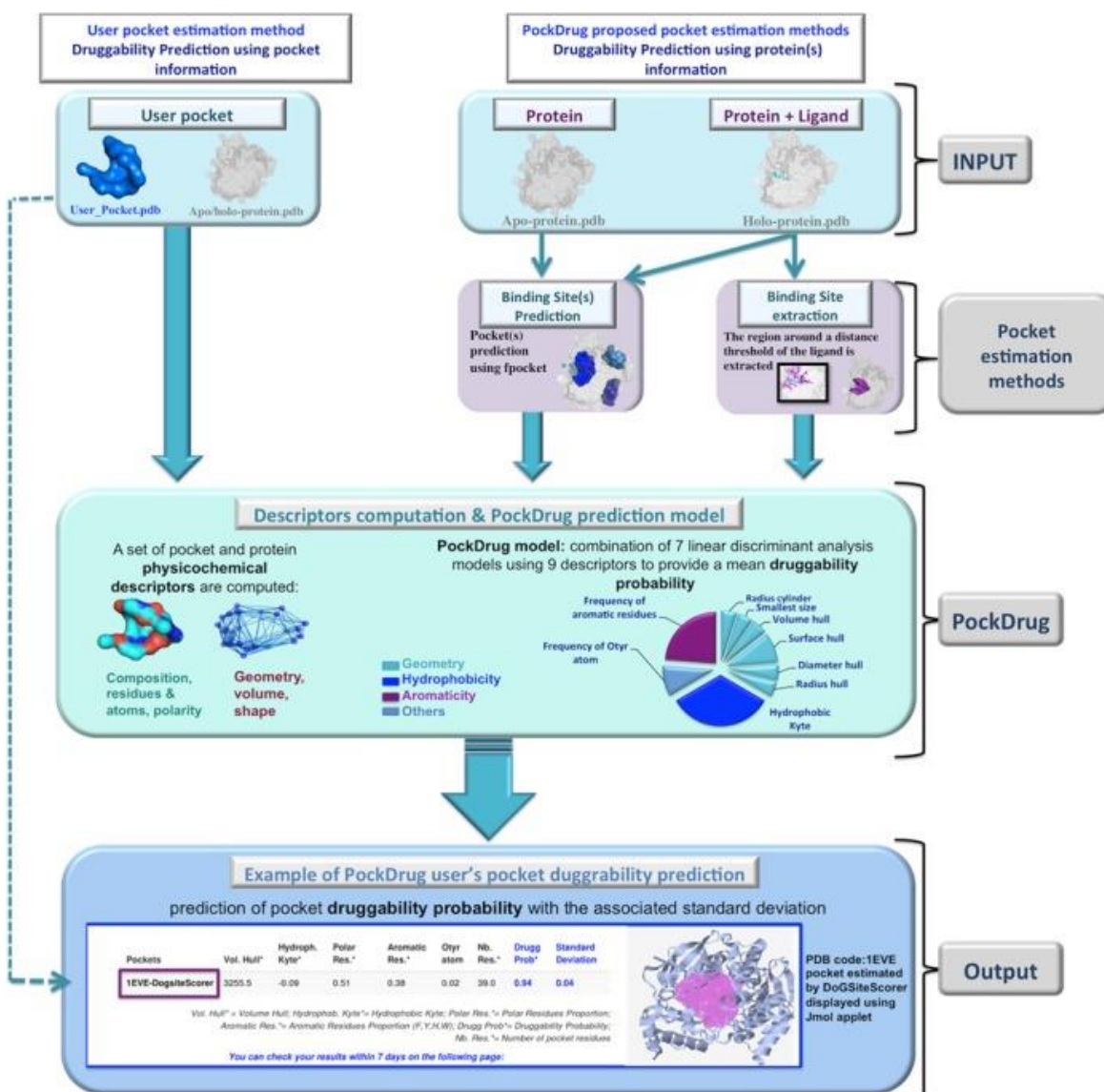


Figure 2: PockDrug Workflow [68]

PockDrug workflow is divided in four sections:

INPUT

Pocket or protein structure is retrieved from PDB.

Pocket estimation methods & PockDrug

Two methods prox and fpocket are used, either one of them or both. Standard deviation is provided. Prox is guided by the ligand position information known as prox4 and prox5.5. This threshold has a strong influence on pocket descriptors. fpocket does not uses ligand position. It is an automated geometry based method. It increases calculation speed and has a good performance.

Output

Accordingly to the input data and estimation method the submitted query can fetch a *single pocket structure or protein* where results can be displayed in a sortable table, visualized using Jmol, or compressed file, or a *list of PDB codes*, specifically a table that

contains protein PDB codes, number of estimated and druggable pockets, and the highest druggability probability and its standard deviation.

PockDrug model accuracy, sensitivity, specificity and Matthew's correlation coefficient (MCC) have been evaluated in Borrel et al. study "PockDrug: a model for predicting pocket druggability that overcomes pocket estimation uncertainties".

4.1.1.3 EDTSurf

Residue depth measures the distance between the residue of interest and its nearest neighboring water molecule or protein surface.

The EDTSurf tool provides a way to describe the protein tertiary structure by calculating the geometric feature depth. It can be used to classify proteins and compare their structures. The structures are retrieved from PDB.

Euclidean distance algorithm has been used in atom protein structure and exhaustive search (ES) to calculate the depth of each atom. Triangulated surface is built by using Marching Cube method. Macromolecular Visualization and Processing (MVP) program is used to visualize the depth.

Two-layer back-propagation neural network predicts the residues. Residue depth is used to recognize protein fold. The scoring function for matching depth of residue or matching 2D structure is the negative logarithm of the probability of a residue type. The scoring equation combines the dot-product of the frequency profile and the log-odds profile from the Position-Specific Substitution Matrix. Dynamic programming, Needleman–Wunsch algorithm, is performed to align protein sequences.

Chains are randomly selected from the PISCES list. The Pearson's correlation coefficients (PCC) that has a high correlation (>0.90) finds depth similarities.

Radius of gyration (RG) and radius of the bounding sphere (RS) are measurements that characterize and predict protein shapes. The RG refers to the Root Mean Square Distance of protein atoms. The 3D visualized structure can be examined.

MUSTER program combines alignments with sequence based 2D structure. The Template Modeling Score (TM-score) can provide alignment accuracy.


Performance of structure based methods is crucial in protein prediction process. **MUSTER** algorithm is a profile-profile alignment based method. Multiple structure features such as 2D structure and residue depth are included in prediction calculations. Their accuracy is the key. Incorrect assignments may risk the algorithm performance. Quality needs to be improved. [78]

Compared to the alpha and beta shapes approach in Master Thesis of Fotopoulos and Papathasiou [1] the MUSTER tool has weak predictive performance and is more time-consuming.


4.1.2 Catalytic sites of enzyme / Classify active sites

An active site is a region of an enzyme where substrate molecules bind and chemical reaction occurs. This means that residues bind with the substrate, named binding site, and catalyze a reaction, named catalytic site. [8]

Table 3: Selected tools for analyzing the catalytic sites of enzymes

Tools	
Search for: Catalytic sites of enzyme / Classify active sites	
Name (publication date)	Description
GASS-WEB Genetic Active Site Search-WEB (2017)	GASS (Genetic Active Site Search) is a user friendly online tool that searches similar active sites in proteins. Active sites are retrieved from the Catalytic Site Atlas. Input data are also selected from PDB and NCBI-VAST. Matthew correlation coefficient and Critical Assessment of protein Structure Prediction (CASP 10) dataset are used. MeGASS is a multi-objective version of GASS. The reports do have information such as enzyme EC number, UNIPROT accession code and structure resolution. Two approaches are available: a protein is searched in the stored database, an active site is searched in protein structure databases. Search process is based on a genetic algorithm . A heuristic search finds matching active sites. The results are ordered by a fitness function , a modified version of RMSD, which indicates structural similarity. The tool is implemented in Python and C. [79]
HotSpot Wizard (2016) 	HotSpot Wizard 2.0 is an online application that identifies hotspots and designs libraries with information regarding stability, catalytic activity, substrate specificity and enantioselectivity, extracted from three databases and twenty tools. Protein structures are visualized with Jsmol applet. The HotSpot workflow has three steps : <i>annotation of a protein</i> , where MakeMultimer tool, CAS and UniProtKB, DSSP algorithm, Shrake and Rupley algorithm with BioJava, Fpocket score, CAVER, BLAST, USEARCH, UCLUST are used, <i>identification of mutagenesis hot spots</i> , and <i>design of the smart library</i> , the SwiftLib tool can be used to calculate codons. [80]
ACFIS Auto Core Fragment In silico	Auto Core Fragment in silico Screening (ACFIS) is a highly valuable tool that performs computer-aided fragment-based drug discovery. It uses active molecules to create fragment structure. The computation is based on the pharmacophore-

Screening (2016)	linked fragment virtual screening (PFVS) method. The tool creates new molecules based on growing algorithm. ACFIS has three modules : <i>PARA_GEN</i> generates parameters, AutoDock Tools, CHIMERA and PYMOL are used, <i>CORE_GEN</i> identifies core fragment structure from a bioactive molecule by using the RCSB protein data bank or performing docking calculations. DAIM program reduces the ligand structure, <i>CAND_GEN</i> links fragments to the basic one and generates candidate, binding free energy calculation method MM_PBSA , that uses the Poisson-Boltzmann equation, or MM_GBSA , that uses the Generalized Born approximation, is performed. [81]
CleavePred (2016)	<p>ASAP (Amino-acid Sequence Annotation Prediction) is a ML framework for predicting residue properties. The founded features are used to train ML classifiers and new classifiers can be constructed. CleavePred is a tool of ASAP and used to design proteins. It is trained to perform Residue Level Prediction (RLP). PSSM profiles are produced using SCRATCH's ProfilPRO. Scikit-learn elimination along with cross-validation (RFECV) are used to identify the features. CD-HIT program is used to reduce data and USEARCH to define the maximum similarity threshold. The tool algorithm combines: Support Vector Machine, Random forest, and Logistic regression. CleavePred is fast. Python API is used. [82]</p> <p>Scikit-learn, Machine Learning in Python, is an efficient tool for data mining and data analysis. The available tools are divided based on their functionality: Classification, Regression, Clustering, Dimensionality reduction, Model selection, Preprocessing. The algorithms include Support Vector Machines, Random Forest, k-means. [83]</p>
iCataly-PseAAC (2015)	A user friendly online tool that predicts, identifies enzymes catalytic sites using sequence evolution information. The grey system model GM(2,1) is used. Jackknife test has been performed for evaluation. Input data is retrieved from the Catalytic Site Atlas (CSA) and protein sequences and their active sites from Uniprot. CD-Hit program is used for cutoff to eliminate homology bias. The peptides with known catalytic sites are classified. The most common algorithms for prediction accuracy of predictor are K-nearest neighbor rule, neural work methods, and Support Vector Machine. In the protein the pseudo amino acid composition (PseAAC) replaces the simple amino acid composition (AAC). Mathew's correlation coefficient is used to evaluate the predictor. The grey model is useful in cases where necessary information is absent. The GM(2,1) model uses the following three approaches : the K-nearest neighbor algorithm that is one of the powerful methods for performing nonparametric classification, the Fuzzy K-NN classifier that is a variation of the K-NN classification family, and the <i>Euclidean</i> metric. The jackknife test was used to examine the predictor's quality, iCataly-

	PseAAC predictor. Results are shown by Receiver Operating Characteristic (ROC) curve. [84]
PatternQuery (2015)	PatternQuery (PQ) is a user friendly online application and it provides detection and extraction of biomacromolecular patterns such as binding sites and catalytic sites. The application uses a unique query language that is <i>similar to Python</i> . Input data are retrieved from PDB. Molecular structures are described based on the nature and relationship of atoms and residues. Different protein structures levels (primary, secondary, tertiary) are accessible. Two approaches are available: real time investigation, based on prior searches. PatternQuery workflow consists of four steps : <i>Query definition</i> , multiple queries can be executed by one run, the maximum number of queries is 10, <i>Input data specification</i> where PDB subset version is defined, <i>Running the PQ query</i> , execution time depends on the given query and data, <i>Visualization and analysis of retrieved patterns</i> , ChemDoodle can be used. [85]
LIBRA Ligand Binding site Recognition Application (2015)	A software tool that predicts and identifies active sites and ligand binding sites. The algorithm is based on a graph theory approach. It locates the largest subset of similar residues. Input data are retrieved from the Catalytic Site Atlas and the Protein Data Bank. The tool is implemented in Java SE 7 with a Swing GUI embedding a JMol applet and it can be run on any operating system that supports Java Virtual Machine (JVM). [86]
RRDistMaps (2015) 	RRDistMaps is a UCSF Chimera tool that calculates maps to identify structural features from 2D structures. Conformational binding site or residue changes, motion between unbound and bound proteins can be visualized. No cutoff threshold is used but a color-coded distance map shows the interactions. The tools approach for max two chains is: select a chain, compute RR chain distances, and display a grey and a color-coded distance map. If more than two chains selected it uses Needleman–Wunsch algorithm , MUSCLE tool and an online RBVI web service. The average α -carbon distance and standard deviation are computed from all chains Chimera's Multalign Viewer tool is used to show sequence alignment. [87]
<p>Other tools:</p> <p>JET2 Viewer (2017), a database of predicted multiple, possibly overlapping, protein-protein interaction sites for PDB structures.</p> <p>L1Base (2017), a database that contains more retrotransposition-active LINE-1s, more mammalian genomes.</p>	

DASP3 (2016), Deacon Active Site Profiler, identification of protein sequences belonging to functionally relevant groups.

KLIFS (2016), Kinase-Ligand Interaction Fingerprints and Structures database.

STarMirDB (2016), a database of microRNA binding sites.

xVis (2015), a web server for the schematic visualization and interpretation of crosslink-derived spatial restraints.

EzCatDB (2015), the enzyme reaction database, classifies enzyme reactions based on enzyme active site structures and their catalytic mechanism.

novPTMenzy (2015), a database for enzymes involved in novel post-translational modifications.

CSA Catalytic Site Atlas (2014), **VarMod Variant Modeler** (2014), **STarMir** (2014), **SAPTA Scoring Algorithm for Predicting TALEN Activity** (2014), **HINT Hmm-based IdeNtification of Tf footprints** (2014), **SparScape** (2014), **MC Path, Monte Carlo path generation** (2013), **ActiveDriver** (2013), **ASMC Active Sites Modeling and Clustering** (2010), **SitesIdentify** (2009)

Detailed description of selected tools is provided in the following paragraphs.

4.1.2.1 HotSpot Wizard

HotSpot Wizard 2.0 is an online application that identifies hotspots and designs libraries with information regarding stability, catalytic activity, substrate specificity and enantioselectivity, extracted from three databases and twenty tools. Protein structures are visualized with Jsmol applet.

The HotSpot workflow has three steps:

annotation of a protein, MakeMultimer tool is used to generate protein target, residue information is downloaded from CAS and UniProtKB. DSSP algorithm finds the 2D structure, Shrake and Rupley algorithm with BioJava finds accessible surfaces. Protein pockets are identified by the Fpocket score and access tunnel with CAVER. Clustal Omega creates multiple sequence alignment. Jensen-Shannon entropy is used to estimate conservation.

identification of mutagenesis hot spots.

design of the smart library, the SwiftLib tool can be used to calculate codons.

The results include residue features, details, and pocket and tunnel information. They can be visualized with JSmol.

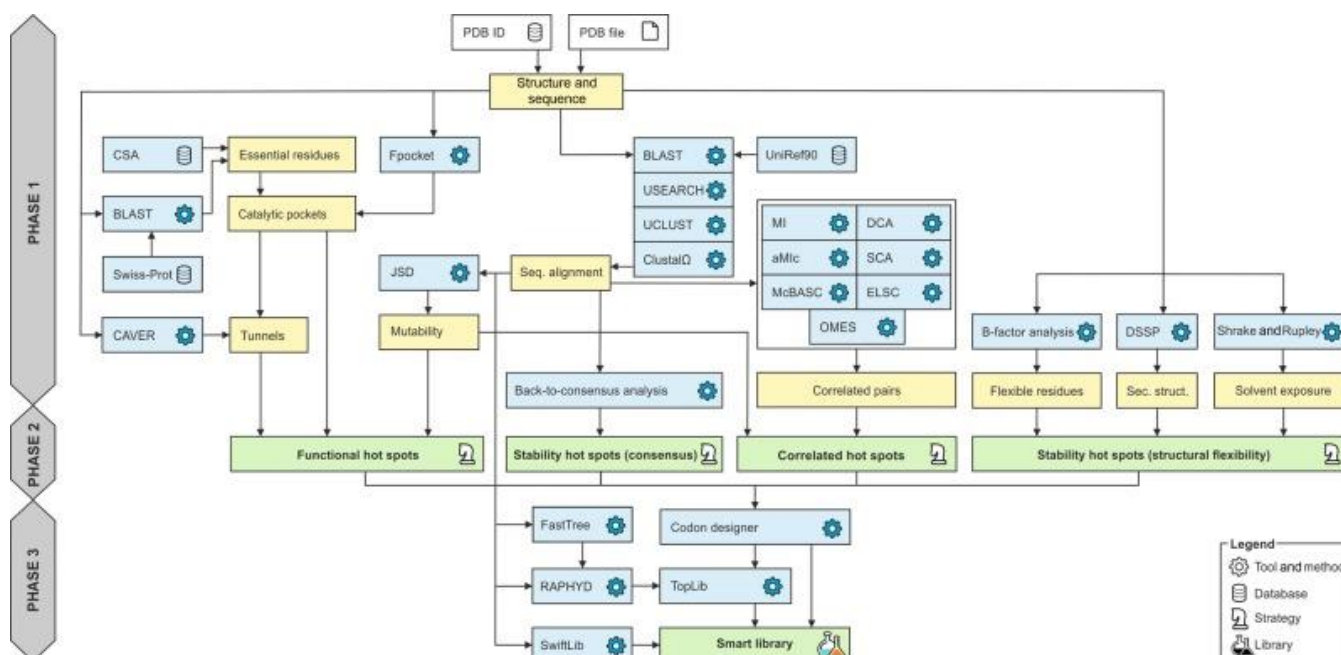


Figure 3: HotSpot Workflow [80]

4.1.2.2 RRDistMaps

UCSF Chimera is an online program for interactive visualization and analysis of molecular structures and related data. Information such as density maps, sequence alignments, docking results, trajectories, and conformational ensembles are included. [88]

The available tools are divided based on their functionality:

General Controls, Viewing Controls, Depiction, Sequence, Amber, Higher Order Structure, Volume Data

Structure Analysis like Angles/Torsions, Metal Geometry

Structure Comparison like **RR Distance Maps**, Morph Conformations

Surface/Binding Analysis like FindHBond, ViewDock, Dock Prep

Structure Editing like Rotamers (**Ramachandran Plot**), Build Structure

MD/Ensemble Analysis like Molecular Dynamics Simulation, Ensemble Cluster

Movement like Rotation, Constraint Move


Utilities like Structure Diagram


The available **shapes** in UCSF Chimera are: sphere, cylinder, and icosahedron.



4.1.3 Protein pocket

A more detailed reference on pockets and their importance to understand protein dynamics and molecular processes is shown in the following table. The pocket identification process finds cavities in protein structures. [10]

Table 4: Protein pocket

Tools	
Search for: Protein pocket	
Name (publication date)	Description
PRNAclass (2017) 	<p>PRNAclass is a platform to identify the RNA-binding residues in 3D protein structures and extract pockets involved in binding site process. Clustering pockets based on a structure based method. Analyzing RNA recognition motifs from protein 3D structures is important due to that fact this knowledge shows the structure groups that lead to a mechanism. RNA-binding protein complexes are retrieved from PDB. Entangle is used to identify binding residues. Protein pockets refer to empty concavities that the solvent can access, known as mouth openings. Accessible pockets are found by CASTp. Local and global structure similarities are measured using structure alignment algorithm SAMO (Structural Alignment by Multi-objective Optimization). Clustering and investigation of RNA-binding pockets and proteins follows. Q-score is used to find pocket similarity. The evaluation uses a P-value that is based on that Q-score, extreme value distribution (EVD). Threshold is set to 0.05. A similarity-network-based strategy is performed, classification of binding pockets finds similarities between pockets by mapping edges of pockets, the topological structure, with a significant structure similarity. Groups are created. Gene Ontology (GO) and DAVID are used to improve or implement functional analysis of these groups. Cytoscape and Pymol are used for structure description and visualization. [89]</p>
JED Java Essential Dynamics (2017)	<p>A package written in Java that compares Essential Dynamics (ED) from multiple protein trajectories performing multivariate statistics. Calculates cumulative overlap (CO), root mean square inner product (RMSIP) and principal angles (PA) to eliminate outliers. Important function of the JED toolkit are the Z-score based on elimination of outliers, distance pair PCA (dpPCA) and the comparative analysis of subspaces. Main JED features are: <i>outlier removal, creation of Pymol scripts to visualize motion over user-selected time scales, creation of free energy</i></p>

	<p>surfaces based on Gaussian kernel density estimation, <i>calculation</i> of the precision matrix from Q, the partial correlation matrix with its eigenvectors and eigenvalues, <i>comparison</i> of dynamics of multiple proteins, <i>multivariate statistical analysis methods</i>. Outliers are removed based on a user-defined threshold. A dynamic trajectory show the various conformations of a protein. JAMA Matrix package is used that calls the KDE. A file JED_Driver.txt includes all required information for the job execution. Chains are retrieved from PDB. Analysis at the coarse-grain level of all alpha carbons is performed. The B-factors in a PDB file are replaced with residue RMSD. Weighted and unweight mean squared fluctuation (MSF) and root mean squared fluctuation (RMSF) are calculated for the covariance (Q), correlation (R) and partial correlation (P) matrices. [90]</p>
<p>SeekR (2017)</p>	<p>Simulation Enabled Estimation of Kinetic Rates (SEEKR) is a package that includes scripts and tools for designing and performing multi-scale ligand and protein binding kinetics calculations. Brownian dynamics and milestoning theory is performed. Simulations can be performed locally or on supercomputers. k_{on}, k_{off}, and ΔG_{bind} are computed and kinetics checked. SeekR combines two techniques to affectively compute binding kinetics and thermodynamics: Molecular Dynamics (MD) a costly method for a “brute force” approach, and Brownian Dynamics (BD) for a binding kinetics approach. The tool is divided in three steps: Prepare input values given by the user are used to produce files. These files are necessary for the other steps. Run submits a job that is put in a SLURM supercomputer queue. BD simulations are performed. Analyze is responsible to execute milestoning and error calculations. Kinetics and thermodynamic information are produced. [91] [92]</p>
<p>FlexPred (2017)</p> 	<p>FlexPred is an online tool to predict absolute per-residue variation from a 3D protein structure. It uses static features of a protein structure to predict MD residue fluctuation in 3D. For evaluation Pearson's correlation coefficient is used. [93]</p>
<p>BALL-SNPgp (2016)</p>	<p>A tool to characterize structural and functional non-synonymous single nucleotide variants (nsSNVs). BALL-SNPgp is based on SNV data, collects information regarding disease relevance and visualizes the 3D protein. Cluster analysis and binding pockets are performed. In other words it improves pathogenicity assessment in computational diagnostics. [94]</p>
<p>locPREFMD local Protein structure REFinement via</p>	<p>locPREFMD, local Protein structure REFinement via Molecular Dynamics is an online tool that uses molecular dynamics simulations to achieve high resolution protein structures. Basic condition score is the MolProbity one. It improves the</p>

Molecular Dynamics (2016) 	<p>stereochemical quality of given models while Ca coordinates are constant. Missing atoms are included by the complete.pl tool from the MMTSB Tool Set. Ca positions are reconstructed and a residue is converted to the SICHO (SlideCHain-Only) coarse-grained model. The side chain is ready. In case a side chain is missing the tool SCWRL is used for reconstruction. CHARMM all-atom modeling program adds hydrogen atoms. CMA potential was modified to increase penalties for phi and chi angles outside the preferred Ramachandran map areas. Fractions of rotamer and backbone torsion (Ramachandran) outliers are reduced. Electrostatic and Lennard-Jones interactions were cut off. The Berendsen thermostat was used in the MD simulations. SHAKE was applied to constrain atom-hydrogen distances. All of the models were downloaded from the CASP Web site. PROCHECK measures stereochemistry and the G-factor is a log-odds score of bonds, angles, and torsion angles. [95]</p>
HTMD High-Throughput Molecular Dynamics (2016)	<p>HTMD is a platform in Python to improve data generation and increase reproducibility. Input data protein structures are retrieved from PDB. The process leads and produces information regarding conformation stability and kinetic rates. Thousands of simulations are handled. Available components are CHARMM and AMBER, projection methods, clustering, molecular simulation, an Amazon cloud interface, Markov state models and visualization. [96]</p>
MSMBuilder Markov State Models Builder (2016) 	<p>Statistical Models for Biomolecular Dynamics. MSMBuilder, Markov State Models (MSMs), is a software package for constructing statistical models. Analyzes protein folding and conformational change. Includes algorithms related to Hidden Markov Models (HMMs) and time-structure based Independent Component Analysis (tICA). The package is developed in Python and C. The API is based on scikit-learn estimation. MSMBuilder provides a fast analysis of large molecular dynamics datasets. The production process has three transition steps: Raw Cartesian coordinate's □ Features □ Kinetic coordinates □ State labels. MSMBuilder supports algorithms such as SparseTICA, learning algorithms like principal components analysis (PCA), SparsePCA, MiniBatchSparsePCA, MiniBatchKMeans, K-Means (KCenters, KMedoids, MiniBatchKMedoids), hierarchical clustering. To examine the dynamics between active and inactive conformations Robust Perron Clustering Analysis (PCCA+) on MSM is performed. Generalized matrix Rayleigh quotient (GMRQ) score measures the ability of a model to capture the slowest dynamics of a system. [97]</p>
UNRES (2016)	<p>UNited RESidue (UNRES) is a tool to predict protein structure. Target proteins are replicated and the methods are using a multi extensive simulations exchange of them. Two interaction sites are present the united side chain and peptide group per residue. [98]</p>

iGNM (2016)	Gaussian network model (GNM) is a user-friendly interface and database for investigating the dynamics of proteins and their complexes. Structures are retrieved from PDB and structural dynamics data are performed. Visualization of 2D or 3D structure helps to get information about residues, cross-correlations, motion, hinge sites and energy localization spots. iGNM has seven basic components : X-ray crystallographic B-factors (3D/2D) , it provides the relationship between X-ray crystallographic B-factors and GNM-predicted mean square profiles and they can be presented via color coded JSmol. Mode shapes (3D/2D) , residue mobility can be shown with color coded diagrams. Domain separations by dynamics (3D/2D) , each residue moves to a positive or negative direction in a mode axis. GNM connectivity model (3D/2D) , displays the topology of the network as an interactive 3D network model. Cross-correlations (3D/2D) , displays the orientation correlations between pairs of nodes. Collectivity (2D) , measures the degree of cooperatively between residues. All results can be downloaded, visualized and analyzed. Relational tables can be exported in tsv, csv or excel format. [99]
Other tools: BALL-SNP (2015) , combines genetic and structural information to identify candidate non-synonymous single nucleotide polymorphisms. AlloPred (2015) , prediction of allosteric pockets on proteins using normal mode perturbation analysis. DoGSiteScorer (2015) , a web server for automatic binding site prediction, analysis and druggability assessment. CafeMol (2015) , a Coarse-Grained Biomolecular Simulator for Simulating Proteins at Work. MDLovoFit (2015) , automatic identification of mobile and rigid substructures in molecular dynamics simulations and fractional structural fluctuation analysis. RsiteDB (2015) , a database of protein binding pockets that interact with RNA nucleotide bases. READ DB (2015) , RNA Binding Protein Expression and Disease Dynamics database. PARS Protein Allosteric and Regulatory Sites (2014) , eMatchSite (2014) , DynaMine (2014) , iMODS (2014) , Evol (2014) , relax (2014) , pocketzebra (2014) , APoc Alignment of Pockets (2013) , TargetATPsite (2013) , aCSM atomic Cutoff Scanning Matrix (2013) , GSAtools (2013) , ShereKhan (2013) , Dynamic Analysis of Nucleosome and Protein Occupancy by Sequencing DANPOS (2013) , PocketAnnotate (2012) , PocketQuery (2012) , PocketOptimizer (2012) , BSP-SLIM (2012) , Provar Probability of variation (2012) , GUARDD (2012) , KOSMOS (2012) , PCA PocketAnalyzer (2011) , MDAnalysis (2011) , SiMMap Site-Moiety-Map (2010) , PARIS Pocket Alignment in Relation to Identification of Substrates (2010) , ghecom grid-based HECOMi finder (2010) , OptCDR (2010) ,	

PocketPicker (2007), **CASTp** (2006), **LIGSITEcsc** (2006)

4.1.3.1 PRNAClass

A platform that identifies RNA-binding structure motifs and classifies them.

As shown in the following flow diagram the algorithm has four steps:

Extract RNA-binding pockets retrieved from RBPs.

Comparing structures and create groups based on similarity calculated by the algorithm SAMO (Structural Alignment by Multi-objective Optimization).

Separate these groups into subgroups via a community detection method.

Identify RNA-binding structure motifs and analyze them.



Figure 4: PRNAClass Process Flow [89]

The community detection method is a hierarchical clustering one that improves the measures between groups considering space pocket similarity.

The purpose of the SAMO algorithm, a multiple pocket alignment is to find a consensus structure from these pairwise alignments and perform a **one-to-one residue matching** process between any pocket and the consensus structure. It is a **greedy approach**. To a group with maximal consistent aligned residue pairs from three pockets are all the other pockets added to maximize the objective function.

To find pocket similarity **Q-score** is calculated. For small structures it produces many false positives. It is a complex characteristic which takes into account both the alignment length and RMSD. It shows that higher Q-score indicates better matches compared to RMSD.

Extreme value distribution (**EVD**) is used that provides lower misclassification rates and rapidly execution. No large amounts of Monte Carlo sampling is required to set a novelty threshold. [100]

4.1.3.2 FlexPred

FlexPred is a tool that predicts MD residue from 3D protein structure. It estimates if a protein chain is flexible, calculates RMSD between Ca atoms in MD simulations and produces PDB files.

It combines features such as B-factor, 2D and residue distance from the protein center of mass, residue lower or upper half-sphere exposure.

These features are clustered by using Support Vector Regression SVR (LIBSVM tool, a library for Support Vector Machines).

Predictions are evaluated using Pearson's correlation coefficient and Root Mean Square Error.

In Master Thesis of Fotopoulos and Papathasiou [1] radar plot is used for the visualization of protein B-factor. It helps the user to understand and spot B-factors by the specific region's temperature.

4.1.3.3 locPREFMD

locPREFMD, local Protein structure REFinement via Molecular Dynamics is an online tool that uses molecular dynamics simulations to achieve high resolution protein structures. Molecular dynamics are used to improve the structural protein model of a problematic residue. Computational effort is medium to improve the MolProbity score. locPREFMD rebuilds problematic regions. Conformations are selected with the minimum MolProbity score and closer to the initial conformation based on C α RMSD.

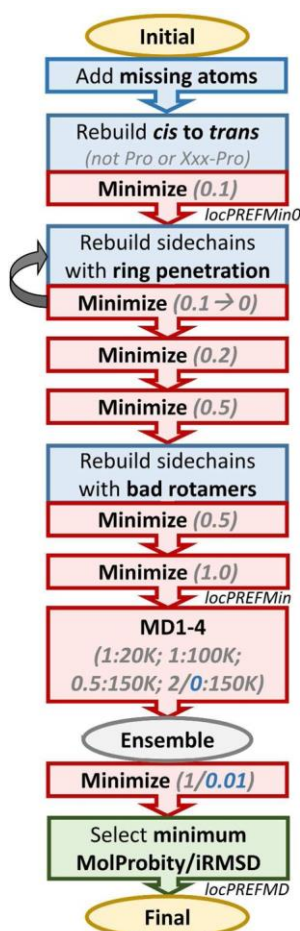


Figure 5: locPREFMD Flow [95]

Missing atoms are included by the *complete.pl* tool from the MMTSB Tool Set. Ca positions are reconstructed and a residue is converted to the SICHO (SldeCHain-Only) coarse-grained model. The side chain is ready. In case a side chain is missing the tool SCWRL is used for reconstruction. CHARMM all-atom modeling program adds hydrogen atoms.

CMAP potential was modified to increase penalties for phi and chi angles outside the preferred Ramachandran map areas. Fractions of rotamer and backbone torsion (Ramachandran) outliers are reduced.

Electrostatic and Lennard-Jones interactions were cut off.

The Berendsen thermostat was used in the MD simulations.

SHAKE was applied to constrain atom-hydrogen distances.

All of the models were downloaded from the CASP Web site.

PROCHECK measures stereochemistry and the G-factor is a log-odds score of bonds, angles, and torsion angles.

The Seok group reports with GalaxyRefine to refine models using the ROSETTA method.

4.1.3.4 MSMBuilder

MSMBuilder, Markov State Models (MSMs), is a software package for constructing statistical models. Analyzes protein folding and conformational change. Includes algorithms related to Hidden Markov Models (HMMs) and time-structure based Independent Component Analysis (tICA). The package is developed in Python and C. The API is based on scikit-learn estimation. MSMBuilder provides a fast analysis of large molecular dynamics datasets.

The production process has three transition steps: Raw Cartesian coordinate's, Features, Kinetic coordinates, State labels.

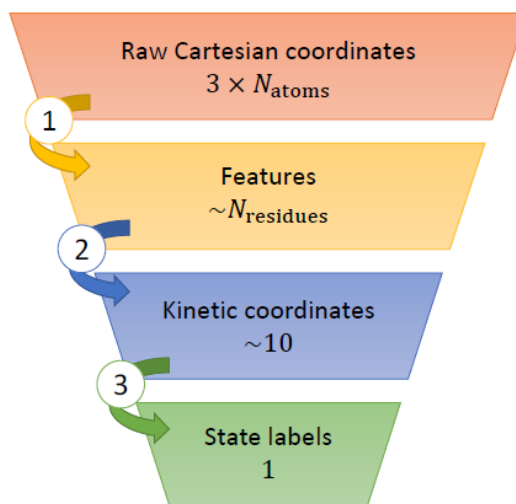


Figure 6: MSMBuilder process steps [97]

Raw Cartesian coordinates are transformed into vectors that contain only information according to translation and rotation to reduce the dimensionality of the data.

The second step although it improves statistical qualities it may discard important information. Slow coordinates between known active and inactive conformations are found by time-structure based independent component analysis (tICA).

MSMBuilder supports algorithms such as SparseTICA, learning algorithms like principal components analysis (PCA), SparsePCA, MiniBatchSparsePCA, MiniBatchKMeans, K-Means (KCenters, KMedoids, MiniBatchKMedoids), hierarchical clustering.

To examine the dynamics between active and inactive conformations Robust Perron Clustering Analysis (PCCA+) on MSM is performed. PCCA+ is a spectral clustering method. Generalized matrix Rayleigh quotient (GMRQ) score measures the ability of a model to capture the slowest dynamics of a system.

The MSMBuilder keeps information regarding **rotation** and **translation** of the structures in opposition with the Master Thesis of Fotopoulos and Papathasiou. [1] They presenting a method that uses multidimensional k-means clustering without rotations and translations of protein structures.


4.1.4 In silico docking

A protein must be flexible. The folded shape is an ensemble of multiple conformers and is referred to as proteins conformation or structure. Statistics can be used to find side chain conformational flexibility. Smaller chains tend to favor beta sheets while longer chains tend to favor the α -helix.

Protein side-chain conformation is related to their biological functions. The side-chain prediction is used in protein design and docking and structure optimization. [101]

Table 5: In silico docking

Tools	
Search for: In silico docking	
Name (publication date)	Description
Octopus (2017)	Octopus is a tool for docking simulations of an <i>unlimited number of compounds</i> into a set of molecular targets. It uses the semi-empirical method PM7 to improve geometry approach. It integrates MOPAC2016, MGLTools, PyMOL, and AutoDock Vina. Ligands are generated in PDB format. The simulations can be used in Our Own Molecular Targets (OOMT) databank. Best binding energies are kept in a standard table. Has a friendly Linux -based user interface. [102]
DUck Dynamic Undocking (2016)	DUck Dynamic UNdocking method that calculates where a ligand breaks the most important native contact with the receptor. It evaluates structural stability. It delivers several fold improvements. [103]
bSiteFinder (2016)	bSiteFinder is a tool to identify protein binding sites. Stability of Complex is a new criterion in clustering algorithm to improve the accuracy. Homology and chain length indexing increase efficiency of the structural alignment. Structural alignment uses Combinatorial Extension (CE) algorithm that defines continuous residues in the sequence as aligned fragment pairs (AFPs). Ligands are clustered and the largest one is defined as the center of the binding site. Datasets are retrieved from LIGSITEcsc. Accuracy and Matthews Correlation Coefficient (MCC) used for evaluation. [104]
SCAR Steric-Clashes Alleviating	SCARs (steric-clashes alleviating receptors) tool is a docking program used to design novel covalent ligands. It eliminates steric clashes between covalently bonding atoms. Already known docking methods are modified and parameterized.

Receptor (2016)	[105]
VirtualToxLab (2016) 	VirtualToxLab is a tool used to check if a potential drug is toxic, find known or suspected side effects. Binding modes of compounds known as off-targets are analyzed in 3D or 4D real time. Simulates interactions by using flexible docking combined with multi-dimensional QSAR (mQSAR). The 3D structures are retrieved from the Cambridge Structure Database (CSD). Calculation of descriptors related to pharmacokinetics uses the Schrodinger's QikProp program , molecular weight (MW), polar surface area (PSA) and the VCC Lab AlogPs algorithm . Desmond Software checks dynamic stability of molecular interactions so that appropriate ligand targets are selected. A client-server on a Linux cluster is used for simulations. Lipinski's rule-of-five approach is applied. It increases likelihood for a compound. 4D analysis shows multiple poses with different orientation and changes of side chain conformations. [106]
WATsite (2014)	WATsite is a program with a graphical user interface based on PyMOL. The results can be displayed in PyMOL and can be used for ligand docking. The enthalpy and entropy of the water molecule are calculated and estimate the free energy profile of each hydration site. [107]
ChemBioServer (2012)	ChemBioServer is an online tool for mining and filtering chemical compounds. Pre and post processing of compounds is available. It has six sections : <i>basic search</i> using ChemmineR package, <i>filtering</i> where Lipinski Rule of Five is applied, advanced filtering Open Babel is used to convert Simplified Molecular Input Line Entry Specification (SMILES) files, <i>clustering hierarchical</i> and AP are provided and clustering is repeated until a maximum condition is reached, <i>customize</i> pipeline and <i>visualize</i> compounds' properties. The application is implemented in R language and PHP . JchemPaint and Jmol are used to visualize 2D and 3D structures. Compound Fingerprints are generated with Open Babel. Input files are in SDF or MOL format. [108]
Other tools: PREDICTA (2015), Predict DNA-Drug Interaction strength by Computing ΔT_m and Affinity of binding. LigMatch (2011), LigandExpo (2004)	

4.1.4.1 VirtualToxLab

Molecular docking is a method to determine compound's binding. This method can also find hypothetical compounds. A 3D structure target macromolecule is required. Flexibility can be visualized.

VirtualToxLab uses the QikProp program to predict molecular properties. Can be executed in two modes the normal and fast one in which some calculation parameters are excluded. QikProp was developed with the following process: The BOSS program and the OPLS-AA force field were used to perform Monte Carlo statistical mechanics. After evaluation atom characteristics are analyzed. Properties solubility (logS) and permeability (logP) are used to improve prediction process because. They consider molecular space along with experimental measurements. Results can be exported in CSV format files. **QikProp** predicts log BB (brain & blood) values but other outliers cannot be fitted by any regression method. [109]

VCC Lab AlogPs algorithm provides an online prediction of logP, water solubility and pKa(s) of compounds. Molecules are retrieved from molecules from PHYSPROP database. The user can create his own library. The algorithm can increase its prediction for the user's molecules up to 5 times. Also logD values are predicted. The logP prediction accuracy is Root Mean Squared Error equal to 0.35 and Standard Mean Error equal to 0.26. [110]

Desmond Software checks dynamic stability of molecular interactions so that appropriate ligand targets are selected. Minimization option can be selected to speed up job execution. This option uses a hybrid method of the steepest decent and the limited-memory Broyden-Fletcher-Goldfarb-Shannon (LBFGS) algorithms. To apply bond constraints Shake algorithm is performed. Ewald method evaluates electrostatic characteristics. Desmond Software replicates exchange simulations by copying the simulations in different temperature. [111]

4.1.5 Molecular dynamics simulations

Molecular dynamics simulations do help understand the relationship between macromolecular structure and their function.

A protein must be flexible. The folded shape is an ensemble of multiple conformers and is referred to as proteins conformation or structure. Statistics can be used to find side chain conformational flexibility. Smaller chains tend to favor β -sheets while longer chains tend to favor the α helix.

Protein side-chain conformation is related to their biological functions. The side-chain prediction is used in protein design and docking and structure optimization. [101]

Table 6: Molecular dynamics simulations

Tools	
Search for: Molecular dynamics simulations	
Name (publication date)	Description
PrimaDORAC (2017)	PrimaDORAC is a reliable online Interface for the Assignment of Partial Charges, Chemical Topology, and Bonded Parameters in Organic or Drug Molecules. The interface is written in FORTRAN90 and works on the last Generalized Amber Force Field parameter set (GAFF2). The Public Domain MOPAC7 program is used to compute Bond charge corrections (BCC). Creates small molecule drugs (SMDs) topology and parameter files. Increases the accuracy and can be helpful in MD simulation process. [112]
ReFOLD (2017)	ReFOLD is a tool that identifies and fixes errors in user supplied 3D models of proteins. The user can analyze the refined models by visualizing the residue locations that were improved and compare specific regions between those models. It was initially developed for the CASP12 experiment. The approach combines molecular dynamics simulations with NAMD and refinement with i3Drefine . It uses the quality estimation method ModFOLD . CHARMM22/27 was used as the parameter file combined with default TIP3P water model. Scores are ranked. For the visualization JSmol/HTML5 framework are used. [113]
Bio3D-web (2016)	A web application that is based on the Bio3D and Shiny R packages and used for analyzing the sequence, structure and conformational heterogeneity of protein families. The results are input information for analysis, mapping, clustering and prediction of structural dynamics. Bio3D-web provides principal component

	analysis (PCA) for relationship mapping and ensemble normal mode analysis (eNMA) to predict internal dynamics between protein families. No programming knowledge is required. Structure and sequence annotations are retrieved derived from RCSB, PDB and PFAM databases. Available sections are: Search tab identifies sequence similar structures, Align tab selects structures for multiple alignment, similarity and conservation analysis, Fit tab provides superimposed structures which means conservation analysis and RMSD based clustering, PCA tab displays the relationship between all structures in terms of the principal displacements, and eNMA tab displays dynamics of nucleotide exchange. [114]
<p>Other tools:</p> <p>FATSLiM (2017), Fast Analysis Toolbox for Simulations of Lipid Membranes.</p> <p>MD-TASK (2017), a software suite for analyzing molecular dynamics trajectories.</p> <p>WAXSiS (2015), a web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics.</p> <p>do_x3dna (2015), a tool to analyze structural fluctuations of dsDNA or dsRNA from molecular dynamics simulations.</p> <p>GalaxyRefine (2016), GROMACS (2015), pyMDmix - python Molecular Dynamics simulations with mixed solvents (2014), MEMBPLUGIN MEMBrane PLUGIN (2014), g_mmpbsa (2014), AMBER Assisted Model Building with Energy Refinement (2013), MDWeb provides most popular MD packages (2012), MOIL (2011), Meredys MESoscopic REaction DYnamics Simulator (2010), Abalone (2010), Desmond</p>	

4.1.6 Protein conformers / Side chain angle

A protein must be flexible. The folded shape is an ensemble of multiple conformers and is referred to as proteins conformation or structure. Statistics can be used to find side chain conformational flexibility. Smaller chains tend to favor β -sheets while longer chains tend to favor the alpha helix.

Protein side-chain conformation is related to their biological functions. The side-chain prediction is used in protein design and docking and structure optimization. [101]

Table 7: Tools for analyzing Protein conformers

Tools	
Search for: Protein conformers / Side chain angle	
Name (publication date)	Description
IBS Illustrator of Biological Sequences (2015)	Illustrator of Biological Sequences (IBS), a user-friendly Java software package with a dual-mode user interface. User can produce their own illustrative diagrams for either protein or nucleotide sequences. Provides graphical elements, such as polygons, brackets, curves and polylines to represent functional elements or regulatory molecules. Multiple options are available and biological sequences can be manipulated, recolored or rescaled. Protein mode : the protein sequences and their functional domains can be defined by specifying the starting and ending positions. Annotations are retrieved from UniProt database, NCBI GenBank and TCGA. Nucleotide mode : the color and thickness of polylines do represent different gene functionality. Image formats JPEG, PNG and TIFF are supported. The schematic diagram can be exported as a vector image in SVG format. All drawing details can be stored in a .xml file. The standalone package of IBS was implemented in JAVA and is supported by Windows, Linux and Mac OS. The web application was implemented in HTML5 and JavaScript. [115]
Jzy3d (2014)	Jzy3d is an open source java library that provides an application programming interface (API) to draw 3D data into surfaces, multiple scatter plots and bar charts, 2D and 3D graphs charts, spheres, triangles, polygons. Can easily deploy native OpenGL charts on Windows, UNIX, and MacOS and integrate into Swing, AWT, SWT or JavaFX. Jzy3d is available for other languages or platforms such as Scala, Groovy, Matlab, C#. Algorithms used are: Grid based and Delaunay surface pattern methods, 3D structure, 2D envelopes (Convex hulls), Polygon ordering for improved

	transparency rendering. Dual depth peeling: scene graph order independent transparency, Matlab-like array processors and statistics tools, Experimental Support Vector Machine integration (Svm3d). [116]
GOAP Generalized Orientation-dependent All-atom Potential (2011)	Generalized Orientation-dependent all-Atom Potential (GOAP) statistical potential for protein structure prediction that depends on the relative pairwise atoms orientation. GOAP has distance and angle dependent contributions . DFIRE, RWplus and dDFIRE are worse than GOAP for the distance-dependent component. The GOAP potential is extracted from known protein structures based on the inverse Boltzmann equation . A Monte Carlo simulation is used. Alanine peptide bond lengths and angles taken from CHARMM. GOAP can be included in the QMEAN score. [117]
GASV Geometric Analysis of Structural Variants (2009)	Geometric Analysis of Structural Variants (GASV), a tool for identifying structural variants from paired-end sequencing data. GASV identifies, classifies and compares structural variants. Structural variant is represented as a polygon and intersections are computed. Two techniques have been used to identify structural variants in the human genome: array comparative genomic hybridization (aCGH) and paired-end mapping. Neither of them measure the breakpoints of a structural variant exactly. A breakpoint is localized only to the distance between the genomic probes. The plane sweep algorithm is performed to determine whether n line segments in the plane have any intersections. [118]
VTK Visualization Toolkit (2008)	Generates 3D computer graphics, image processing, and visualization. VTK supports a wide variety of visualization algorithms. The toolkit supports parallel processing and integrates with various databases on graphical user interface (GUI) toolkits. Three conceptual stages : Modality-dependent Segmentation, Modality-independent Segmentation, Intermediate Slice Interpolation. Uses a method based on active geodesic contours. Segmentation effort is based on a variant of Voronoi diagram. Delaunay triangles dual to the Voronoi map is used. Estimation uses Bayesian formulation. The images are categorized based on statistical shape models using Leventon. Framework is implemented using the National Library of Medicine's (NIH/NLM) Insight Segmentation and Registration Toolkit (ITK) and the Visualization Toolkit (VTK) from Kitware Inc. 2.5GHz Pentium machines running Linux with 1GB main memory. [119]
ArchPRED (2006)	An online application for predicting loop conformations. Algorithm selects candidate loop fragments, exhaustive conformational fragments, from a multidimensional library called "Search Space". The library is updated by analyzing all structures in Protein Data Bank (PDB). Structures are organized in a three level hierarchy and defined by DSSP. Basic method steps are Selection :

	<p>selection conditions are length of loop, motif geometry, distance of ending points, Filtering: apply filters such as fit accessed by Root Mean Square Deviation (RMSD) of stem regions, steric fitting, Ranking: composite Z-score. Provides a higher coverage compared to FREAD. Implemented on an Apache server running Fedora core 3 operating system. CGI Perl and java script coded web interface. Data stored in a MySQL relational database, DBI–DBD (Database Interface–Database Driver). [120]</p>
--	--

5. CONCLUSIONS

In this study through a literature review approaches based on molecular cavity detection, analysis and visualization have been reviewed and organized.

The primary target was to find the tools used in these approaches and give a brief description of the algorithms and methods they perform on cavity detection and analysis. Several tools have been developed to date.

Direct comparison of the results obtained from the literature research with the work previously performed by Mr. Fotopoulos and Mr. Papathanasiou showed that multidimensional k-means clustering is a competitive advantage. This algorithm uses superposition of the protein and does not compare pairwise differences, leading to a faster computation. It should be noted that a one-to-one residue matching is a greedy approach and more time-consuming.

Performance and accuracy are crucial in the protein prediction process. Multiple structural features are included in prediction calculation and an approach like superposition of a protein, using secondary elements such as α -helices and β -sheets, show a strong predictive performance and is less time-consuming. 3D visualization techniques are used to analyze workflows. Color coded distance maps display interactions and are more user friendly. No tool was found that visualizes structural boundaries using polygon shapes.

It is concluded that some tools such as Scikit-learn and UCSF Chimera provide a wide variety of methods, techniques and algorithms. In general, a platform that offers a package of tools can be most helpful in drug design approaches.

The results of the present study confirm that molecular topology and molecular geometry are important as it has been previously reported in the literature. Structure, shape, location, distances and angles of binding sites may affect ligand binding affinity. A number of factors also, most of which are interconnected affect the binding site process. Since different parameters should be thoroughly investigated further analysis of the impact that alteration of each of these may have is required in order to assess whether a binding pocket should be targeted for applying the method of structure-based ligand design.

APPENDIX I

DRUG DEVELOPMENT & PROTEIN STRUCTURE

Drugs are chemicals and affect each body differently. A major problem is their side effects. Drug design is focused on this part.

Essentially a drug is a molecule that bounds in different cell areas, that is why drug design is also known as ligand design.

It enables or disables a biomolecular function such as a protein, receptor, enzyme, channel etc.

The objective of drug design is to find, meaning predict, and binding affinity.

Two approaches exist a) computer-based and b) structure-based drug design.

The computational approach is referred to as **computer-aided drug design**. The other one depends on the 3D biomolecular target and is referred to as **structure-based drug design**.

Their purpose of computer-aided drug design is to upgrade features like affinity, selectivity and stability of proteins. It estimates how strong a molecule binds to the target. This process is called molecular dynamics.

The activation or inhibition of a biomolecular function requires analysis of pathways of a disease or pathology condition. The target molecule is under investigation.

Firstly we need to find a target molecule that is disease or pathology related and check if it is “druggable”.

These molecules can be found in libraries of potential drug compounds. The purified protein can be visualized. Scoring methods are performed for evaluation. Scoring methods are linear regression, machine learning, and neural networks. The consensus scoring and cluster analysis play an important key role.

Structure-based drug design as the name itself says is based on the 3D protein structure. Receptors can be found in 3D structures of small molecules libraries. Docking programs check and find new binding pockets. Known ligands can be optimized by evaluating binding cavity. [122]

It is helpful to mention some definitions of molecular geometry and protein structure.

Molecular geometry is the 3D placement of the atoms that form a molecule.

Covalent bonds held molecules like a glue.

Types of bonds are shown in Figure 7.

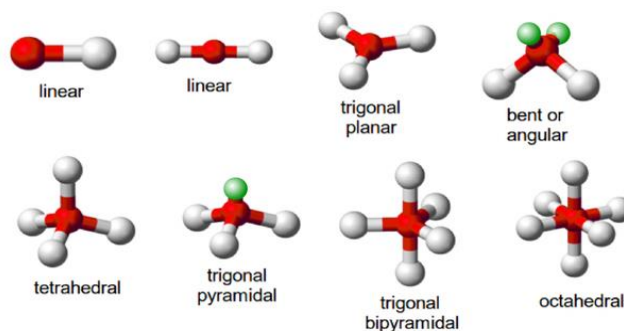


Figure 7: molecule bonds [123]

Isomers are types of molecules that use a chemical formula and do have different geometry viz different properties. [123]

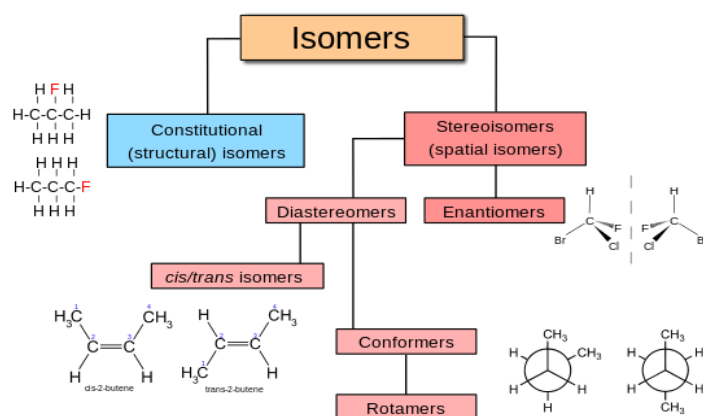


Figure 8: different types of isomers [123]

Steric collisions do forbid some combinations of the phi and psi angles of an alpha carbon atom. These combinations are represented in a two dimensional plot knows as **Ramachandran plot**. [2]

Protein Structure

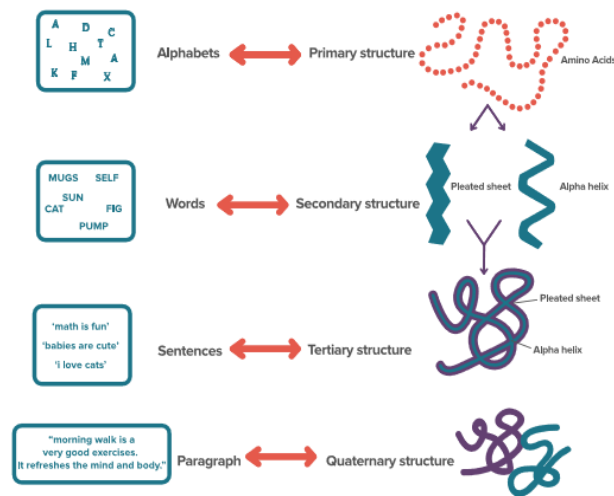


Figure 9: protein structure [3]

Proteins are made up of chains of amino acids and they determine protein's shape and interaction.

There are 20 common amino acids.

They consist of an amino group, a carboxyl group and an alpha carbon. On the alpha carbon bonds a group known as R group that defines the variety of amino acids. This side chain influences the protein structure.

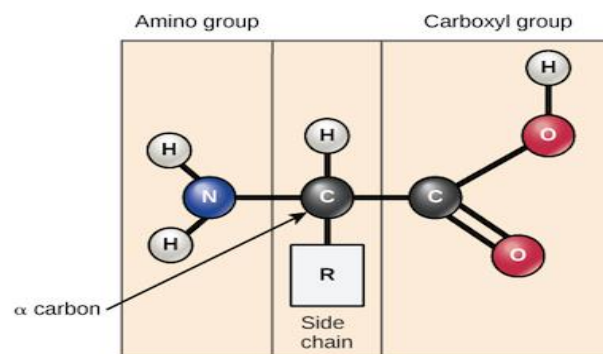


Figure 10: amino acid [3]

A peptide is an ensemble of amino acids and they are connected by peptide bonds. The peptide bonds can rotate.

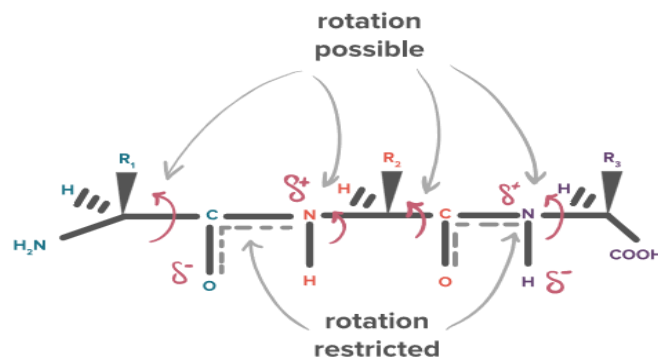


Figure 11: protein rotation [3]

Four level of protein structure do exist. The first one is the **primary structure** and it is the sequence of the amino acids. How a polypeptide is bend to do the things that it is need to do. The **secondary structure** is due to interactions of the peptide backbone. Beta pleated sheet. The backbone is going in a helical structure. Hydrogen bonds are between the different layers of the helix. So called a-helix.

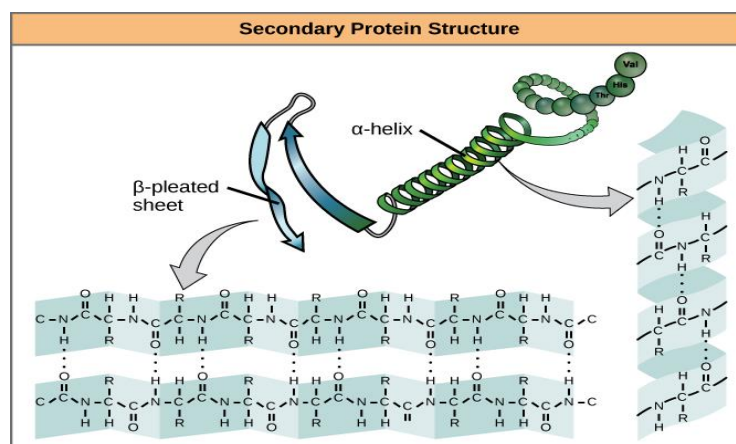


Figure 12: secondary structure [3]

The **tertiary structure** is due to interactions of side chains.

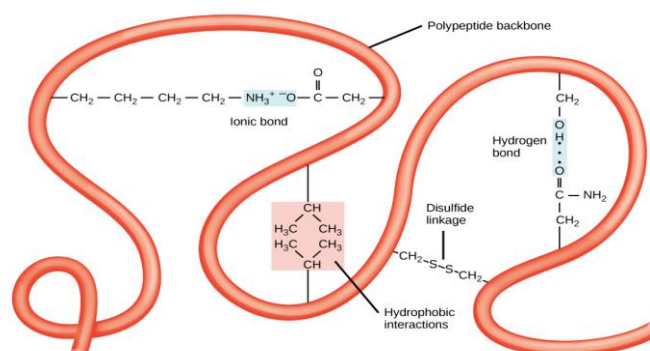


Figure 13: tertiary structure [3]

For more than one polypeptide this is the **quaternary structure** how the come together. Arrangement of multi-peptide chain.

REFERENCES

- [1] Development of data mining tools for identifying structural determinants that dictate protein-ligand interactions Anaxagoras A. Fotopoulos, Athanasios V. Papathanasiou MASTER-THESIS June 2015.
- [2] University of Waterloo-Computational Techniques in in Structural Bioinformatics.
- [3] Khan Academy-Free Online Courses <https://www.khanacademy.org/> [accessed 10/07/17].
- [4] Bienfait B., Ertl P., 013, "JSME: a free molecule editor in JavaScript", J Cheminform, 2013, doi: 10.1186/1758-2946-5-24.
- [5] Ye Fang, Yun Ding, Wei P. Feinstein, David M. Koppelman, Juana Moreno, J. Ramanujam, Michal Brylinski, "GeauxDock: Accelerating Structure-Based Virtual Screening with Heterogeneous Computing", PLoS One, 2016, doi: 10.1371/journal.pone.0158898.
- [6] Xie T, Song S, Li S, Ouyang L, Xia L, Huang J., "Review of natural product databases", Cell Prolif., 2015, doi: 10.1111/cpr.12190.
- [7] Hannu Raunio, Mira Kuusisto, Risto O. Juvonen, Olli T. Pentikäinen, "Modeling of interactions between xenobiotics and cytochrome P450 (CYP) enzymes", Front. Pharmacol., 2015, doi: 10.3389/fphar.2015.00123.
- [8] Active Site https://en.wikipedia.org/wiki/Active_site [accessed 10/07/17].
- [9] Computational Approaches for Studying Enzyme Mechanism, Part 2 [book].
- [10] Schmidtke P, Le Guilloux V, Maupetit J, Tufféry P., "fpocket: online tools for protein ensemble pocket detection and tracking", Nucleic Acids Res., 2010, doi: 10.1093/nar/gkq383.
- [11] Tanya Singh, D.Biswas, B.Jayaram, "AADS - An Automated Active Site Identification, Docking, and Scoring Protocol for Protein Targets Based on Physicochemical Descriptors", J.Chem.Inf.Model., 2011, doi: 10.1021/ci200193z.
- [12] Yitav Glantz-Gashai, Tomer Meirson, Abraham O.Samson, "Normal Modes Expose Active Sites in Enzymes", PLOS Comp.Biol., 2016, doi: 10.1371/journal.pcbi.1005293.
- [13] Miguel D.Toscano, Kenneth J.Woycechowsky, Donald Hilvert, "Minimalist Active-Site Redesign:Teaching Old Enzymes New Tricks", Angew.Chem.Int.Ed., 2007, doi: 10.1002/anie.200604205.
- [14] Doctor Thesis of Noe Sturm, "Biosynthetic moldings give potent biological activities to natural products", Eskitis Institute for Drug Discovery, Griffith University, 2015.
- [15] Sing Mei Lim et al., "Structural and dynamic insights into substrate binding and catalysis of human lipocalin prostaglandin D synthase", Journal of Lipid Research, 2013, doi: 10.1194/jlr.M035410.
- [16] Laura DIAZ-SAEZ, Velupillai SRIKANNATHASAN, Martin ZOLTNER, William N.HUNTER, "Structures of bacterial kynurenine formamidase reveal a crowded binuclear zinc catalytic site primed to generate a potent nucleophile", Biochem.J., 2014, doi: 10.1042/BJ20140511.
- [17] Katia D'Ambrosio, Simone Carradori, Simona M.Monti, Martina Buonanno, Daniela Secci, Daniela Vullo, Claudiu T.Supuran, Giuseppina De Simone, "Out of the active site binding pocket for carbonic anhydrase inhibitors", Chem.Comm., 2015, doi: 10.1039/C4CC07320G.
- [18] Robert Roskoski Jr, "Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes", ELSEVIER, 2015, doi: 10.1016/j.phrs.2015.10.021.
- [19] Enrique Marcos et al., "Principles for designing proteinswith cavities formed by curved b sheets", Science, 2017, doi: 10.1126/science.aah7389.
- [20] Labbé CM, Kuenemann MA, Zarzycka B, Vriend G, Nicolaes GA, Lagorce D, Miteva MA, Villoutreix BO, Sperandio O, "iPPI-DB: an online database of modulators of protein-protein interactions", Nucleic Acids Res., 2016, doi: 10.1093/nar/gkv982.
- [21] Yong-Cui Wang, Yong-Cui Wang, Nai-Yang Deng, Yong Wang, "Computational probing protein-protein interactions targeting small molecules", Bioinformatics, 2016, doi: 10.1093/bioinformatics/btv528.
- [22] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP., "ChEMBL: a large-scale bioactivity database for drug discovery", Nucleic Acids Res., 2012, doi: 10.1093/nar/gkr777.
- [23] Saladin A, Rey J, Thévenet P, Zacharias M, Moroy G, Tufféry P., "PEP-SiteFinder: a tool for the blind identification of peptide binding sites on protein surfaces", Nucleic Acids Res., 2014, doi: 10.1093/nar/gku404.
- [24] Hasup Lee, Lim Heo, Myeong Sup Lee, Chaok Seok, "GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization", Nucleic Acids Res., 2015, doi: 10.1093/nar/gkv495.
- [25] Ashish Kabra, Salman Shahid, Ravi Kant Pal, Rahul Yadav, S.V.S. Rama Krishna Pulavarti, Anupam Jain, Sarita Tripathi, Ashish Arora, "Unraveling the stereochemical and dynamic aspects of the catalytic site of bacterial peptidyl-tRNA hydrolase", 2016, doi: 10.1261/rna.057620.116.

- [26] Thomas Fober, Marco Mernberger, Gerhard Klebe, EykeHullermeier, "Efficient Similarity Retrieval for Protein Binding Sites based on Histogram Comparison", Department of Mathematics and Computer Science.
- [27] Pankaj Kumar et al., "Non-classical transpeptidases yield insight into new antibacterials", *Nature Chemical Biology*, 2016, doi: 10.1038/NCHEMBIO.2237.
- [28] Wallace K.B.Chan, Hongjiu Zhang, Jianyi Yang, Jeffrey R.Brender, Junguk Hur, Arzucan Özgür, Yang Zhang, "GLASS: a comprehensive database for experimentally validated GPCR-ligand associations", *Bioinformatics*, 2015, doi: 10.1093/bioinformatics/btv302.
- [29] Lee GR, Seok C., "Galaxy7TM: flexible GPCR-ligand docking by structure refinement", *Nucleic Acids Res.*, 2016, doi: 10.1093/nar/gkw360.
- [30] Costello et al., "A community effort to assess and improve drug sensitivity prediction algorithms", *Nat Biotechnol.*, 2014, doi: 10.1038/nbt.2877.
- [31] Bui QC, Sloot PM, van Mulligen EM, Kors JA., "A novel feature-based approach to extract drug-drug interactions from biomedical text", *Bioinformatics*, 2014, doi: 10.1093/bioinformatics/btu557.
- [32] Santiago Vilar, Eugenio Uriarte, Lourdes Santana, Tal Lorberbaum, George Hripcsak, Carol Friedman, Nicholas P Tatonetti, "Similarity-based modeling in large-scale prediction of drug-drug interactions", *Nat Protoc.*, 2014, doi: 10.1038/nprot.2014.151.
- [33] Pan Tong, Kevin R. Coombes, Faye M. Johnson, Lauren A. Byers, Lixia Diao, Diane D. Liu, J. Jack Lee, John V. Heymach, Jing Wang, "drexplorer: A tool to explore dose-response relationships and drug-drug interactions", *Bioinformatics*, 2015, doi: 10.1093/bioinformatics/btv028.
- [34] He L, Wennerberg K, Aittokallio T, Tang J., "TIMMA-R: an R package for predicting synergistic multi-targeted drug combinations in cancer cell lines or patient-derived samples", *Bioinformatics*, 2015, doi: 10.1093/bioinformatics/btv067.
- [35] Yang F, Xu J, Zeng J, "Drug-target interaction prediction: databases, web servers and computational models", *Brief Bioinform.*, 2016, PMID: PMC4730876.
- [36] Di Veroli GY, Fornari C, Wang D, Mollard S, Bramhall JL, Richards FM, Jodrell DI., "Combeneft: an interactive platform for the analysis and visualization of drug combinations", *Bioinformatics*, 2016, doi: 10.1093/bioinformatics/btw230.
- [37] Bansal et al., "A community computational challenge to predict the activity of pairs of compounds", *Nat Biotechnol.*, 2014, doi: 10.1038/nbt.3052.
- [38] Liu H, Hou T., "CaFE: a tool for binding affinity prediction using end-point free energy methods", *Bioinformatics*, 2016, doi: 10.1093/bioinformatics/btw215.
- [39] Thomas Sander, Joel Freyss, Modest von Korff, Christian Rufener, "DataWarrior: an open-source program for chemistry aware data visualization and analysis", *J Chem Inf Model*, 2015, doi: 10.1021/ci500588j.
- [40] Shu-Dong Zhang, Timothy W Gant, "sscMap: an extensible Java application for connecting small-molecule drugs using gene-expression signatures", *BMC Bioinformatics*, 2009, doi: 10.1186/1471-2105-10-236.
- [41] Daniel J.Rigden, "From Protein Structure to Function with Bioinformatics", Springer, 2017, doi: 10.1007/978-94-024-1069-3.
- [42] Douglas E. V. Pires, Tom L. Blundell, David B. Ascher, "pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures", *J Med Chem.*, 2015, doi: 10.1021/acs.jmedchem.5b00104.
- [43] Shao CY, Su BH, Tu YS, Lin C, Lin OA, Tseng YJ, "CypRules: a rule-based P450 inhibition prediction server", *Bioinformatics*, 2015, doi: 10.1093/bioinformatics/btv043.
- [44] Jed Zaretski, Charles Bergeron, Tao-wei Huang, Patrik Rydberg, S. Joshua Swamidass, Curt M. Breneman, "RS-WebPredictor: a server for predicting CYP-mediated sites of metabolism on drug-like molecules", *Bioinformatics*, 2013, doi: 10.1093/bioinformatics/bts705.
- [45] Hannu Raunio, Mira Kuusisto, Risto O.Juvonen, Olli T.Pentikäinen, "Modeling of interactions between xenobiotics and cytochrome P450 (CYP) enzymes", 2015, doi: 10.3389/fphar.2015.00123.
- [46] Legehar A, Xhaard H, Ghemtio L., "IDAAPM: integrated database of ADMET and adverse effects of predictive modeling based on FDA approved drug data", *J Cheminform*, 2016, doi: 10.1186/s13321-016-0141-7.
- [47] Lim CN, Liang S, Feng K, Chittenden J, Henry A, Mouksassi S, Birnbaum AK., "PhxnIme: An R package that facilitates pharmacometric workflow of Phoenix NLME analyses", *Comput Methods Programs Biomed.*, 2017, doi: 10.1016/j.cmpb.2016.12.002.
- [48] Tetko et al., "Prediction of logP for Pt(II) and Pt(IV) complexes: Comparison of statistical and quantum-chemistry based approaches", *J Inorg Biochem*, 2016, doi: 10.1016/j.jinorgbio.2015.12.006.
- [49] Oberhauser N, Nurisso A, Carrupt PA., "MLP Tools: a PyMOL plugin for using the molecular lipophilicity potential in computer-aided drug design", *J Comput Aided Mol Des.*, 2014, doi: 10.1007/s10822-014-9744-0.
- [50] Arwa B. Raies, Vladimir B. Bajic, "In silico toxicology: computational methods for the prediction of chemical toxicity", *Wiley Interdiscip Rev Comput Mol Sci.*, 2016, doi: 10.1002/wcms.1240.

- [51] Horvath M.P. et al., "Structure of the lutein-binding domain of human StARD3 at 1.74 resolution and model of a complex with lutein", *Struct.Biol.Comm.*, 2016, doi: 10.1107/S2053230X16010694.
- [52] Anthony Nash, Helen L.Birch, Nora H.de Leeuw, "Mapping intermolecular interactions and active site conformations: from human MMP-1 crystal structure to molecular dynamics free energy calculations", *Journal of Biomolecular Structure and Dynamics*, 2016, doi: 10.1080/07391102.2016.1153521.
- [53] Maciej Wójcikowski, Piotr Zielenkiewicz, Pawel Siedlecki, "Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field", *J Cheminform.*, 2015, doi: 10.1186/s13321-015-0078-2.
- [54] Moreno et al., "BiNChE: a web tool and library for chemical enrichment analysis based on the ChEBI ontology", *BMC Bioinformatics*, 2015, doi: 10.1186/s12859-015-0486-3.
- [55] <https://pc1664.pharmazie.uni-marburg.de:433/index.php?topic=research> [accessed 10/07/17].
- [56] <http://www.caver.cz/index.php?sid=133> [accessed 10/07/17].
- [57] Xiao-chen Bai, Eeson Rajendra, Guanghui Yang, Yigong Shi, Sjors HW Scheres, "Sampling the conformational space of the catalytic subunit of human γ -secretase", 2015, doi: 10.7554/eLife.11182.
- [58] Master Thesis of Markus Lumipuu, "Computer-aided identification of the binding sites of protein-ligand complexes", UNIVERSITY OF EASTERN FINLAND, 2013.
- [59] <https://www.flashcardmachine.com/bio-200.html> [accessed 10/07/17].
- [60] Peter Schmidtke, Axel Bidon-Chanal, F.Javier Luque Xavier Barril, "MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories", *Bioinformatics*, 2011, doi: 10.1093/bioinformatics/btr550.
- [61] https://www.tutorialspoint.com/data_mining/ [accessed 24/07/17].
- [62] Muhammad Aamer Mehmood, Ujala Sehar and Niaz Ahmad, "Use of Bioinformatics Tools in Different Spheres of Life Sciences", *J Data Mining Genomics Proteomics*, 2014, doi: 10.4172/2153-0602.1000158.
- [63] Jae-Kwan Kim, Youngsong Cho, Mokwon Lee, Roman A.Laskowski, Seong Eon Ryu, Kokichi Sugihara, Deok-Soo Kim, "BetaCavityWeb: a webserver for molecular voids and channels", *Nucleic Acids Res.*, 2015, doi: 10.1093/nar/gkv360.
- [64] Antonia Stank, Daria B. Kokh, Max Horn, Elena Sizikova, Rebecca Neil, Joanna Panecka, Stefan Richter, Rebecca C. Wade, "TRAPP webserver: predicting protein binding site flexibility and detecting transient binding pockets", *Nucleic Acids Res.*, 2017, doi: 10.1093/nar/gkx277.
- [65] <http://trapp.h-its.org/> [accessed 10/07/17].
- [66] Czirájk G., "PrinCCes: Continuity-based geometric decomposition and systematic visualization of the void repertoire of proteins", *J Mol Graph Model*, 2015, doi: 10.1016/j.jmgm.2015.09.013.
- [67] Elías D. López, Juan Pablo Arcon, Diego F. Gauto, Ariel A. Petruk, Carlos P. Modenutti, Victoria G. Dumas, Marcelo A. Marti, Adrian G. Turjanski, "WATCLUST: a tool for improving the design of drugs based on protein-water interactions", *Bioinformatics*, 2015, doi: 10.1093/bioinformatics/btv411.
- [68] Hiba Abi Hussein, Alexandre Borrel, Colette Geneix, Michel Petitjean, Leslie Regad, Anne-Claude Camproux, "PockDrug-Server: a new web server for predicting pocket druggability on holo and apo proteins", *Nucleic Acids Res.*, 2015, doi: 10.1093/nar/gkv462.
- [69] Paramo T, East A, Garzón D, Ulmschneider MB, Bond PJ., "Efficient Characterization of Protein Cavities within Molecular Simulation Trajectories: trj_cavity", *J Chem Theory Comput.*, 2014, doi: 10.1021/ct401098b.
- [70] Dong Xu, Hua Li, Yang Zhang, "Protein Depth Calculation and the Use for Improving Accuracy of Protein Fold Recognition", *J Comput Biol.*, 2013, doi: 10.1089/cmb.2013.0071.
- [71] Lee PH, Helms V., "Identifying continuous pores in protein structures with PROPORES by computational repositioning of gating residues", *Proteins*, 2012, doi: 10.1002/prot.23204.
- [72] Olechnovic K, Margelevicius M, Venclovas C., "Voroprot: an interactive tool for the analysis and visualization of complex geometric features of protein structure", *Bioinformatics*, 2011, doi: 10.1093/bioinformatics/btq720.
- [73] Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA., "Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure", *PLoS Comput Biol.*, 2009, doi: 10.1371/journal.pcbi.1000585.
- [74] Pellegrini-Calace M, Maiwald T, Thornton JM., "PoreWalker: a novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure", *PLoS Comput Biol.*, 2009, doi: 10.1371/journal.pcbi.1000440.
- [75] Nayal M, Honig B., "On the nature of cavities on protein surfaces: application to the identification of drug-binding sites", *Proteins*, 2006, doi: 10.1002/prot.20897.
- [76] B, Jonassen I., "J-Express: exploring gene expression data using Java", *Bioinformatics*, 2001, PMID: 11301307.
- [77] Ayyala DN, Lin S., "GrammR: graphical representation and modeling of count data with application in metagenomics", *Bioinformatics*, 2015, doi: 10.1093/bioinformatics/btv032.

- [78] Renxiang Yan, Dong Xu, Jianyi Yang, Sara Walker, Yang Zhang, "A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction", *SCIENTIFIC REPORTS*, 2013, doi: 10.1038/srep02619.
- [79] João P. A. Moraes, Gisele L. Pappa, Douglas E. V. Pires, Sandro C. Izidoro, "GASS-WEB: a web server for identifying enzyme active sites based on genetic algorithms", *Nucleic Acids Research*, 2017, doi: 10.1093/nar/gkx337.
- [80] Jaroslav Bendl et al., "HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering", *Nucleic Acids Res.*, 2016, doi: 10.1093/nar/gkw416.
- [81] Ge-Fei Hao et al., "ACFIS: a web server for fragment-based drug discovery", *Nucleic Acids Res.*, 2016, doi: 10.1093/nar/gkw393.
- [82] Nadav Brandes, Dan Ofer, Michal Linial, "ASAP: a machine learning framework for local protein properties", *Database (Oxford)*, 2016, doi: 10.1093/database/baw133.
- [83] <http://scikit-learn.org/stable/index.html> [accessed 10/07/17].
- [84] Xiao X, Hui MJ, Liu Z, Qiu WR., "iCataly-PseAAC: Identification of Enzymes Catalytic Sites Using Sequence Evolution Information with Grey Model GM (2,1)", *J Membr Biol.*, 2015, doi: 10.1007/s00232-015-9815-8.
- [85] David Sehnal, Lukáš Pravda, Radka Svobodová Vařeková, Crina-Maria Ionescu, Jaroslav Koča, "PatternQuery: web application for fast detection of biomacromolecular structural patterns in the entire Protein Data Bank", *Nucleic Acids Res.*, 2015, doi: 10.1093/nar/gkv561.
- [86] Hung le V, Caprari S, Bizai M, Toti D, Polticelli F., "LIBRA: LIgand Binding site Recognition Application", *Bioinformatics*, 2015, doi: 10.1093/bioinformatics/btv489.
- [87] Jonathan E. Chen, Conrad C. Huang, Thomas E. Ferrin, "RRDistMaps: a UCSF Chimera tool for viewing and comparing protein distance maps", *Bioinformatics*, 2015, doi: 10.1093/bioinformatics/btu841.
- [88] <https://www.cgl.ucsf.edu/chimera/> [accessed 10/07/17].
- [89] Zhi-Ping Liu, Shutang Liu, Ruitang Chen, Xiaopeng Huang, Ling-Yun Wu, "Structure alignment-based classification of RNA-binding pockets reveals regional RNA recognition motifs on protein surfaces", *BMC Bioinformatics*, 2017, doi: 10.1186/s12859-016-1410-1.
- [90] Charles C. David, Ettayapuram Ramaprasad Azhagiya Singam, Donald J. Jacobs, "JED: a Java Essential Dynamics Program for comparative analysis of protein trajectories", *BMC Bioinformatics*, 2017, doi: 10.1186/s12859-017-1676-y.
- [91] Votapka LW, Jagger BR, Heyneman AL, Amaro RE., "SEEKR: Simulation Enabled Estimation of Kinetic Rates, A Computational Tool to Estimate Molecular Kinetics and Its Application to Trypsin-Benzamidine Binding", *J Phys Chem B.*, 2017, doi: 10.1021/acs.jpcc.6b09388.
- [92] <https://amarolab.ucsd.edu/seekr/> [accessed 10/07/17].
- [93] Peterson L, Jamroz M, Kolinski A, Kihara D., "Predicting Real-Valued Protein Residue Fluctuation Using FlexPred", *Methods Mol Biol.*, 2017, doi: 10.1007/978-1-4939-6406-2_13.
- [94] Mueller SC, Backes C, Gress A, Baumgarten N, Kalinina OV, Moll A, Kohlbacher O, Meese E, Keller A., *Bioinformatics*. 2016, doi: 10.1093/bioinformatics/btw084.
- [95] Michael Feig, "Local Protein Structure Refinement via Molecular Dynamics Simulations with locPREFMD", *J Chem Inf Model*, 2016, doi: 10.1021/acs.jcim.6b00222.
- [96] Doerr S, Harvey MJ, Noé F, De Fabritiis G., "HTMD: High-Throughput Molecular Dynamics for Molecular Discovery", *J Chem Theory Comput.*, 2016, doi: 10.1021/acs.jctc.6b00049.
- [97] Matthew P Harrigan et al., "MSMBuilder: Statistical Models for Biomolecular Dynamics", *bioRxiv*, 2016, doi: 10.1101/084020.
- [98] Krupa P et al., "Performance of protein-structure predictions with the physics-based UNRES force field in CASP11", *Bioinformatics*, 2016, doi: 10.1093/bioinformatics/btw404.
- [99] Hongchun Li, Yuan-Yu Chang, Lee-Wei Yang, Ivet Bahar, "iGNM 2.0: the Gaussian network model database for biomolecular structural dynamics", *Nucleic Acids Res.*, 2016, doi: 10.1093/nar/gkv1236.
- [100] David A. Clifton, Samuel Hugueny, Lionel Tarassenko, "A comparison of approaches to multivariate extreme value theory for novelty detection", *IEEE Workshop on Statistical Signal Processing*, 2009, doi: 10.1109/SSP.2009.5278652.
- [101] Zhichao Miao, Yang Cao, "Quantifying side-chain conformational variations in protein structure", *Scientific Reports*, 2016, doi: 10.1038/srep37024.
- [102] Maia EH, Campos VA, Dos Reis Santos B, Costa MS, Lima IG, Greco SJ, Ribeiro RI, Munayer FM, da Silva AM, Taranto AG., "Octopus: a platform for the virtual high-throughput screening of a pool of compounds against a set of molecular targets", *J Mol Model.*, 2017, doi: 10.1007/s00894-016-3184-9.
- [103] Sergio Ruiz-Carmona et al., "Dynamic undocking and the quasi-bound state as tools for drug discovery", *Nature Chemistr*, 2016, doi: 10.1038/nchem.2660.
- [104] Jun Gao, Qingchen Zhang, Min Liu, Lixin Zhu, Dingfeng Wu, Zhiwei Cao, Ruixin Zhu, "bSiteFinder, an improved protein-binding sites prediction server based on structural alignment: more accurate and less time-consuming", *J Cheminform.*, 2016, doi: 10.1186/s13321-016-0149-z.

- [105] Yuanbao Ai, Lingling Yu, Xiao Tan, Xiaoying Chai, Sen Liu, "Discovery of Covalent Ligands via Noncovalent Docking by Dissecting Covalent Docking Based on a "Steric-Clashes Alleviating Receptor (SCAR)" Strategy", *J. Chem. Inf. Model.*, 2016, doi: 10.1021/acs.jcim.6b00334.
- [106] Smieško M, Vedani A, "VirtualToxLab: Exploring the Toxic Potential of Rejuvenating Substances Found in Traditional Medicines", *Methods Mol Biol.*, 2016, doi: 10.1007/978-1-4939-3609-0_7.
- [107] Hu B, Lill MA., "WATsite: hydration site prediction program with PyMOL interface", *J Comput Chem.*, 2014, doi: 10.1002/jcc.23616.
- [108] Emmanouil Athanasiadis, Zoe Cournia, George Spyrou, "ChemBioServer: a web-based pipeline for filtering, clustering and visualization of chemical compounds used in drug discovery", *George Spyrou, Bioinformatics*, 2012, doi: 10.1093/bioinformatics/bts551.
- [109] QuikProp User Manual
- [110] <http://www.vcclab.org/lab/alogps/> (accessed 10/07/17).
- [111] Desmond User Manual.
- [112] Procacci P., "PrimaDORAC: A Free Web Interface for the Assignment of Partial Charges, Chemical Topology, and Bonded Parameters in Organic or Drug Molecules", *J Chem Inf Model.*, 2017, doi: 10.1021/acs.jcim.7b00145.
- [113] Ahmad N. Shuid, Robert Kempster, Liam J. McGuffin, "ReFOLD: a server for the refinement of 3D protein models guided by accurate quality estimates", *Nucleic Acids Research*, 2017, <https://doi.org/10.1093/nar/gkx249>.
- [114] Skjærven L, Jariwala S, Yao XQ, Grant BJ., "Online interactive analysis of protein structure ensembles with Bio3D-web", *Bioinformatics*, 2016, doi: 10.1093/bioinformatics/btw482.
- [115] Liu W, Xie Y, Ma J, Luo X, Nie P, Zuo Z, Lahrmann U, Zhao Q, Zheng Y, Zhao Y, Xue Y, Ren J., "IBS: an illustrator for the presentation and visualization of biological sequences", *Bioinformatics*, 2015, doi: 10.1093/bioinformatics/btv362.
- [116] <http://www.jzy3d.org/> [accessed 27/07/17].
- [117] Zhou H, Skolnick J., "GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction", *Biophys J.*, 2011, doi: 10.1016/j.bpj.2011.09.012.
- [118] Sindi S, Helman E, Bashir A, Raphael BJ., "A geometric approach for classification and comparison of structural variants", *Bioinformatics*, 2009, doi: 10.1093/bioinformatics/btp208.
- [119] Mosaliganti K, Cooper L, Sharp R, Machiraju R, Leone G, Huang K, Saltz J., "Reconstruction of cellular biological structures from optical microscopy data", *IEEE Trans Vis Comput Graph*, 2008, doi: 10.1109/TVCG.2008.30.
- [120] Narcis Fernandez-Fuentes, Jun Zhai, Andrés Fiser, "ArchPRED: a template based loop structure prediction server", *Nucleic Acids Res*, 2006, doi: 10.1093/nar/gkl113.
- [121] DNAnexus <https://www.dnanexus.com/company> [accessed 27/07/17].
- [122] Drug design https://en.wikipedia.org/wiki/Drug_design [accessed 10/07/17].
- [123] Molecular geometry https://en.wikipedia.org/wiki/Molecular_geometry [accessed 10/07/17].
- [124] Furnham N, Holliday GL, de Beer TA, Jacobsen JO, Pearson WR, Thornton JM, "The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes", *Nucleic Acids Res.*, 2014, doi: 10.1093/nar/gkt1243.