Masaryk University
Faculty of Informatics

Visualization and Visual Analysis of Intermolecular Interactions of Proteins

Rigorous Thesis

Katarína Furmanová

Advisor: prof. Jiří Sochor
Co-Advisor: assoc. prof. Barbora Kozlíková

Brno, Spring 2017

Signature of Thesis Advisor

Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out on my own. All sources, references, and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Katarína Furmanová

Advisor: prof. Jiří Sochor
Co-Advisor: assoc. prof. Barbora Kozlíková

Acknowledgement

Abstract

Keywords

protein, protein-protein interactions, visualization, cavity, protein void, tunnel, contact zone, CAVER Analyst

Contents

1 Introduction

Proteins are highly complex macromolecules that are vital to biochemical processes taking place in each living organism. Weather alone or as a part of multi-unit complexes, they facilitate vast field of functions such as catalysing chemical reactions, transporting molecules across the cells or replication of DNA. In these processes the ability of a protein to interact with other molecules plays a defining role.

Since the proper understanding of protein interactions contributes to advances in medicine, pharmaceutics or even agriculture, the study of interaction patterns of proteins has been at the forefront of biochemical research for decades. Unfortunately, the complexity of protein structures and the necessity for expensive and time consuming in-vitro experiments make the progress in the area slow. Many computational tools aim to support this research by simulating the experiments in-silico and thus reducing the costs. However, these tools can produce a vast amounts of data. For example molecular dynamics simulations can mimic the movement of millions of atoms over a given period of time. It is virtually impossible to identify significant patterns by simply observing such simulation. Another example are the protein-protein docking simulations that predict the possible ways two or more proteins interact together. Here the output often comprises of tens to hundreds of possible conformations that the domain expert needs to analyse individually one by one.

Therefore, visualization and visual analysis tools became inherent part of proteomic research both as guidance during the experiments as well as for validation and analysis of results by the domain experts. The main aim of these tools is to speed up the analysis process by - often interactively - extracting the important features of the data and conveying them in such way, that previously hardly observable patterns and relationships become more prominent. Although much has been done in the field of molecular visualization in the past decades, there are still areas and problems that are not currently addressed.

1.1 Biochemical Definitions

Although this thesis deals with the research in the field of visualization and visual analysis, it also ventures into the field of biochemistry. It is therefore inevitable to clarify the basic biochemical terms that will occur throughout

the thesis and are important for its proper understanding. This section shall provide the reader with all the necessary knowledge.

1.1.1 Protein Structures

Proteins are complex molecules formed by one or more chains of amino acids. Amino acids are basic building blocks of all living organisms. There are approximately 500 known amino acids, but only 20 standard amino acids are encoded in genetic code. Each of them consists of *carboxyl group* ($-COOH$), an *amino group* ($-NH_2$) and a unique *side chain* ($-R$) that defines its properties. The three groups are connected by a carbon atom, also called an *alpha carbon* C_α. See figure 1.1 a).

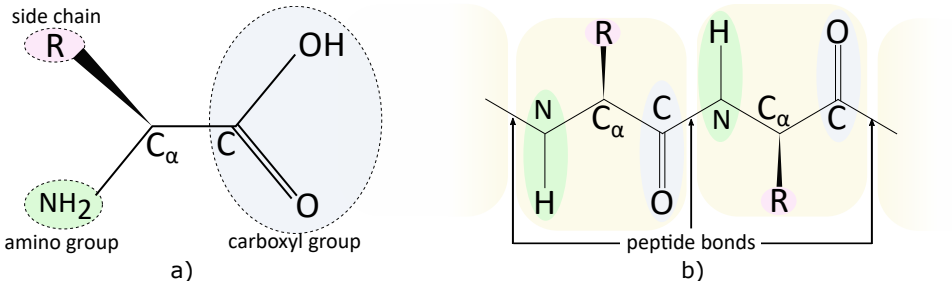

Figure 1.1: a) Illustration of a basic amino acid structure. b) Amino acid residues connected into polypeptide chain. It can be noted, that the amino and carboxyl groups are missing atoms, which were released during the formation of peptide bonds as H_2O molecules.

During a protein synthesis amino acids are joined together by peptide bonds (covalent bonds), forming polypeptide chains. A peptide bond is formed in a reaction between carboxyl group of one amino acid and amino group of another amino acid (see figure 1.1 b)). As both groups loose atoms that are released as molecule of water during this reaction, the amino acids bonded in polypeptide chains are refereed to as *amino acid residues*.

Each protein contains at least one long polypeptide chain. This sequence of amino acids, connected by rigid peptide bonds, also known as *backbone*, forms *primary structure* of the protein.

Unlike the peptide bonds, the bonds linking the carboxyl and amino groups to the alpha carbon are free to rotate. Based on these rotations and the patterns of hydrogen bonds that form between hydrogen from amino group and oxygen from corboxyl group, the segments of polypeptide chain can take on various 3D formations. The two most common of those are $\alpha -$

helices and $\beta - sheets$, which are formed by laterally connected $\beta - strands$. These local formations of polypeptide chain are called *secondary protein structures*. Parts of polypeptide chain with absent secondary structures are called *random coils*. See figure 1.2.

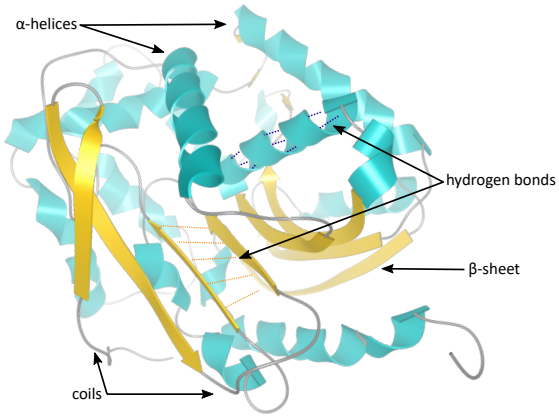

Figure 1.2: Typical secondary structures of protein: $\alpha - helices$ (blue), $\beta - strands$ (orange) forming $\beta - sheet$ and *coils*.

Various side chains of amino acid residues can interact together during the formation of protein. As a result, the secondary structures of the protein are bended and shaped into a unique 3D structure until the protein attains its minimal energy state. This process is called *protein folding* and it results in a *tertiary protein structure*. The tertiary structure defines the complete spatial arrangement of atoms of one polypeptide chain. Interactions between amino acids of multiple polypeptide chains than define their *quaternary protein structure*.

1.1.2 Properties of Proteins

Previous section described the process of protein attaining its 3D structure. This structure directly influences the way protein is behaving with regards to other molecules and its ability to function properly.

Example of this are the inner voids of the protein. When protein folds, there is naturally some empty space left inside. Depending on the shape of the space we classify four types of inner voids (figure 1.3): *cavities* – void space buried deeply inside the protein, *tunnels* – connecting cavities with surface of the protein, *channels* – passing through the whole protein and *pockets* – shallow dents on the surface of the protein.

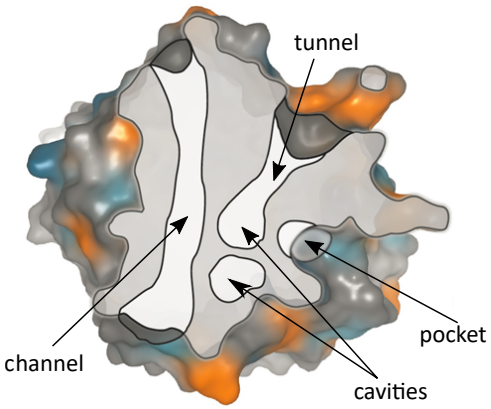

Figure 1.3: Types of inner voids of protein. Image adapted from [63]

These inner voids can significantly influence the reactivity of the protein since they contain *active site*. Active site is a region of reactive amino acids, where other smaller molecules can bind to protein and undergo a chemical reaction that changes their properties. This place is often buried deep inside the protein and its accessibility is thus limited by the size, shape and physico-chemical properties of the tunnels leading to it. However, the binding site can be located also in shallower pockets on protein surface. In several types of proteins, these binding sites serve for interacting with other proteins.

On the other hand, proteins containing channels (also called pores) occupy entirely different function. They are often found in the membranes of the cells, where the geometry and properties of the channels are responsible for regulating the molecules that can pass through the cell membrane. They are often specific to one type of molecule – e.g. water, and no other molecules can pass through them in or out of the cell.

As noted above, the reactivity and functions of the proteins are given by their geometry as well as by their physico-chemical properties. These properties are analogous to the properties of the their amino acids:

- *Polarity and Partial Charge*
  In a molecule of water, hydrogen atoms are bound to highly electronegative oxygen atom. The electronegativity of oxygen causes higher concentration of electrons on its side of hydrogen bonds and thus a separation of positive and negative electric charge (electric dipole). This phenomenon occurs also in several so called *polar* amino acids. The amount of separated charge is usually lower than fundamental charge, therefore it is called *partial charge*.

- *Donor / Acceptor*
  Amino acids participating in hydrogen bonds can be classified as *hydrogen donors* or *hydrogen bond donors* if they contain the hydrogen atoms participating in this bonds. Amino acids on the other side of the bond are called *hydrogen acceptors*. Note that amino acids participating in multiple hydrogen bonds can be donors and acceptors at the same time.
- *Hydrophobicity*
  Amino acids are called *hydrophobic* if they seemingly repel water. Unlike *hydrophilic* amino acids, they are not polar and thus cannot create bonds with polar molecules of water.

Most of the proteins contain hydrophobic amino acids at their core, while their surface is covered by polar amino acids. They are in contact with outer environment – *solvent*, where they can form hydrogen bonds.

So far, when discussing the properties of proteins, we have assumed the static 3D structure. However, due to constant physical forces taking place between millions of atoms of proteins and surrounding solvents, the structure of the protein is not static and when studying the proteins one has to consider so called *molecular dynamics (MD)*. This term generally denotes the simulation or the captured interval of atom movement that constantly changes not only the shape but consequently also properties of observed proteins.

1.2 Problem Formulation

Now that the reader is familiar with basic biochemical terminology, we can formulate the specific problems that will be the focus of this thesis. It was already hinted that proteins can participate in various kinds of intermolecular interactions. In this thesis we will focus on two typical types of interactions: a) protein-ligand interactions and b) protein-protein interactions.

1.2.1 Protein-Ligand Interactions

In biochemical terminology ligand denotes a small molecule that binds to a protein, where the consequential reaction changes both, the target protein as well as the ligand itself. Analysis of protein-ligand docking (the act of ligand travelling through the protein tunnel and binding to the active site) has application in different fields of biochemistry such as protein engineering or drug design. The typical goal of protein engineering research is changing of protein properties by mutating some of its amino acids to make it, e.g. more

stable under high temperature conditions or more reactive with a particular type of ligand. In drug design the goal is to find or adjust protein-ligand combinations, such that their mutual reaction would synthesize new drug from the ligand. However, in both cases the researchers are looking for the answers to the following questions:

- Can the ligand pass through the tunnel leading to the active site?
- If not, which parts of the tunnels are causing problems?
- Is it the geometrical bottleneck, that prevents the ligand from passing through the tunnel?
- Are the physico-chemical properties of the tunnel amino acids responsible for repelling the ligand from the active site?
- Can these problems can be resolved by mutating the protein amino acids?

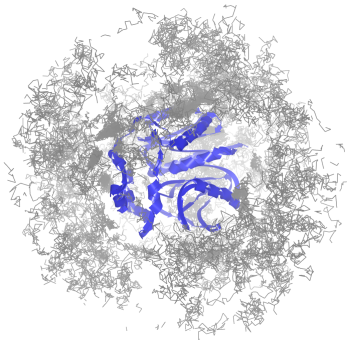

Figure 1.4: Ligand trajectory (gray) in a simulation containing 50 000 time steps. The protein chain is depicted in blue.

There is already great amount of published work aiming to answer these questions either by studying the tunnel properties or by directly simulating the transportation of the ligand to the active site. However, with the complexity of protein structures in combination with the ever-changing molecular dynamics, the answers are not trivial. Figure 1.4 for example depicts trajectory of a ligand in a MD simulation consisting of 50 000 time steps. It is apparent, that further analysis is necessary to identify significant parts and patterns in this simulation.

1.2.2 Protein-Protein Interactions

Most of the proteins responsible for various functions in cellular life are operating in larger multi-protein complexes. For example a family of SMC complexes (structural maintenance of chromosomes) govern the organisation of DNA in the cell nucleus. However, in order to interpret their functions properly, it is vital to understand the way the protein are interacting together in these complexes. Mapping the *contact zones* consisting of surface amino acids interacting between the proteins is time consuming process that requires expensive laboratory experiments. Several computational tools there-

fore aim to reduce the amount of necessary experiments by predicting the possible docking conformations of given proteins. These tools can produce tens to hundreds of possible solutions and it is than up to biochemists to identify the plausible ones. To determine this, the researchers are trying to answer following questions:

- Which pairs of interacting amino acids are present in a given configuration?
- Which configurations contain a specific interacting pair of amino acids?
- How close are the amino acids in the contact zone and which are the closest ones?
- How similar and different are the contact zones in different configurations?
- What are the physico-chemical properties of the amino acids in the contact zone?

The identification of relevant docking conformations is currently not well supported and the domain experts performing this task by visually comparing the 3D representations of the docked proteins (see figure 1.5). This approach suffers from high visual complexity, occlusion and imprecise identification of contact pairs of amino acids. It is therefore very difficult to answer the posed questions.

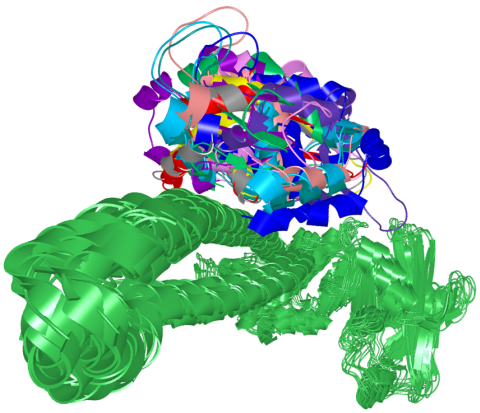

Figure 1.5: Superposition of several possible conformations between two proteins. The set of green protein instances corresponds to one of the proteins in the interaction, the colored components represent the second protein in different conformations.

1.2.3 Summary

In-silico simulations of chemical processes such as molecular docking reduce the time and costs necessary for in-vitro experiments. Yet, the complexity of the generated data almost always calls for further analysis. It is usually up to domain expert to judge the soundness of data and derive conclusions. Without popper tools, this can be difficult and tedious assignment. However, visualization metaphors supporting particular research tasks and their combinations in an interactive visual analytics's system can significantly speed up the analysis procedure and help with relieving interesting patterns and relationships in data.

In this work we will present the current state of the art techniques in the visualization and visual analysis of intermolecular interactions of proteins and analyse, how they address the questions posed in the previous sections. We will identify the unsolved problems occurring in the literature, then present the proposed solutions and results that have already been achieved. We will also outline the possibilities for further research.

2 State of the Art

In this chapter we will present the state of the art work present in bioinformatical literature with regards to the visualization and analysis of proteins and their interactions with other molecules. We will start with overview of existing molecular visualization techniques, then continue with the work related to protein-ligand interactions, where literature covers a substantial amount of diverse research. The end of this chapter will be dedicated to protein-protein interactions.

2.1 Molecular Visualization

Many different molecular representations have developed to cater for diverse needs of molecular biologists. Although some new representations are still emerging, the research in this area is currently more focused on development of fast visualization algorithms and GPU-based acceleration of traditional ones, in order to represent large and dynamic molecular data. Here we will provide the overview of typical molecular representations and they state of the art execution. There are, however, countless of other approaches that exceed the capacity of this work. The detailed study concerning molecular representation can be found in state of the art report by Kozlikova et al. [27]. Available is also the report by Patané and Spagnuolo focusing on modeling of molecular surfaces [53].

2.1.1 Atomistic and Bond-Centric Models

We can say that history of molecular visualization dates back to the 19th century. In 1808 John Dalton published his atomic theory [12], where he represented atoms and simple molecules with circular shapes. Couple of decades later, around 1860, August Wilhelm von Hoffmann started using first 3D models of molecules in his lectures at Royal Institution of Great Britain [54] – see Figure 2.1. This type of molecular representation is called *ball-and-stick* model, where balls represent atoms and sticks represent bonds between them. With couple of modifications this representation is commonly used also nowadays (Figure 2.2 b)).

MARSH-GAS.

Figure 2.1: Hoffman's methane (Marsh-gas) representation [54].

Over the years other derivations of ball-and-stick model emerged. In 1959 André Dreiding introduced molecular modelling kit using *stick-only* model [13]. Here the atoms were not represented by balls, but merely as connection points between sticks. Nowadays this model is called also *liquorice* or *Dreiding's model* (Figure 2.2 a)). The colouring of the sticks is often used to indicate atoms or their properties.

Although several researchers, including Dalton and Hoffman, claimed that different atoms have different radii, it wasn't until 1873 that the sizes of atoms were experimentally derived by Johannes Diderik van der Waals [68]. In later years this discovery led to so called *space-filling* molecular representations, also called *callote* or *CPK models* after chemists Robert Corey, Linus Pauling, and Walter Koltun [9]. In this representation, full "space-filling" sizes of atoms are used, which provides the overview of molecular surface (Figure 2.2 c)).

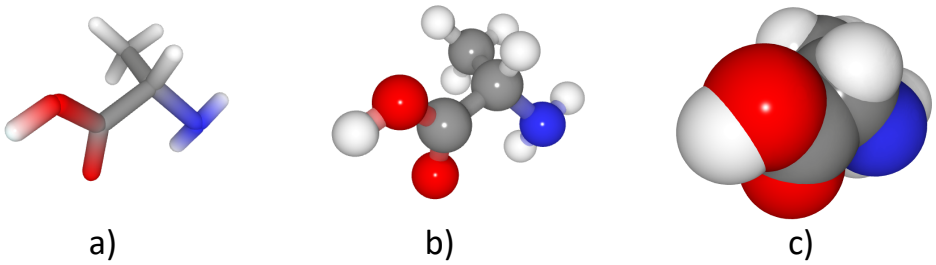

Figure 2.2: Types of molecular representations in modern visualization tools: a) liquorice model b) ball-and-stick model c) space-filling model

Atom and bond based representations of molecules can be decomposed into primitive shapes such as spheres and cylinders, which makes them suitable for GPU-based ray casting. Most of the state of the art rendering techniques stem from glyph ray casting introduce by Gumhold et al. [15], however many performance speedups focusing on rendering of large dynamic molecular structures exist. Since these techniques are focusing on large data samples, they often utilize level of detail (LOD) strategies. Example of this can be the two-level approach of Lampe et al. [33] where residues are each residue is represented by one vertex and the atoms in the residues are generated on-the-fly on the GPU. Another approach is used by Le Muzic et al. [37], where atom positions are stored in a texture and reconstructed using tessellation and geometry shaders.

2.1.2 Protein Architecture

The afore mentioned representations of molecules provide detail information about arrangement of atoms in a molecule. However, for proteins, which can consist of thousands of atoms, this representations can be too cluttered. Therefore, several schematic visualizations were developed.

One of the simplest representations of protein structure is called *alpha trace*. It depicts only the backbone of the protein, as it is derived from the positions of α-carbons (Figure 2.3 a)). This representation provides coarse overview of tertiary and quaternary structure of the protein – spatial arrangement of the polypeptide chains. However, it can be difficult to identify secondary structures from the alpha trace.

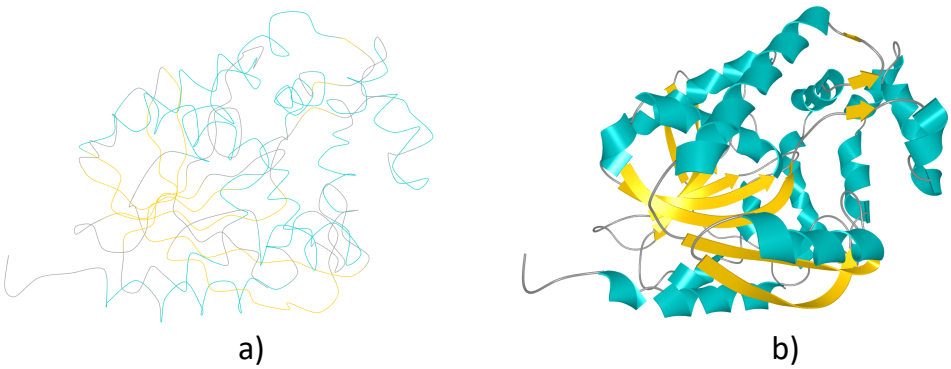

Figure 2.3: Types of molecular representations in modern visualization tools: a) alpha trace b) ribbon diagrams

In 1981 Jane S. Richardson published *cartoon* illustrations of all then known protein structures [60]. In these schematic representations, known today as *ribbon diagrams*, she used consistent and intuitive illustrations of secondary structures to demarcate their position along protein backbone. Although ribbon diagrams were originally hand drawn, they are nowadays part of every molecular visualization software (Figure 2.3 b)).

Currently the fastes approaches to visualization of ribbon diagrams include the two stage approach by Wahle and Birmanns [69] where first the backbone tube is generated on CPU and than the vertices are adjusted on GPU to form the final geometry. Another adaptive method by Hermosilla et al. [18] takes advantage of the tessellation shader and generates only the geometry needed for current viewpoint.

2.1.3 Surface Representations

These representations communicate the internal structure of the protein. However in many cases the focus of interest is on the surface of the protein, since the surface is the part of protein that is in contact with outer environment. It is therefore important for biochemists to identify the boundaries of the proteins that are accessible to ligands or interacting with other proteins.

We have already mentioned one type of surface defined by atom spheres of van der Walls radii and therefore called *van der Waals (vdW) surface* [59] – Figure 2.4 (blue). This surface indicates the precise molecular volume (Figure 2.6 a)).

Another type of surface – *solvent accesible surface (SAS)* was developed to show the regions of molecule accesible by a solvent molecules [38]. Here, the solvent molecule is approximated by a spherical probe, which rolls rolls over the vdW surface. The center of the probe than defines the SAS surface – Figure 2.4 (yellow). In other words, solvent accesible surface is equal to a vdW surface inflated by the radius of probe.

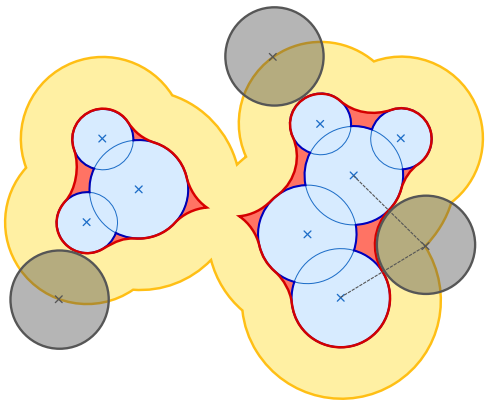

Figure 2.4: Schematic representation of molecular surfaces: vdW surface (blue), SES (red) and SAS (yellow). The SES and SAS are defined a probe (grey) rolling over vdW surface. Image taken from [27].

Solvent excluded surface (SES) [59] is defined in similar manner to SAS. However, instead of the center of the probe, its outer shell defines the surface – Figure 2.4 (red). It was the first smooth surface defined and thanks to the close approximation of molecular volume it is one of the most used surface representations (Figure 2.6 b)). Many algorithms for its computation and visualization have been developed over the years. Currently the fastest soulutions include paralelization of contour-buildap algorithm [64] – algorithm that computes track of the probe on atom surfaces – by Lindow et al. [46] and Krone et al. [29] as well as a grid-based approach by Hermosilla at al. [19] that utilizes progressive surface refinement for rendering of dynamic models on the fly.

Ligand exluded surface (LES) is a relatively new generalization of SES proposed by Lindow et al. [45]. Instead of using an approximate probe, it uses

full geometry of ligand to generate the surface. It thus illustrates the precise accessibility, however it is very computationally demanding (Figure 2.6 c)).

Yet another type of molecular surface – *molecular skin surface (MSS)* was proposed by Edelsbrunner [14] (Figure 2.6 c)). The shape of MSS depends on the single parameter s – shrink factor. The advantage of MSS over SES is full C^1 continuity. Among the fastest aproaches to generation of MSS belong the ones by Lindow et al. [46] and Yan et al. [71].

In 1982 Blinn [2] proposed use of a Gaussian convolution kernel to blend atom potentials to achieve an approximation of molecular surface. This technique callled *convolution surface model (CSM)* is more commonly known as Metaballs (Figure 2.6 e)). As with other techniques, improvements and new kernels have been proposed over the years (e.g. by Krone et al. [32]) and the resulting techniques belong to the fastest surface rendering approaches.

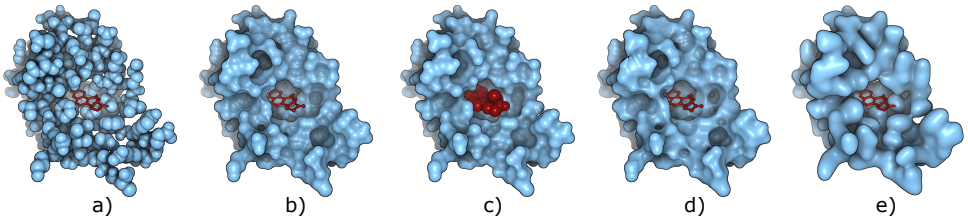

Figure 2.5: Comparison between different molecular surfaces of the protein isomerase: a) vdW surface b) SES with probe radius 1:4 Å c) LES for equilenine d) MSSwith shrink factor 0.35 e) Gaussian surface with standard deviation equal to the atom radius. Image taken from [27].

Surface Simplification

With larger kernels, the CMS can be used for simplification of molecular surface, showing just the general shape of the protein. This is sometimes required, as molecular surfaces are often used for mapping of other properties of proteins and atomistic models with many occlusions and high visual complexity are not suitable for this purpose.

Couple of other approaches for surface simplification have been proposed. *Coarse graining* [40] is method that groups several atoms (e.g. one amino acids) together. These groups are than represented by single sphere. Another approach is mapping the molecular surface to spherical coordinates [57].

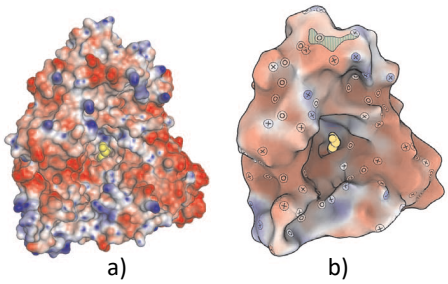

Figure 2.6: Simplified surface representation proposed by Cipriano and Gleicher. a) SES b) simplified surface. Image taken from [7].

Cipriano and Gleicher [7] proposed a method that uses a combination of filters and mesh restructuring to smooth the low frequency parts of the surface and generate a simplified representation of overall shape of the protein (see Figure 2.6). Markings are then placed on the surface to indicate the important removed details as well as other features of the surface, such as the physico-chemical properties of surface amino acids.

Visual Enhancements

Another way of dealing with high visual complexity of molecular structures is introduction various visual enhancements. One the most employed visual cues in molecular rendering is certainly *ambient occlusion (AO)* [49]. The goal of this technique is to calculate, how each object in the scene is exposed to the ambient lighting. It is computationaly very expensive, however the latest results from Hermosilla et al. [17] enable real time rendering of AO for MD simulations.

In their work, they also cover rendering of another common effect – *halos*. Haloes are highlights extending from selected object boundaries (used e.g. to indicate ligand in an MD simulation). Similar to them are also *depth-dependent silhouettes*, that contour the edges detected from scene depth map. Another effect for guidance of attention was adapted from photography – *Depth of Field* simulation blurs the objects that are out of focus. Other rendering techniques, such as *toon shading*, *line drawing* and *hatching* are commonly used for molecular rendering as well, since they can be adjusted to emphasize important features [27].

2.2 Analysis of Protein Voids

We have described many possible representations that support the analysis of proteins in general. In this section, we will focus on one of the most important features of proteins – their inner voids. As we mentioned before, the active site – the reactive area of the protein is often buried deeply inside of the protein structure and accessible only via those voids, namely tunnels.

Therefore, the extraction and analysis of these tunnels is vital for the study of protein-ligand binding. Here, we will mention only several most videly used principles of the approaches used in this area, as the extensiveness of this work exceeds the capacity of this thesis. Complete overview of the published tools for detection and analysis of biomolecular cavities can be found in state of the art reports by Krone et al. [30] and Simões et al. [61].

2.2.1 Detection of Protein Voids

There are several methods for extracting the shape of the protein voids in general, as well as numerous ones focusing on tunnels specifically. The algorithms can be classified into several categories, depending on the approach they use: grid-based, probe-based, Voronoi-based, surface-based, path analysis and ligand based. Most of the algorithms combine several approaches, in order to achieve better results. Moreover, we can differentiate between algorithms applicable only for static structures and algorithms taking into account molecular dynamics.

Many algorithms for void detection use a voxel grid to subdivide the 3D space containing the protein. The basic ide of purely *grid-based approaches* is to split voxels into two groups – those that lie inside a protein atoms and those that lie in a void space. The grid points from the second group are then assigned value based on different properties – distance to the protein atoms ([39, 56, 16]), interaction energies ([1, 35, 20]) or protein and solvent residence probability extracted from MD simulation ([58, 28, 25, 51]). The evaluated grid is further processed to extract the cavities using e.g. flood-fill segmentation or path analysis to find the cheapest path from active site to the surface of the protein.

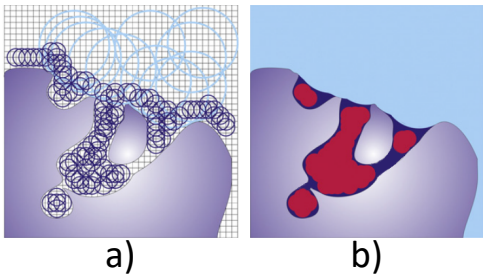

Figure 2.7: Rolling probe method. a) Two kinds of probes are placed in each point of the grid. b) The identified volumes are divided into the protein surrounding (light blue), and internal voids (red). Dark blue areas indicate undetected internal volume. Image adapted from [3].

Second group of algorithms utilizes *grid-based approaches* in combination with *probe-based approaches*. For example, HOLLOW [21] and 3V [67] algorithms use two probe spheres of different sizes that are placed in each node of the grid (see Figure 2.7). The probes that do not intersect with protein atoms define

the surface of the protein – large probe defines the outer surface, while the smaller one defines also the inner voids. This approach is also referred to as *rolling probe* principle. Different combinations probe-based and grid-based approaches can be found e.g. in [24, 11, 34, 72].

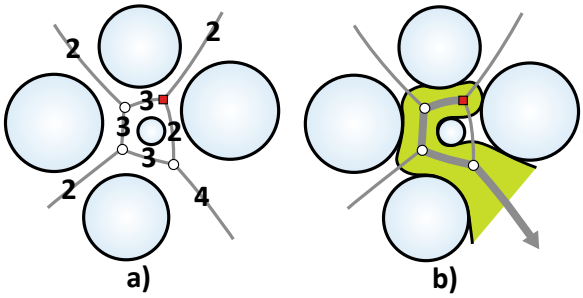

Figure 2.8: Example of Voronoi-based tunnel detection. a) Evaluated edges. b) Path with highest score found by Dijkstra's algorithm. Red square indicates active site. Image adapted from [48]

Accuracy of grid-based algorithms strongly depends on the resolution of the voxel grid. At the same time, high resolution of the grid leads to high memory demands of these algorithms. *Voronoi-based* algorithms in combination with *path analysis* address these drawbacks by utilizing Voronoi diagrams to subdivide the 3D space of protein structure (see Figure 2.8). Each atom of the protein forms a center of Voronoi cell. The edges are then evaluated by cost function, which assigns the value based on distance of the edge from the cell centres (i.e. atom centres). Then, Dijkstra's algorithm is used on the edge graph to find the best path from the active site towards protein surface. This principle is used and improved upon in [55, 48, 70]. Chovancova et al. [6] extend this approach for detection of tunnels taking into account the movement of the protein in CAVER 3.0 algorithm. It computes tunnel paths for each time frame of MD simulation. Than the corresponding paths are clustered. Thus it is possible to track the evolution of the tunnels in time.

Another set of void detection methods is based on theory of α-shapes. First, they compute Voronoi diagram on the atoms of the protein in a same manner as previous methods and transform it to Delaunay triangulation. Than, all the triangles that do not lie completely inside the protein atoms are deleted, resulting in an α-shape of the protein. The cavities can than be easily extracted from the α-shape. This method was first used in CAST [41] and was later extended and generalized by [62, 23, 47].

In 2011 two new approaches [50, 44] appeared using weighted *Voronoi diagrams* of atom spheres in combination with *probes*. For each Voronoi vertex of these diagrams an empty sphere tangent to four atom spheres exists. If the sphere is larger than a size of probe it is consider to be a cavity. Since

this approach considers the atom radii, it produces geometrically optimal results for probe spheres.

In Section 2.1.3 we have described several methods for extraction of molecular surfaces. Based on these methods, several *surface-based* approaches for detection of cavities have been proposed. For example, Jurčík et al. [22] extend the method proposed for computation of SES by Krone et al. [29] to detect also closed cavities. Coleman and Sharp [8] use triangulation of SES to detect channels in proteins. Krone et al. [31] proposed a method for real time GPU accelerated extraction of cavities based on Ambient Occlusion. The LES method proposed by Lindow et al. [45] enables extraction of cavities based on the actual geometry of the ligand.

The tunnel detection can be alternatively viewed as a path planning problem – the task is to find a collision free path for ligand starting at active site and leading to the surface. Path planning approaches to tunnel detection usually employ Rapidly Exploring Random Trees [36] (RRT), a technique that builds a tree of collision free configurations moving through the defined space. This approach can be applied directly on an MD simulation, thus it avoids the computationally expensive step of traditional, e.g. Voronoi-based approaches, where correspondence between geometry computed in different MD snapshots has to be found. Most notable work in this are includes [10, 65, 66].

2.2.2 Visual Analysis of Protein Tunnels

Spatial visual representations of molecular tunnels correspond to traditional molecular visualizations. The most simple tunnel representation consist of set of spheres placed along the tunnel centreline – the radius of each sphere corresponds to the maximal probe that fits into the tunnel at given place without intersecting with molecular atoms. Optionally tunnel centreline can be displayed as well.

We have already mentioned several surface-based methods for cavity detection that consequently use surface visualization to depict the detected cavities. Those methods also propose visual enhancements to highlight the cavities within the context of molecule – e.g. customized clipping planes, modulated coloring [52] or transparency of molecular srface [22] or selective lighting [44].

Moreover, several systems interactively combine 3D view and abstracted 2D plots to support the analysis of detected tunnels. Lindow et al. [42, 43] proposed an interactive tool for analysis of changes in the cavities in molecular dynamics. The tool includes interactive relational graph, that plots the

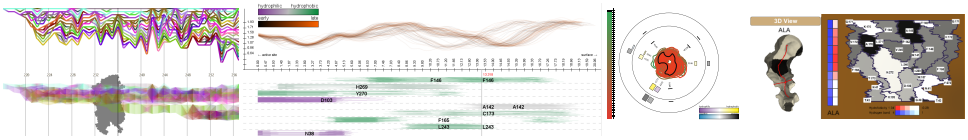

Figure 2.9: a) Relational graph by Lindow et al. [43] b) AnimoAminoMiner[5] c)MoleCollar [4], d) Tunnel unfolding by Koles8r et al. [26].

evolution of the position and size of the cavities, as well as their splits and merges. On demand it offers 3D overview of aggregated shaped of selected cavity traced through the MD simulation. Similar principles were also utilized by Krone et al. krone2014visual.

Byška et al. [5] presented a tool that combines line plot showing the tunnel width profile with a list of amino acids that influence the tunnel at different time steps of an MD simulation. By interaction it is possible to find the amino acids that might cause tunnel bottlenecks. Byška et al. [4] also presented another tool for exploration of evolution of molecular tunnel. In this work they employed filterable heat maps to provide width overview for multiple detected tunnels as well as more detailed view based on tunnel cross-cuts which again plots surrounding amino acids.

Yet another approach to tunnel analysis was presented by Kolesár er al. [26]. This work presents a method for comparison and analysis of tunnels based on unfolding of a 3D surface and flattening it into 2D map-like representation that denotes the amino acids surrounding the tunnel. These 2D representations can than be clustered and compared using image processing methods.

2.3 Protein-Protein Interactions

2.3.1 Docking

2.3.2 Visual Analysis

3 Aims of the Thesis

4 Achieved Results

5 Author's Publications

Bibliography

[1] Jianghong An, Maxim Totrov, and Ruben Abagyan. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & Cellular Proteomics*, 4(6):752–761, 2005.

[2] James F Blinn. A generalization of algebraic surface drawing. *ACM transactions on graphics (TOG)*, 1(3):235–256, 1982.

[3] Jan Brezovsky, Eva Chovancova, Artur Gora, Antonin Pavelka, Lada Biedermannova, and Jiri Damborsky. Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnology advances*, 31(1):38–49, 2013.

[4] J Byška, A Jurčík, M Eduard Gröller, Ivan Viola, and Barbora Kozlíková. Molecollar and tunnel heat map visualizations for conveying spatio-temporo-chemical properties across and along protein voids. In *Computer Graphics Forum*, volume 34, pages 1–10. Wiley Online Library, 2015.

[5] Jan Byška, Mathieu Le Muzic, M Eduard Gröller, Ivan Viola, and Barbora Kozlikova. Animoaminominer: Exploration of protein tunnels and their properties in molecular dynamics. *IEEE transactions on visualization and computer graphics*, 22(1):747–756, 2016.

[6] Eva Chovancova, Antonin Pavelka, Petr Benes, Ondrej Strnad, Jan Brezovsky, Barbora Kozlikova, Artur Gora, Vilem Sustr, Martin Klvana, Petr Medek, et al. Caver 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS computational biology*, 8(10):e1002708, 2012.

[7] Gregory Cipriano and Michael Gleicher. Molecular surface abstraction. *IEEE transactions on visualization and computer graphics*, 13(6):1608–1615, 2007.

[8] Ryan G. Coleman and Kim A. Sharp. Finding and characterizing tunnels in macromolecules with application to ion channels and pores. *Biophysical Journal*, 96(2):632–645, 2009.

[9] Robert B Corey and Linus Pauling. Molecular models of amino acids, peptides, and proteins. *Review of Scientific Instruments*, 24(8):621–627, 1953.

[10] Juan Cortés, Thierry Siméon, V Ruiz de Angulo, David Guieysse, Magali Remaud-Siméon, and Vinh Tran. A path planning approach for computing large-amplitude motions of flexible molecules. *Bioinformatics*, 21(suppl_1):i116–i125, 2005.

[11] Gábor Czirják. Princces: Continuity-based geometric decomposition and systematic visualization of the void repertoire of proteins. *Journal of Molecular Graphics and Modelling*, 62:118–127, 2015.

[12] John Dalton. *A new system of chemical philosophy*, volume 1. Cambridge University Press, 1808.

[13] Andre S Dreiding. Einfache molekularmodelle. *Helvetica Chimica Acta*, 42(4):1339–1344, 1959.

[14] Herbert Edelsbrunner. Deformable smooth surface design. *Discrete & Computational Geometry*, 21(1):87–115, 1999.

[15] Stefan Gumhold. Splatting illuminated ellipsoids with depth correction. In *VMV*, pages 245–252, 2003.

[16] Manfred Hendlich, Friedrich Rippmann, and Gerhard Barnickel. Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, 15(6):359–363, 1997.

[17] Pedro Hermosilla, Victor Guallar, Alvar Vinacua, and Pere-Pau Vázquez. High quality illustrative effects for molecular rendering. *Computers & Graphics*, 54:113–120, 2016.

[18] Pedro Hermosilla, Víctor Guallar, Álvaro Vinacua Pla, and Pere Pau Vázquez Alcocer. Instant visualization of secondary structures of molecular models. In *VCBM 15: Eurographics Workshop on Visual Computing for Biology and Medicine*, pages 51–60. European Association for Computer Graphics (Eurographics), 2015.

[19] Pedro Hermosilla, Michael Krone, Victor Guallar, Pere-Pau Vázquez, Àlvar Vinacua, and Timo Ropinski. Interactive gpu-based generation of solvent-excluded surfaces. *The Visual Computer*, pages 1–13, 2017.

[20] Marylens Hernandez, Dario Ghersi, and Roberto Sanchez. Sitehound-web: a server for ligand binding site identification in protein structures. *Nucleic acids research*, 37(suppl_2):W413–W416, 2009.

[21] Bosco K. Ho and Franz Gruswitz. HOLLOW: Generating accurate representations of channel and interior surfaces in molecular structures. *BMC Structural Biology*, 8(1):49, 2008.

[22] Adam Jurčík, Július Parulek, Jiří Sochor, and Barbora Kozlíkova. Accelerated visualization of transparent molecular surfaces in molecular dynamics. In *Pacific Visualization Symposium (PacificVis), 2016 IEEE*, pages 112–119. IEEE, 2016.

[23] Deok-Soo Kim, Youngsong Cho, Jae-Kwan Kim, and Kokichi Sugihara. Tunnels and voids in molecules via voronoi diagrams and beta-complexes. In *Transactions on Computational Science XX*, pages 92–111. Springer, 2013.

[24] Gerard J Kleywegt and T ALWYN Jones. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallographica Section D: Biological Crystallography*, 50(2):178–185, 1994.

[25] Daria B Kokh, Stefan Richter, Stefan Henrich, Paul Czodrowski, Friedrich Rippmann, and Rebecca C Wade. Trapp: a tool for analysis of transient binding pockets in proteins, 2013.

[26] Ivan Kolesár, Jan Byška, Julius Parulek, Helwig Hauser, and Barbora Kozlíková. Unfolding and interactive exploration of protein tunnels and their dynamics. In *Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine*, pages 1–10. Eurographics Association, 2016.

[27] Barbora Kozlikova, Michael Krone, Norbert Lindow, Martin Falk, Marc Baaden, Daniel Baum, Ivan Viola, Julius Parulek, Hans-Christian Hege, et al. Visualization of biomolecular structures: State of the art. In *Eurographics Conference on Visualization (EuroVis)-STARs*, pages 061–081. The Eurographics Association, 2015.

[28] Michael Krone, Martin Falk, Sascha Rehm, Jürgen Pleiss, and Thomas Ertl. Interactive exploration of protein cavities. *Computer Graphics Forum*, 30(3):673–682, 2011.

[29] Michael Krone, Sebastian Grottel, and Thomas Ertl. Parallel contour-buildup algorithm for the molecular surface. In *Biological Data Visualization (BioVis), 2011 IEEE Symposium on*, pages 17–22. IEEE, 2011.

[30] Michael Krone, Barbora Kozlíková, Norbert Lindow, Marc Baaden, Daniel Baum, Julius Parulek, H-C Hege, and Ivan Viola. Visual analysis of biomolecular cavities: State of the art. In *Computer Graphics Forum*, volume 35, pages 527–551. Wiley Online Library, 2016.

[31] Michael Krone, Guido Reina, Christoph Schulz, Tobias Kulschewski, Jürgen Pleiss, and Thomas Ertl. Interactive extraction and tracking of biomolecular surface features. In *Computer Graphics Forum*, volume 32, pages 331–340. Wiley Online Library, 2013.

[32] Michael Krone, John E Stone, Thomas Ertl, and Klaus Schulten. Fast visualization of gaussian density surfaces for molecular dynamics and particle system trajectories. *EuroVis-Short Papers*, 2012:67–71, 2012.

[33] Ove Daae Lampe, Ivan Viola, Nathalie Reuter, and Helwig Hauser. Two-level approach to efficient visualization of protein dynamics. *IEEE transactions on visualization and computer graphics*, 13(6):1616–1623, 2007.

[34] Roman A Laskowski. Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of molecular graphics*, 13(5):323–330, 1995.

[35] Alasdair TR Laurie and Richard M Jackson. Q-sitefinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics*, 21(9):1908–1916, 2005.

[36] Steven M LaValle. Rapidly-exploring random trees: A new tool for path planning. 1998.

[37] Mathieu Le Muzic, Julius Parulek, Anne-Kristin Stavrum, and Ivan Viola. Illustrative visualization of molecular reactions using omniscient intelligence and passive agents. In *Computer Graphics Forum*, volume 33, pages 141–150. Wiley Online Library, 2014.

[38] Byungkook Lee and Frederic M Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of molecular biology*, 55(3):379IN3–400IN4, 1971.

[39] David G Levitt and Leonard J Banaszak. Pocket: a computer graphies method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of molecular graphics*, 10(4):229–234, 1992.

[40] Michael Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of molecular biology*, 104(1):59–107, 1976.

[41] Jie Liang, Clare Woodward, and Herbert Edelsbrunner. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein science*, 7(9):1884–1897, 1998.

[42] Norbert Lindow, Daniel Baum, Ana-Nicoleta Bondar, and Hans-Christian Hege. Dynamic channels in biomolecular systems: Path analysis and visualization. In *IEEE Symposium on Biological Data Visualization*, pages 99–106. IEEE, 2012.

[43] Norbert Lindow, Daniel Baum, Ana-Nicoleta Bondar, and Hans-Christian Hege. Exploring cavity dynamics in biomolecular systems. *BMC Bioinformatics*, 14(Suppl 19):S5, 2013.

[44] Norbert Lindow, Daniel Baum, and Hans-Christian Hege. Voronoi–based extraction and visualization of molecular paths. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2025–2034, 2011.

[45] Norbert Lindow, Daniel Baum, and Hans-Christian Hege. Ligand excluded surface: A new type of molecular surface. *IEEE transactions on visualization and computer graphics*, 20(12):2486–2495, 2014.

[46] Norbert Lindow, Daniel Baum, Steffen Prohaska, and Hans-Christian Hege. Accelerated visualization of dynamic molecular surfaces. In *Computer Graphics Forum*, volume 29, pages 943–952. Wiley Online Library, 2010.

[47] Talha Bin Masood, Sankaran Sandhya, Nagasuma Chandra, and Vijay Natarajan. Chexvis: a tool for molecular channel extraction and visualization. *BMC bioinformatics*, 16(1):119, 2015.

[48] Petr Medek, Petr Beneš, Jiří Sochor, Vicent Vivanloc, Jean-Christophe Hoelt, Coong Binh Hong, Mathias Paulin, Jonas Spillmann, M Becker, M Teschner, et al. Computation of tunnels in protein molecules using delaunay triangulation, 2007.

[49] Gavin Miller. Efficient algorithms for local and global accessibility shading. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 319–326. ACM, 1994.

[50] Kliment Olechnovič, Mindaugas Margelevičius, and Česlovas Venclovas. Voroprot: an interactive tool for the analysis and visualization of complex geometric features of protein structure. *Bioinformatics*, 27(5):723–724, 2010.

[51] Teresa Paramo, Alexandra East, Diana Garzón, Martin B Ulmschneider, and Peter J Bond. Efficient characterization of protein cavities within molecular simulation trajectories: trj_cavity. *Journal of chemical theory and computation*, 10(5):2151–2164, 2014.

[52] Julius Parulek, Cagatay Turkay, Nathalie Reuter, and Ivan Viola. Implicit surfaces for interactive graph based cavity analysis of molecular simulations. In *Biological Data Visualization (BioVis), 2012 IEEE Symposium on*, pages 115–122. IEEE, 2012.

[53] Giuseppe Patané and Michela Spagnuolo. State-of-the-art and perspectives of geometric and implicit modeling for molecular surfaces. In *Computational Electrostatics for Biological Applications*, pages 157–176. Springer, 2015.

[54] James A Perkins. A history of molecular representation. part one: 1800 to the 1960s. *J Biocommun*, 31(1):1, 2005.

[55] Martin Petřek, Pavlína Košinová, Jaroslav Koča, and Michal Otyepka. MOLE: a voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure*, 15(11):1357–1363, 2007.

[56] Martin Petřek, Michal Otyepka, Pavel Banáš, Pavlína Košinová, Jaroslav Koča, and Jiří Damborskỳ. Caver: a new tool to explore routes from protein clefts, pockets and cavities. *BMC bioinformatics*, 7(1):316, 2006.

[57] Nicolay Postarnakevich and Rahul Singh. Global-to-local representation and visualization of molecular surfaces using deformable models. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 782–787. ACM, 2009.

[58] Martin Raunest and Christian Kandt. dxtuber: detecting protein cavities, tunnels and clefts based on protein and solvent dynamics. *Journal of Molecular Graphics and Modelling*, 29(7):895–905, 2011.

[59] Frederic M Richards. Areas, volumes, packing, and protein structure. *Annual review of biophysics and bioengineering*, 6(1):151–176, 1977.

[60] Jane S Richardson. The anatomy and taxonomy of protein structure. *Advances in protein chemistry*, 34:167–339, 1981.

[61] Tiago Simões, Daniel Lopes, Sérgio Dias, Francisco Fernandes, João Pereira, Joaquim Jorge, Chandrajit Bajaj, and Abel Gomes. Geometric detection algorithms for cavities on protein surfaces in molecular graphics: A survey. In *Computer Graphics Forum*. Wiley Online Library.

[62] Raghavendra Sridharamurthy, Talha Bin Masood, Harish Doraiswamy, Siddharth Patel, Raghavan Varadarajan, and Vijay Natarajan. Extraction of robust voids and pockets in proteins. In *Visualization in Medicine and Life Sciences III*, pages 329–349. Springer, 2016.

[63] Ondřej Strnad. *Algorithms for Detecting Pathways in Large Protein Structures and Their Ensembles*. PhD thesis, Masaryk University, Faculty of Informatics, 2014.

[64] Maxim Totrov and Ruben Abagyan. The contour-buildup algorithm to calculate the analytical molecular surface. *Journal of structural biology*, 116(1):138–143, 1996.

[65] Vojtěch Vonásek and Barbora Kozlíková. Application of sampling-based path planning for tunnel detection in dynamic protein structures. In *Methods and Models in Automation and Robotics (MMAR), 2016 21st International Conference on*, pages 1010–1015. IEEE, 2016.

[66] Vojtěch Vonásek and Barbora Kozlíková. Tunnel detection in protein structures using sampling-based motion planning. In *Robot Motion and Control (RoMoCo), 2017 11th International Workshop on*, pages 185–192. IEEE, 2017.

[67] Neil R Voss and Mark Gerstein. 3v: cavity, channel and cleft volume calculator and extractor. *Nucleic acids research*, 38(suppl_2):W555–W562, 2010.

[68] Johannes Diderik van der Waals. *Over de Continuiteit van den Gas–en Vloeistoftoestand*. PhD thesis, 1873.

[69] Manuel Wahle and Stefan Birmanns. Gpu-accelerated visualization of protein dynamics in ribbon mode. In *Visualization and Data Analysis*, page 786805, 2011.

[70] Eitan Yaffe, Dan Fishelovitch, Haim J Wolfson, Dan Halperin, and Ruth Nussinov. Molaxis: efficient and accurate identification of channels in macromolecules. *Proteins: Structure, Function, and Bioinformatics*, 73(1):72–86, 2008.

[71] Ke Yan, Ho-Lun Cheng, Zhiwei Ji, Xin Zhang, and Huijuan Lu. Accelerating smooth molecular surface calculation. *Journal of Mathematical Biology*, Jul 2017.

[72] Jian Yu, Yong Zhou, Isao Tanaka, and Min Yao. Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics*, 26(1):46–52, 2009.