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** 1
  1.1 *Biochemical Definitions* 1
    1.1.1 Protein Structures 2
    1.1.2 Properties of Proteins 3
  1.2 *Problem Formulation* 5
    1.2.1 Protein-Ligand Interactions 5
    1.2.2 Protein-Protein Interactions 7
    1.2.3 Summary 8
2 **State of the Art** 9
  2.1 *Molecular Visualization* 9
  2.2 *Detection of Protein Voids* 10
  2.3 *Protein-Protein Interactions* 13
3 **Aims of the Thesis** 15
4 **Achieved Results** 17
5 **Author's Publications** 19

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).

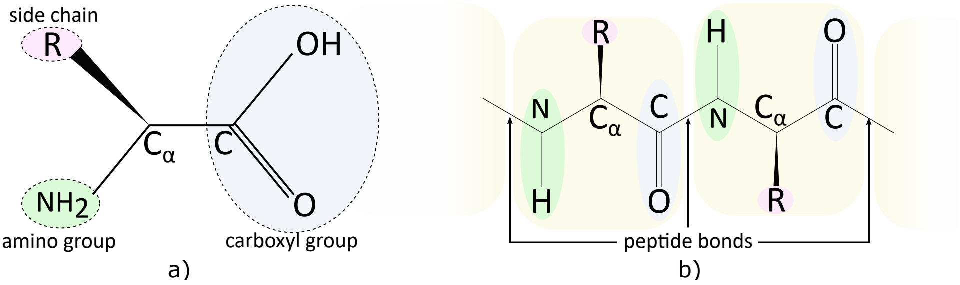

Figure 1.1: a) Illustration of a basic amino acid structure. b) Amino acid residues connected into polypeptide chain. It can be noted, that the amino and carboxyl groups are missing atoms, which were released during the formation of peptide bonds as H_2O molecules.

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.

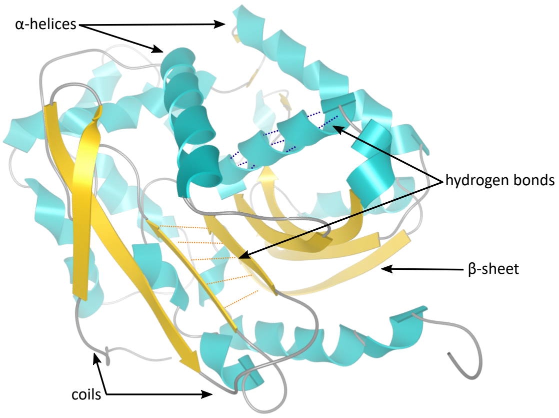

Figure 1.2: Typical secondary structures of protein: $\alpha - helices$ (blue), $\beta - strands$ (orange) forming $\beta - sheet$ and *coils*.

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.

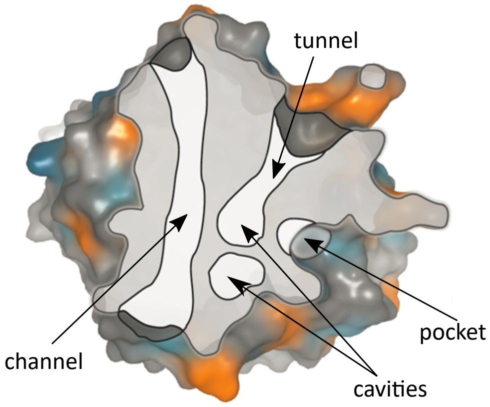

Figure 1.3: Types of inner voids of protein. Image adapted from [22]

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?

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.

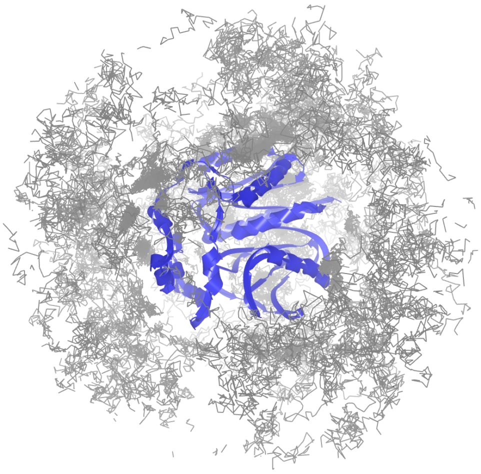

Figure 1.4: Ligand trajectory (gray) in a simulation containing 50 000 time steps. The protein chain is depicted in blue.

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 therefore 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.

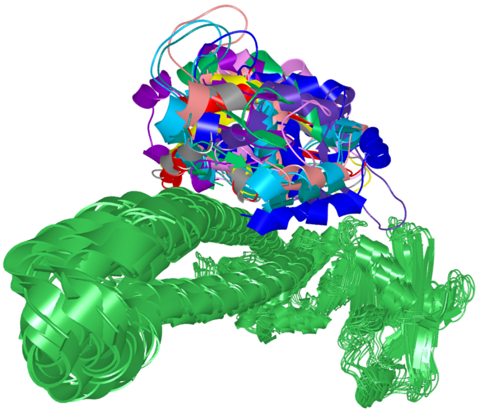

Figure 1.5: Superposition of several possible conformations between two proteins. The set of green protein instances corresponds to one of the proteins in the interaction, the colored components represent the second protein in different conformations.

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 intermolecular interactions of proteins. 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. Then we will continue with analysis of protein-protein interactions. This field is however only sparsely covered in literature.

2.1 Molecular Visualization

We can say that history of molecular visualization dates back to the 19th century. In 1808 John Dalton published his atomic theory [5], 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 [16] – 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)).

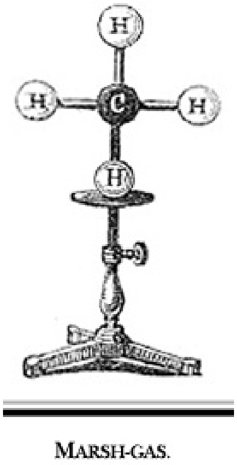

Figure 2.1: Hoffman's methane (Marsh-gas) representation.

Over the years other derivations of ball-and-stick model emerged. In 1959 André Dreiding introduced molecular modelling kit using *stick-only* model [6]. 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 [24]. In later years this discovery led to so called *space-filling* molecular representations, also called *CPK models* after chemists Robert Corey, Linus Pauling, and

Walter Koltun [4]. In this representation, full "space-filling" sizes of atoms are used, which provides the overview of molecular surface (Figure 2.2 c)).

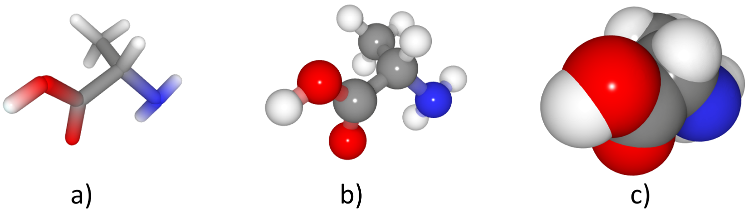

Figure 2.2: Types of molecular representations in modern visualization tools: a) liquorice model b) ball-and-stick model c) space-filling model

The afore mentioned representations of molecules provide detail information about structure 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.

In 1981 Jane S. Richardson published a consistent hand-drawn illustrations of all then known protein structures [20]. These schematic representations, known today as *ribbon diagrams*, provide intuitive overview of protein secondary structures (see section 1.1.1) as well as overall arrangement of the protein. Although they were originally hand drawn, they are nowadays part of every molecular visualization software.

Another, yet more simple representation of proteins is called *alpha trace*. It can be derived from the ribbon diagrams stripped from all the secondary structures – it depicted only the chain of α-carbons.

Surfaces, tools (py mol, analyst)

2.2 Detection of Protein 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 protein tunnels. Therefore extraction and analysis of these tunnels is vital for the study of protein-ligand binding. 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, surface-based, Voronoi-based, ligand based and path analysis. 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.

Due to the extensiveness of the work related to this topic, we will name only several representatives to demonstrate the basic principles of these approaches. Complete overview of the published tools for detection and analysis of biomolecular cavities can be found in state of the art report by Krone et al. [10].

Many algorithms for void detection use a voxel grid to subdivide the 3D space containing the protein. An exaple of a **grid-based approach** that also utilizes **path analysis** is the first version of CAVER algorithm [18]. Here, each node of the grid is assigned a cost, based on the maximal radius of a hypothetical ball that can be inserted into a node without intersecting voxels occupied by protein atoms – the larger the radius, the lower the cost. A graph searching algorithm than searches for cheapest path leading from user defined active site to the outer boundary of protein. This path is than taken as tunnel centreline and the detected tunnel is tan represented as a set of maximal spheres placed on centreline.

A slightly dofferent approach is utilized by HOLLOW [8] and 3V [23] algorithms. These algorithms use a combination of **grid-based** and **probe-based approach**. Two probe spheres of different sizes are placed in each node of the grid. 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 (see Figure 2.3). This approach is also referred to as *rolling probe* principle.

Other approaches that utilize voxelized grid and sphere probes include POCKET [12], VOIDOO [9], LIGSITE [7], SURFNET [11], Roll [26] or dxTuber [19].

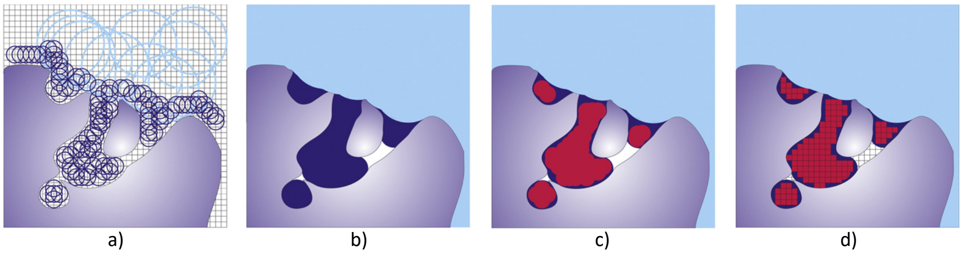

Figure 2.3: a) Rolling probe method. The surface and inner voids of the protein are defined by the placement of probes (large for surrounding and small for voids) in each point of the grid. b) The identified volumes are divided into the protein surrounding (light blue), and internal voids (dark blue), while undetected internal volumes are white. c) HOLLOW represents identified voids using the dummy atoms fitting these voids (red spheres). d) 3V represents detected internal void by voxels (red squares). Image adapted from [1].

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 2.4). 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.

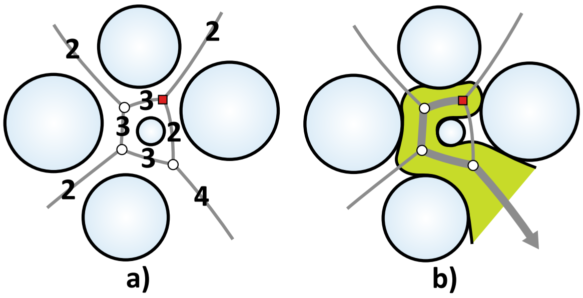

Figure 2.4: Example of Voronoi-based tunnel detection. a) Evaluated edges. b) Path with highest score found by Dijkstra's algorithm. Red square indicates active site. Image adapted from [14]

This principle is used in MOLE [17] and by Medek et al. [14]. MolAxis [25] increases the precision of this algorithm by approximating atom radii, which previous approaches omitted.

While all these approaches offer a solution for static molecules, CAVER 3.0 [2] extends them and allows for detection of tunnels taking into account the movement of the protein. 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.

Path analysis algorithms

HOLE [21] CHUNNEL [3] POREWALKER [15]

CAST [13]

ligand md simulations

2.3 Protein-Protein Interactions

3 Aims of the Thesis

4 Achieved Results

5 Author's Publications

Bibliography

[1] 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.

[2] 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.

[3] 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.

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

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] Marialuisa Pellegrini-Calace, Tim Maiwald, and Janet M. Thornton. PoreWalker: A novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PLOS Computational Biology*, 5(7):e1000440, 2009.

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

[17] 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.

[18] 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.

[19] 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.

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

[21] Oliver S Smart, Joseph G Neduvelil, Xiaonan Wang, BA Wallace, and Mark SP Sansom. Hole: a program for the analysis of the pore dimensions of ion channel structural models. *Journal of Molecular Graphics*, 14(6):354–360, 1996.

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

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

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

[25] E. Yaffe, D. Fishelovitch, H.J. Wolfson, D. Halperin, and R. Nussinov. MolAxis: Efficient and accurate identification of channels in macromolecules. *Proteins*, 73(1):72–86, 2008.

[26] 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.