

MoleCollar and Tunnel Heat Map Visualizations for Conveying Spatio-Temporo-Chemical Properties Across and Along Protein Voids

J. Byška¹ and A. Jurčík¹ and M. E. Gröller^{2,3} and I. Viola^{2,3} and B. Kozlíková^{1,2}

¹Masaryk University, Czech Republic, ²Vienna University of Technology, Austria, ³University of Bergen, Norway

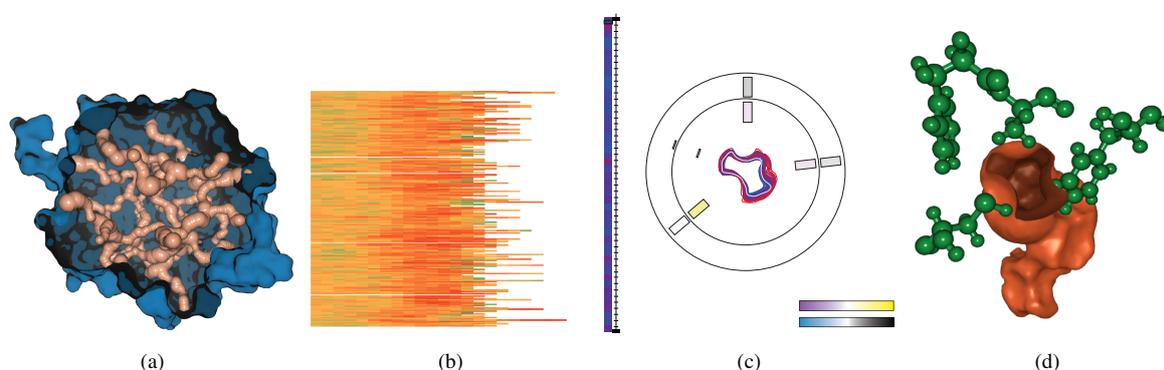


Figure 1: Proposed methods for a set of tunnels in protein structures computed within an ensemble of molecular dynamics. (a) One timestep of DhaA haloalkane dehalogenase with 39 tunnels leading to one active site; (b) heat map of one selected tunnel (STH) showing the evolution of the tunnel width over time (the left side represents the active site, the right side represents the tunnel gorge, and the vertical axis represents the time); (c) detailed exploration of the tunnel bottleneck contour (collar) over time; (d) 3D tunnel cut right at the bottleneck shown together with the surrounding amino acids.

Abstract

Studying the characteristics of proteins and their inner void space, including their geometry, physico-chemical properties and dynamics are instrumental for evaluating the reactivity of the protein with other small molecules. The analysis of long simulations of molecular dynamics produces a large number of voids which have to be further explored and evaluated. In this paper we propose three new methods: two of them convey important properties along the long axis of a selected void during molecular dynamics and one provides a comprehensive picture across the void. The first two proposed methods use a specific heat map to present two types of information: an overview of all detected tunnels in the dynamics and their bottleneck width and stability over time, and an overview of a specific tunnel in the dynamics showing the bottleneck position and changes of the tunnel length over time. These methods help to select a small subset of tunnels, which are explored individually and in detail. For this stage we propose the third method, which shows in one static image the temporal evolution of the shape of the most critical tunnel part, i.e., its bottleneck. This view is enriched with abstract depictions of different physico-chemical properties of the amino acids surrounding the bottleneck. The usefulness of our newly proposed methods is demonstrated on a case study and the feedback from the domain experts is included. The biochemists confirmed that our novel methods help to convey the information about the appearance and properties of tunnels in a very intuitive and comprehensible manner.

Categories and Subject Descriptors (according to ACM CCS): I.3.6 [Computer Graphics]: Picture/Image Generation—Line and curve generation

1. Introduction

Proteins are highly complex systems contributing to all functions in living organisms. The understanding of their structural and other properties is essential for the development of various chemical compounds, such as drugs. This is predominantly important for the health sector and pharmaceutical industry. In principle, the drug designers are searching for a proper combination of a small molecule (ligand) and a protein, which leads to the chemical reaction changing the ligand or the protein properties as desired. This search process consumes an enormous amount of time and resources because of the vast number of possible protein-ligand combinations to be tested. Thus, several in-silico predictive methods already appeared, which focus mostly on the analysis of the protein structure. It helps to evaluate the occurrence probability of a chemical reaction between the protein and given ligands. Ligands can bind to the protein at an active site which can be either located on the protein surface or is buried deep inside the protein. In the latter case, the in-silico analysis methods are often based on the detection of protein entrance paths which the ligand can follow from the outer solvent to the protein active site. The presence and properties of these paths, called tunnels, is fundamental for the computation of the occurrence probability of the chemical reaction between the scrutinized protein-ligand pair.

Due to the internal protein dynamics, tunnels and their characteristics may change significantly over time [KM02], so it is essential to study the tunnels within molecular dynamics simulations [KPK*09, LSL*11]. This increases the accuracy of the predictive methods substantially. The presence of a given tunnel in molecular dynamics signifies that the entrance path is opened within the simulation and the ligand can actually follow this tunnel to reach the protein active site. Generally, the longer the molecular dynamics is simulated and analyzed, the more relevant results the biochemists can obtain. The current computational power highly increased the capabilities of capturing simulations consisting of hundreds of thousands of timesteps. However, such lengthy molecular dynamics simulations introduced also an enormous complexity to the evaluation of analysis results performed on these simulations. In consequence, it is impossible for biochemists to observe all timesteps of such molecular dynamics in order to see the modification of tunnels over time. Such a situation requires the development of methods enabling a fast, accurate, and intuitive exploration of tunnels within the whole molecular dynamics, their properties and compatibility with the given ligand. This paper introduces novel methods for an interactive and visual tunnel exploration aiming to cover this gap. The proposed methods focus on two distinct stages of the workflow of biochemists in order to improve the effectiveness of the entire process of molecular dynamics data exploration.

The applicability of the proposed methods is demonstrated on a case study described in the biochemistry liter-

ature [KCB*13]. The authors aimed to increase the stability and resistance of DhaA haloalkane dehalogenase with respect to organic cosolvents. This is done by mutating the amino acids surrounding the main tunnel in order to narrow the tunnel bottleneck. We present the feedback from the domain experts who evaluated the benefits and drawbacks of our proposed methods.

2. Related Work

The proposed methods have been designed for the visual exploration of tunnels in proteins. Existing molecular tunnel analyses and visualizations are focusing on 3D representation that gives a realistic view on the overall structure as such. It fails however to provide essential information about the most critical part of its structure which is provided by our visualization design. These two methods, 3D visualization vs. 2D schematic visualization, are complementary to each other and can and should be used in combination.

In this section we first review the existing approaches concerning tunnel computation and geometric representation. The protein void space does not contain only tunnels. It can be classified into other categories as well, e.g., cavities, pockets, channels, pores. This is done according to their position with respect to the protein surface. Thus, we mention also techniques, which have been proposed for the analysis and exploration of these different types of voids. Finally, existing methods and techniques, which concentrate on 3D data projections and reformations, will be briefly described as well.

2.1. Computation and Representation of Tunnels in Protein Structures

The presence of tunnels in protein structures influences the protein's reactivity with other small chemical compounds. Several geometric solutions for proteins have been already proposed. The early solutions were based on a grid approach [POB*06]. The protein was covered by a regular grid and each cell was marked according to its distance to the nearest atom. Then, the Dijkstra algorithm was launched from the empty grid cell closest to the active site position. It computed the widest paths to the protein outer environment. The size and shape of the resulting tunnels were highly dependent on the grid resolution, which was the main limitation of this approach. This solution was then replaced by algorithms using the Voronoi diagram for representing the void space between protein atoms [MBS07, PKK007, SSVB*13]. Here each atom defines a Voronoi cell and the final path is reconstructed from Voronoi vertices and edges. The edges are evaluated and labeled with the distances to the closest atoms. Then, similarly to the grid approach, the Dijkstra algorithm is used to search the widest paths leading from the protein active site to the outer solvent. To rank the detected tunnels, the techniques propose different cost functions, which always take into account the width of the tunnel.

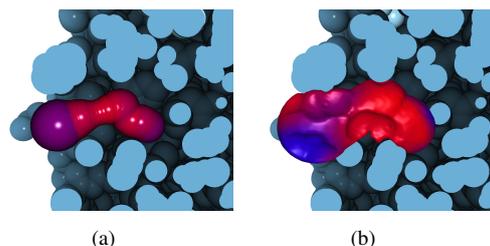


Figure 2: Different types of tunnel surface representations. (a) set of intersecting spheres, (b) asymmetric tunnel representation. The hydrophobicity is mapped onto the surfaces of both representations (hydrophobic = red, hydrophilic = blue, neutral = violet).

One of the most important descriptors of a tunnel is the size of its bottleneck, i.e., the narrowest point of the tunnel along its centerline. Lindow et al. [LBH11] introduce a method for the computation of tunnels utilizing Voronoi diagrams of spheres. Individual atoms are represented by spheres of varying radii. They are able to describe the protein inner void space more precisely than the previous methods. Kim et al. [KCKS13] build their approach to compute tunnels and voids in molecules on Voronoi diagrams, beta-complexes and quasi-triangulation.

These algorithms are able to compute tunnels only for static molecules. As the accuracy of the results can be highly increased by studying tunnels in molecular dynamics, several newer approaches concentrate mainly on this task. They are again based on the Voronoi diagram of spheres [LBBH12]. Alternatively they extend the standard Voronoi diagram on points with a spherical approximation of bigger atoms in order to take into account different atom sizes [CPB*12, YFW*08]. Using these tools, the domain experts are able to track individual tunnels within long molecular dynamics simulations.

A computed tunnel can be represented by its centerline, which is derived from a sequence of Voronoi faces. The tunnel width is limited by the surrounding atoms. A simpler tunnel representation samples the centerline regularly and each sample then defines the center of an inscribed sphere of maximal radius with respect to the surrounding atoms (see Figure 2a). The tunnel is then represented as a set of intersecting spheres and the sphere with the smallest radius defines the tunnel bottleneck. This representation only roughly approximates the tunnel shape and inner void space. Byška et al. [BJS13] introduce the definition and representation of so called asymmetric tunnels which describe the irregular tunnel shape more properly (Figure 2b).

Another possible representation comes from the analysis of tunnels in molecular dynamics [LBH11]. The ligand passage is traced through the tunnel over time and the ligand positions are accumulated. Then the authors overlap these

positions and create the final representation of the whole tunnel. The tunnel is again represented by a set of spheres.

2.2. Interactive Visual Analysis of Protein Inner Space

The analysis of protein structures can lead to the detection of void spaces of different types, such as tunnels, channels, pores, pockets, or cavities. They differ mainly in their location with respect to the protein surface. Pockets are defined as small protrusions on the protein surface which can contain the active site. Cavities represent the inside void space which is inaccessible from the surface. Deeply buried active sites are located in cavities. Channels and pores are defined as paths leading through the protein and having two gorges on the protein surface. Pores have a rather straight shape and play a role of transmembrane channels.

The algorithms dedicated to the detection of these void spaces can be based on geometric approaches. However, the geometric representation of a void space is often insufficient for the understanding of its complex properties and functions. Parulek et al. [PTRV13] use an implicit function sampling and graph algorithms for the extraction of cavities and combine it with their characterization by 3D graphs and by chemical properties. Lindow et al. [LBBH13] proposed a two-stage approach for the detection of cavities. In the pre-processing stage, the Voronoi diagram of spheres is used for computing cavities in the dynamics data. An interactive stage follows, where the user can compute, select and visualize the dynamic cavities. Krone et al. [KFR*11, KKRE14] present another approach for the detection of cavities and their visual analysis in dynamics. Cavities are tracked in molecular dynamics to show their stability over time. They combine several interactively linked views onto the cavities and their parameters.

2.3. Data Projections and Reformations

Large datasets, which have to be analyzed and understood, appear more and more in almost all research fields. As it becomes impossible for the users to explore all aspects of the data, it is necessary to provide them with more abstracted or simplified views onto the most interesting portions of the data.

In our first approach we extend the classic heat map visualization technique, which is very popular for displaying statistical information about the data. Its variations are widely used in bioinformatics [OCO107], but they were not yet applied to tunnel representations in proteins.

Our second proposed method enables the intuitive exploration of the tunnel bottleneck and the physico-chemical properties of the surrounding amino acids by exploring a 2D view extracted from the 3D tunnel representation. In principle, the 2D projection of the tunnel can be performed either along or across the tunnel centerline. Both types are used

in the method described by Mistelbauer et al. [MMV*13]. The authors introduce an aggregated view of cross-sectional planes of a vessel along its centerline. Such a representation shows the calcification of the arterial wall at different positions and can serve as a guidance for treatment planning. Our proposed method adapts the cross-section projection. It distantly resembles the bull's eye plot which was successfully adopted, e.g., by the medical visualization literature [MGP*93, TBB*07]. Slightly related to our proposed solution is also the prototype interface proposed by Lopez-Hernandez et al. [LHGMB10], which visualizes and interacts with univariate time-dependent digital signals. Wu et al. [WZQ*14] propose a tool for the visual analysis of 2D boundary changes where the boundary shape can resemble the shape of a tunnel cross-section. The flattening of a curved object along its centerline is commonly used in medical visualization [VBWKG01, HGQ*06] to study, e.g., the virtual unfolding of a colon.

3. Workflow Overview

In the following sections we describe the main principles as well as details about our newly proposed methods enhancing the process of filtering and exploration of tunnels in molecular dynamics. These methods are described with respect to their embedding into the workflow of biochemists. Firstly we discuss in detail the limitations of the current workflow and how our methods help to overcome them. Then, the description of the individual methods follows. Figure 3 describes the typical workflow of biochemists when searching for the compatibility between a given protein and ligand. This compatibility is defined by the probability of the mutual chemical reaction between the protein and ligand. The analysis phase of the workflow (Figure 3, 1) consists of the computation of tunnels in the input molecular dynamics. Several solutions to this analysis are available, mostly based on Voronoi diagrams [CPB*12, SSVB*13, YFW*08]. In our case we compute the tunnels using the CAVER 3.0 algorithm [CPB*12]. For each timestep, the algorithm computes a set of tunnels satisfying the input parameters. Then the correspondence between tunnels in individual timesteps is determined using clustering. As a result, the algorithm produces dozens of tunnels which change over time. The task is to study their properties and temporal behavior. The most suitable tunnel for ligand transportation must be determined or it is concluded that no tunnel can serve as the path for the given ligand. The second case implies that the protein and ligand are incompatible.

In the filtering step (Figure 3, 2), only a few tunnels should be selected and further scrutinized. This selection is based on the percentage of molecular dynamics timesteps when the tunnel is open (tunnel stability over time). Only the tunnels which are open for most of the simulation duration can serve as transport paths. We propose a novel interactive heat map visualization, described in detail in Section 4, which

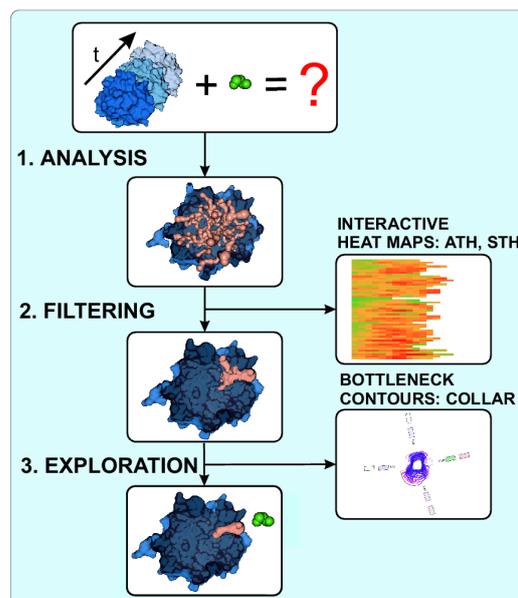


Figure 3: Illustration of the biochemist's workflow along with our proposed methods enhancing this process.

helps to reveal such tunnels without the necessity to observe all timesteps of the simulation. The entire tunnel system is shown in just one image. In the exploration step (Figure 3) the selected tunnels and their properties have to be explored in more detail. We propose another visual method, which is described in Section 5. The tunnel accessibility for ligands is largely governed by their size, shape, amino acid composition and physico-chemical properties. Therefore we offer the biochemists a novel method for the exploration of the most critical part of the tunnel, i.e., its bottleneck. The bottleneck is the narrowest cross-section of the tunnel and limits the size and shape of the admissible ligands. The method illustrates the shape of the bottleneck (contour) of the selected tunnel and its changes over time. This interactive representation includes the mapping of physico-chemical properties of the amino acids surrounding the contour and is tightly linked to the 3D representation of the tunnels.

4. ATH and STH Maps: Filtering of Tunnels

The tunnel analysis usually produces dozens of tunnels. When studying tunnels in one timestep, it is impossible to determine the stability of individual tunnels over time. A tunnel can be open in only one or a few timesteps, or it can be persistent during the entire simulation. In the first case, such a tunnel cannot serve as the ligand transport path and can be excluded from a further detailed exploration. In order to filter out such tunnels without the necessity to observe all timesteps of the molecular dynamics, we propose a fast and accurate representation of all dynamic tunnels in one static

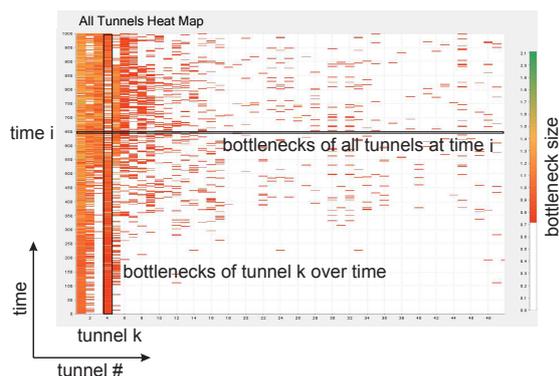


Figure 4: ATH (All Tunnels Heat Map) representation of all tunnels in a molecular dynamics. Each column corresponds to one tunnel, the vertical axis represents time. White spaces denote that a given tunnel is closed at the corresponding timestep.

image. For this we propose a specific heat map, called All Tunnels Heat Map (ATH) (see Figure 4).

The ATH shows all detected tunnels in the molecular dynamics. A rectangle in the heat map corresponds to the bottleneck of a tunnel at a specific point in time. The color encodes the bottleneck size. The color slider to the right defines the mapping of bottleneck size to color. A lower limit of the admissible bottleneck size can be interactively specified on the color slider. Bottlenecks below this limit are shown in white. A vertical line in the ATH corresponds to the temporal development of a specific tunnel over time. White rectangles on such a line indicate timesteps where the tunnel is either closed or its bottleneck is too small for the investigated ligand. A horizontal line corresponds to all the tunnels given at a specific point in time. Tunnels are sorted according to their priority from left to right, i.e., at the left of the heat map the most promising tunnels are shown. The priority of a given tunnel is calculated by averaging the sum of tunnel throughputs (bottleneck size) over all snapshots. If the tunnel is closed at a given snapshot, a zero value is used. So if a tunnel was open at least in one timestep of the molecular dynamics, it will appear in the ATH. Transient tunnels will be depicted in the right part of the heat map. Moreover, the ATH shows the temporal location of the open part of the tunnel. It makes a visual difference if the tunnel is open for a continuous portion of time or if the time steps when the tunnel is open are scattered discontinuously throughout the molecular dynamics. This is crucial to determine if the ligand will be able to follow the tunnel. From this representation, the user can quickly select only the tunnels which are most stable over time. These are candidates for further exploration.

After the filtering stage, the biochemists have to explore the selected tunnels in more detail. Here, one of the crucial parameters of the tunnel is its bottleneck. It represents the

minimal width of the tunnel, which determines the maximal size of admissible ligands. The biochemists usually study ligands of different sizes, which means that they have to compute the tunnels with different settings of the bottleneck. They have to launch the computations repeatedly using different bottleneck sizes. This is very time consuming as one such analysis takes from hours to days, depending on the molecular dynamics duration and the number of detected tunnels. We approach this problem by introducing an interactive Single Tunnel Heat Map (STH). In this case, it is necessary to perform the tunnel analysis only once, with a small bottleneck radius (e.g., 0.9 Ångströms). The STH representation now shows only one tunnel at a time (see Figure 5). The vertical axis again represents the time domain. The horizontal axis now represents the tunnel length and the color of individual rectangles encodes the width of the tunnel at corresponding positions along the tunnel centerline. The grey rectangles depict that the tunnel is closed, i.e., the bottleneck size is below the user defined threshold. We still can use the precomputed information about the tunnel width in the closed parts as well which is encoded as grey value. Darker grey values correspond to smaller bottleneck widths of the closed tunnel.

The greatest benefit of the STH lies in the possibility to interactively change the bottleneck size, which immediately shows the critical, i.e., narrow, parts of a tunnel and their evolution over time. The biochemists get an overview of the tunnel behavior when changing the minimum bottleneck size, which helps them to find an appropriate threshold value. This value then defines the maximal size of a ligand which could be transported via this tunnel.

An ATH is a much more aggregated representation as compared to the STH. In an ATH a tunnel is reduced to its bottleneck, whereas in an STH the entire tunnel at a specific point in time is given. By studying the throughput of tunnels using the ATH and STH, the biochemists can select candidate tunnels for the transportation of a ligand of a given size to the protein active site.

Working with an ATH and an STH, the user could be confused about the fact that a line in an ATH represents all tunnels at a given time while in an STH it represents the length of a single tunnel. Using different colors for an ATH and an STH, to avoid confusion, was not preferred option as the color in both cases represents width. Instead, both ATHs and STHs are distinctly labeled.

5. 2D Collar: Detailed Exploration of Tunnel Bottleneck

In the next phase, if the biochemists already selected a candidate tunnel for ligand penetration, the detailed exploration of a tunnel's shape and surrounding amino acids follows. The tunnel width alone is not sufficient for determining the compatibility between tunnel and ligand. Also the physicochemical properties of close-by amino acids and their atoms,

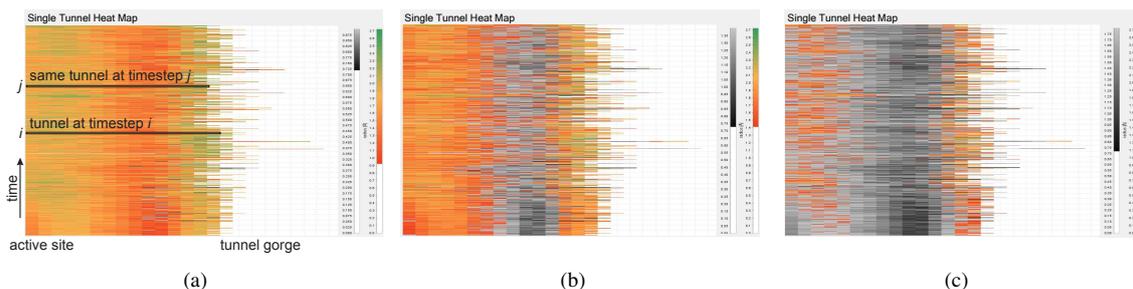


Figure 5: Three STH representing one tunnel over time, with different settings of the bottleneck size (0.9 Å in (a), 1.4 Å in (b), 1.8 Å in (c)). One row of the STH represents one tunnel at a specific timestep. The STH in (c) does not contain enough timesteps with open tunnels for a ligand of size 1.8 Å.

such as hydrophobicity or partial charges, have to be compatible with the corresponding properties of the ligand.

Performing this exploration in one timestep, the 3D representation of the tunnel surface colored according to the selected physico-chemical property may be sufficient (see Figure 2). In this case only one property of the surrounding amino acids can be mapped onto the tunnel surface at once. Moreover, the biochemists are now interested in the shape and area of the tunnel bottleneck. It can be depicted by cutting the tunnel with a clip plane, which creates a cross-section at the bottleneck position. Performing the same exploration tasks in molecular dynamics, the 3D representation becomes too complicated as it is changing over time.

The tunnel shape determines the admissible shape of the ligand, so the shape exploration is a necessary step of the whole process. In this respect, the most crucial part of the tunnel is its bottleneck. Thus, we introduce 2D collars for the reformation of the tunnel bottleneck to a 2D representation. This simplified aggregated view enables to explore a particular site of the tunnel, usually the bottleneck, over time. This is not a restriction to one place along the tunnel, the user can explore also an arbitrary site along the tunnel centerline. The 2D collar takes the shape of the tunnel cross-section, which is derived from the tunnel asymmetric shape, and traces its changes within the molecular dynamics. Individual cross-sections are represented by contours, which are plotted on top of each other within a single image. The bottlenecks are centered so as to provide an uncluttered view on their temporal evolution. The representation is further enriched with radial bars which represent several physico-chemical properties of the amino acids surrounding the tunnel in the corresponding cross-section position (see Figure 6). Each contour shows the bottleneck shape at a specific timestep. The physico-chemical properties of surrounding amino acids are plotted as bars around the contours, their 2D position is derived from their real position around the contour. They are colored with respect to the property they represent. The vertical slider represents the individual timesteps and the contours are colored according to this slider. Neighboring bars

with the same orientation represent amino acids, which influence the same portion of the contour within the molecular dynamics. Their length corresponds to the percentage of their occurrence at the contour over time.

The collar representation is based on the following principle. First, the bottleneck contour is extracted from the 3D tunnel representation. It is derived from the tunnel centerline, which is available from the tunnel computation. Moreover, from the computation phase, the centerline is already sampled in order to provide the domain experts with a visual representation based on a set of intersecting spheres. From this representation we also take the bottleneck position c which corresponds to the minimal tunnel sphere. The bottleneck contour is the intersection of a clip plane and the tunnel surface. The clip plane is positioned at center c and oriented to be orthogonal to the tunnel centerline. The normal vector of this clip plane is computed as an average vector defining the direction of the centerline at the given point c . To robustly compute the normal vector, we take into account up to ten neighboring points of c on the centerline. Then the intersection between this clip plane and the grid representation of the asymmetric tunnel is computed. The asymmetric tunnel is determined using a grid algorithm to determine individual tunnel voxels. Then, the tunnel surface mesh is derived from these voxels using Marching Cubes. Finally, the intersection between the clip plane and the tunnel surface mesh is computed which defines the resulting 2D contour.

The bottleneck contour detected in one timestep has to be subsequently traced through the molecular dynamics. As the tunnel shape changes over time, also the bottleneck position and its surrounding amino acids can be different at various timesteps. Thus, it is problematic just to take all bottlenecks detected in individual timesteps, compute their contours and visualize them. The reason is that they can be from completely different positions along the tunnel centerline. This issue is solved by clustering the examined tunnel in all timesteps according to the spatial position of their bottlenecks. The contours of all timesteps falling into the same

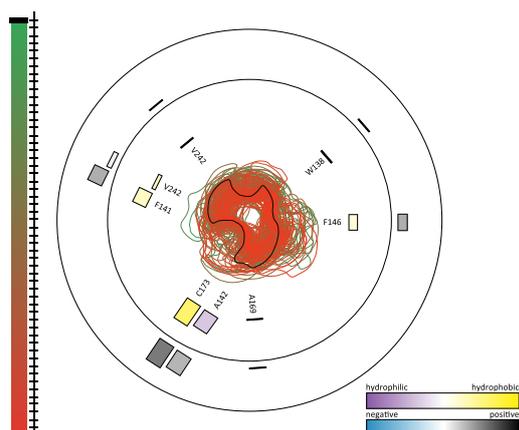


Figure 6: Collar of the bottleneck shape from molecular dynamics.

cluster are then visualized together because they lie in the same location or are very close to each other in the protein.

After the correct contours within the molecular dynamics are computed, they are aggregated into one static image. The center-point of these contours is determined by intersecting their clip planes with the centerline. Initially, the *up* vector of the contour corresponds to the *y* axis of the protein's coordinate system. This alignment can be incorrect as the whole protein can globally rotate over time. So we proposed two alternatives of contour alignment to the biochemists. The first alternative rotates the contours to maximize the empty space inside the bottleneck area. This type of alignment helps to reveal the admissible shape and maximal size of the ligand. However, it omits the shape fluctuation of the contour, which according to the biochemists is an important marker of the bottleneck throughput. The arrangement most favored by the biochemists aligns the contours to maximize the overlap of the surrounding amino acids. This alternative gives the biochemists the information how the amino acids influence the shape of the contour.

The collar representation further includes the schematic and quantitative visualization of the amino acids which surround the tunnel around the bottleneck contour. These amino acids intersect the clip plane which previously served for the definition of the contour. However, only the amino acids in a predefined distance from the contour are taken into account. The default value of the distance is 3 Å, but it can be changed by the user. Each amino acid is represented by a bar of predefined width. Its length corresponds to the number of occurrences of the amino acid around the given contour area within the whole molecular dynamics simulation. The maximal length of the bar corresponds to the situation where the amino acid influences the tunnel boundary around the bottleneck contour in all timesteps of the investigated molecular dynamics. The bar is interactive, i.e., by clicking

on it the user highlights the corresponding amino acid in the 3D view. Its minimal length is defined as well, in order to allow easy picking. The position of the bar corresponds to the position of the amino acid around the contour. The bars are positioned radially around the contours and one amino acid can be represented by several bars. In such a case each bar represents one physico-chemical property of the amino acid (e.g., hydrophobicity, partial charge of atoms). All bars are arranged in several circular levels. The number of levels corresponds to the number of displayed properties and each level corresponds to one property.

It is quite common that some amino acids defining the bottleneck are replaced by other amino acids over time. Then the collar representation has to react to this situation as well. We solve this by placing the bars corresponding to all surrounding amino acids next to each other (see Figure 6). The length of individual bars corresponds to the temporal occurrence of the amino acid at the bottleneck. The sum of their lengths corresponds to the maximal possible length determined for one bar. With this representation, the biochemists can immediately see how often an amino acid participates at the tunnel boundary.

Another feature of this novel collar representation is the possibility to interact with the contours in order to explore their shape and area more closely. By using a vertical time slider (Figure 6 on the left), the domain experts can pick one or several contours, which are then taken for highlighting or filtering out the rest of the contours. This can be achieved by using additional control elements. The color of a contour is determined by its temporal position within the molecular dynamics or by the contour area. Both criteria can be used for filtering as well. The biochemists can utilize several types of filters. The minimal or maximal contour filter highlights only the contour with the minimal or maximal area, respectively. Other filters enable to display only contours with a smaller or a greater area than a defined value. Our domain experts proposed also a filter selecting the representative contour, which is defined as the contour with the area most similar to the average area computed from all contours. Moreover, the time slider can be animated, which enables to track the changes of the contours over time.

The 2D contour representation is tightly linked to the 3D tunnel visualization. In the 3D view, the tunnel is intersected by the clip plane which defines the bottleneck contour (see Figure 7b). By selecting the amino acid bars in the 2D collar view, the corresponding amino acids are highlighted in the 3D representation as well.

Another view onto the bottleneck representation can be achieved by combining the 2D and 3D visualizations. This is done by mapping the 2D collar onto the clip plane visualized in 3D (Figure 7b). The advantage of this integrated visualization is the possibility to combine the 3D representation of surrounding amino acids with the information about their various physico-chemical properties.

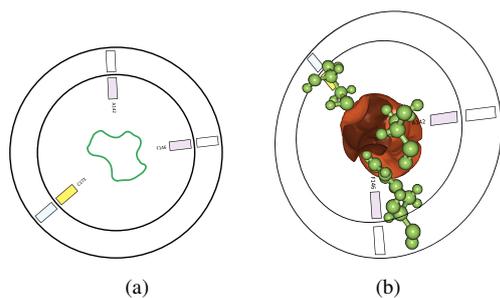


Figure 7: (a) 2D collar representation and (b) its mapping onto the 3D visualization of the clipped tunnel. The 3D representation of the amino acids is supported as well.

6. Demonstration and Results

Our proposed methods were tested and evaluated by domain experts, consisting of one professor, one post doc and two PhD students from a protein and metabolic engineering laboratory focusing mainly on computational biochemistry. They performed the testing on a case study of increasing protein thermostability, kinetic stability and resistance to organic cosolvents [KCB*13].

The proposed mutations around the tunnel focus on changing the bottleneck size, as it defines the size of ligands or other small molecules which can pass through such a tunnel. Thus, when designing the mutations, the biochemists have to study the bottleneck properties during the molecular dynamics in detail. It means that they have to perform the tunnel analysis with different bottleneck settings repeatedly, which is very time consuming. In the subsequent visual analysis, the existing solutions combine the 3D spherical representation of tunnels with a simple graph, showing the changes of the bottleneck width over time. In this case, the user has no information about the bottleneck's real shape and area. Moreover, the biochemists are not able to observe and take into account various properties of the surrounding amino acids. The biochemists confirmed that in these cases our proposed methods help to overcome the aforementioned problems.

In the first phase, the biochemists have to select the best tunnel candidates for further detail exploration. Here the ability to convey this information was tested for the ATH and the STH representations. The testing phase confirmed that the ATH is very helpful in cases where the users study large molecular dynamics and each timestep contains several tunnels. Here the ATH helps to immediately reveal the most stable tunnels over time by taking into account only columns (tunnels) with the most frequent occurrence of colored segments. In combination with the STH, the biochemist can further reduce the set of possibly relevant tunnels. In consequence, the combination of these two methods reduces the time for analysis considerably.

The input dataset for the case study consists of molecular dynamics simulations of the wild type (existing in nature) of DhaA haloalkane dehalogenase and its DhaA80 and DhaA106 mutants. The shape of the main tunnel in wild type does not change substantially over time, which enables the small molecules of the outer solvent to easily penetrate through this tunnel and to destroy the protein. The goal of the proposed mutations was to decrease the bottleneck of the main tunnel so the entrance of solvent molecules is less probable. The DhaA80 mutation narrowed the bottleneck too much so the access to the active site for ligands is more complicated. The finally proposed DhaA106 mutation slightly increased the void space of the bottleneck again so a ligand can penetrate but for the outer solvent it is more complicated to get into the tunnel.

These complex properties of DhaA mutants and bottlenecks of their main tunnel were explored by biochemists using our proposed 2D contour representation (see Figure 8). From this representation, it is clearly visible that the shape and area of the bottleneck of the main tunnel in the DhaA wild type is very stable (Figure 8a). The surrounding amino acids displayed around the contour can be immediately taken as candidates for mutations. The collar representation of the mutated bottleneck in the DhaA80 dehalogenase (Figure 8b) clearly shows that the shape and area of the bottleneck is changing dramatically. Finally, the DhaA106 mutation increased the bottleneck size (Figure 8c) in order to increase the tunnel inner void space and the probability of the ligand penetration. If contours of many time steps are displayed, our representation can suffer from visual clutter. So the implementation enables the user to interactively change the number of displayed contours by selecting a subset of timesteps. Moreover, we enable also the animation of the temporal changes of a contour. Using our representation, the biochemists were able to draw conclusions almost at first sight. Using the previously available visualization methods, this process lasted from hours to days. Communicating the information about the precise shape and area of the tunnel bottleneck and the different properties of the surrounding amino acids was highly rated by our domain experts.

The biochemists concluded that by using our 2D collar representation they can immediately get the information about the bottleneck shape and area and its evolution over time. Such information is not communicated by any of the existing solutions. This representation helps them to verify the first assumption, i.e., if the proposed mutation led to the desirable change of the bottleneck shape. Moreover, the biochemists gain also the information about the bottleneck surroundings and its selected properties. The surrounding amino acids are the direct candidates for further mutations. Most of the information provided by our novel collar representation is very complicated to obtain from or is completely missing in the existing solutions. The ATH and STH were tested by the biochemists on the molecular dynamics simulations, which they used for studying the penetration of

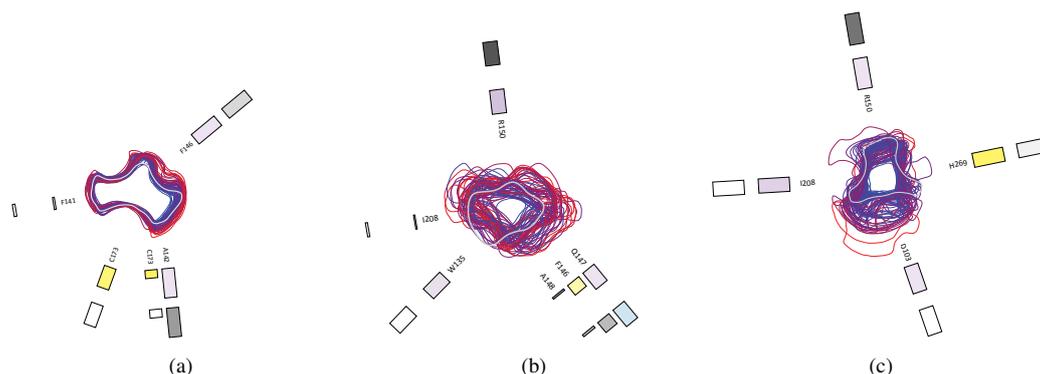


Figure 8: Difference in the bottleneck shape between the wild type of (a) DhaA haloalkane dehalogenase and its (b) DhaA80 and (c) DhaA106 mutations. The changes of the stability of the contour shape over time and the area of the void space correspond to the conclusions made by the biochemists.

ligands of different sizes. Working with interactive changes of the bottleneck size using ATHs and STHs, they were able to observe the probability of the ligand penetration without the necessity to recompute the tunnels with different bottleneck settings. In consequence, this reduces the time spent at this stage of their workflow dramatically.

The biochemists also pointed out few limitations of the 2D collar representation. If exploring long molecular dynamics simulations (thousands or tens of thousands of timesteps), the plotting of the large number of contours is not very comprehensible. In this case the biochemists proposed to draw only the contours from every n -th timestep, where n has to be very small, otherwise we might omit important timesteps. This would give them the overview of the whole dynamics and in the next step, they can select a desired subset which is then studied in full detail. Moreover, the biochemists would appreciate the ability to work with several 2D collar representations simultaneously. This feature would be very useful mainly in cases where a given tunnel has several narrow sites so the bottleneck position changes significantly within the molecular dynamics simulation. In this case, the biochemists have to select the most limiting site along the centerline, i.e., the most important bottleneck, and concentrate on further mutations around it.

All proposed methods were integrated into the CAVER Analyst [KSS*14] visualization tool which served also as a prototyping framework when designing the final appearance of our methods.

7. Conclusions and Future Work

In this paper we proposed three novel methods for the evaluation of the biochemical relevance of tunnels in proteins in large ensembles of molecular dynamics. The first two methods, i.e., ATH and STH, reduce the time and resources spent on filtering-out tunnels with low stability over time and too

small bottleneck sizes. The subset of selected candidates can be further explored by using our third proposed method, i.e., the 2D collar, which shows the bottleneck shape and area throughout the dynamics. It is further furnished with information about the surrounding amino acids and their selected properties. By adding interactivity and linkage to the 3D tunnel representation, the biochemists have various options and variants for exploration, which can finally lead to the correct in-silico evaluation of protein reactivity with respect to a given ligand.

In the future we will concentrate on increasing the usability of the proposed methods by implementing the features which came up from the discussions with the domain experts, such as working with additional windows simultaneously. Among other improvements proposed by the biochemists, there is also the need for computing and mapping the rigidity of the contour. This would differentiate the rather static from the flexible parts of the contour over time.

The applicability of the proposed methods is not strictly bounded to biochemistry alone. They can be successfully adopted to study all kinds of tubular structures where the surroundings plays an important role. An example would be biopsy trajectory planning. Studying the temporal evolution of a tunnel bottleneck is also quite similar to investigating a blood vessel stenosis. Our methods could be used for the exploration of real, e.g., railway, tunnels along a route. By aggregating the contour visualization of the narrowest parts of all tunnels on the route into a single image, we can immediately decide if a train of a given size will be able to use this route.

8. Acknowledgments

This work was supported through grants from the Vienna Science and Technology Fund (WWTF) through project VRG11-010 and the OeAD ICM and MSMT-1492/2015-1 through project CZ 17/2015.

References

- [BJS13] BYŠKA J., JURCIK A., SOCHOR J.: Geometry-based algorithm for detection of asymmetric tunnels in protein molecules. In *TPCG* (2013), pp. 17–24. 3
- [CPB*12] CHOVANCOVA E., PAVELKA A., BENES P., STRNAD O., BREZOVSKY J., KOZLIKOVA B., GORA A., SUSTR V., KLVANA M., MEDEK P., BIEDERMANNOVA L., SOCHOR J., DAMBORSKY J.: CAVER 3.0: A tool for the analysis of transport pathways in dynamic protein structures. *PLoS Computational Biology* 8, 10 (2012). 3, 4
- [HGQ*06] HONG W., GU X., QIU F., JIN M., KAUFMAN A.: Conformal virtual colon flattening. In *Proceedings of the 2006 ACM Symposium on Solid and Physical Modeling* (New York, NY, USA, 2006), SPM '06, ACM, pp. 85–93. 4
- [KCB*13] KOUDELAKOVA T., CHALOUPKOVA R., BREZOVSKY J., PROKOP Z., SEBESTOVA E., HESSELER M., KHABIRI M., PLEVAKA M., KULIK D., KUTA SMATANOVA I., REZACOVA P., ETTRICH R., BORNSCHEUER U. T., DAMBORSKY J.: Engineering enzyme stability and resistance to an organic cosolvent by modification of residues in the access tunnel. *Angewandte Chemie International Edition* 52, 7 (2013). 2, 8
- [KCKS13] KIM D.-S., CHO Y., KIM J.-K., SUGIHARA K.: Tunnels and voids in molecules via Voronoi diagrams and Beta-complexes. In *Transactions on Computational Science XX*, Gavrilova M., Tan C., Kalantari B., (Eds.), vol. 8110 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 92–111. 3
- [KFR*11] KRONE M., FALK M., REHM S., PLEISS J., ERTL T.: Interactive exploration of protein cavities. *Computer Graphics Forum* 30, 3 (2011), 673–682. 3
- [KKRE14] KRONE M., KAUKER D., REINA G., ERTL T.: Visual analysis of dynamic protein cavities and binding sites. In *Pacific Visualization Symposium (PacificVis), 2014 IEEE* (March 2014), pp. 301–305. 3
- [KM02] KARPLUS M., MCCAMMON J. A.: Molecular dynamics simulations of biomolecules. *Nature Structural Biology* 9, 9 (2002), 646–652. 2
- [KPK*09] KLVANA M., PAVLOVA M., KOUDELAKOVA T., CHALOUPKOVA R., DVORAK P., PROKOP Z., STSIAPANAVA A., KUTY M., KUTA-SMATANOVA I., DOHNALEK J., KULHANEK P., WADE R. C., DAMBORSKY J.: Pathways and mechanisms for product release in the engineered haloalkane dehalogenases explored using classical and random acceleration molecular dynamics simulations. *Journal of Molecular Biology* 392, 5 (2009), 1339–1356. 2
- [KSS*14] KOZLIKOVA B., SEBESTOVA E., SUSTR V., BREZOVSKY J., STRNAD O., DANIEL L., BEDNAR D., PAVELKA A., MANAK M., BEZDEKA M., BENES P., KOTRY M., GORA A. W., DAMBORSKY J., SOCHOR J.: CAVER Analyst 1.0: Graphic tool for interactive visualization and analysis of tunnels and channels in protein structures. *Bioinformatics* 30, 18 (2014), 2684–5. 9
- [LBBH12] LINDOW N., BAUM D., BONDAR A., HEGE H.: Dynamic channels in biomolecular systems: Path analysis and visualization. In *Biological Data Visualization (BioVis), 2012 IEEE Symposium on* (Oct 2012), pp. 99–106. 3
- [LBBH13] LINDOW N., BAUM D., BONDAR A.-N., HEGE H.-C.: Exploring cavity dynamics in biomolecular systems. *BMC Bioinformatics* 14, S-19 (2013), S5. 3
- [LBH11] LINDOW N., BAUM D., HEGE H.-C.: Voronoi-based extraction and visualization of molecular paths. *Visualization and Computer Graphics, IEEE Transactions on* 17, 12 (Dec 2011), 2025–2034. 3
- [LHGMB10] LOPEZ-HERNANDEZ R., GUILMAINE D., MCGUFFIN M., BARFORD L.: A layer-oriented interface for visualizing time-series data from oscilloscopes. In *Pacific Visualization Symposium (PacificVis), 2010 IEEE* (March 2010), pp. 41–48. 4
- [LSL*11] LI W., SHEN J., LIU G., TANG Y., HOSHINO T.: Exploring coumarin egress channels in human cytochrome p450 2a6 by random acceleration and steered molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics* 79, 1 (2011), 271–281. 2
- [MBS07] MEDEK P., BENES P., SOCHOR J.: Computation of tunnels in protein molecules using Delaunay triangulation. *Journal of WSCG* 15, 1-3 (2007), 107–114. 2
- [MGP*93] MINOVES M., GARCIA A., PAVIA J., HERRANZ R., SETOAIN J., MAGRINA J.: Evaluation of myocardial perfusion defects by means of "bull's eye" images. *Clinical Cardiology* 16, 1 (1993), 16–22. 4
- [MMV*13] MISTELBAUER G., MORAR A., VARCHOLA A., SCHERNTHANER R., BACLIIA I., KÖCHL A., KANITSAR A., BRUCKNER S., GRÖLLER M. E.: Vessel visualization using curvilinear feature aggregation. *Computer Graphics Forum* 32, 3 (June 2013), 231–240. 4
- [OCO107] OZEN M., CREIGHTON C. J., OZDEMIR M., ITTMANN M.: Widespread deregulation of microRNA expression in human prostate cancer. *Oncogene* 27 (2007), 1788–1793. 3
- [PKKO07] PETREK M., KOSINOVA P., KOCA J., OTYEPKA M.: MOLE: A Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure* 15, 11 (2007), 1357–1363. 2
- [POB*06] PETREK M., OTYEPKA M., BANAS P., KOSINOVA P., KOCA J., DAMBORSKY J.: CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics* 7 (2006). 2
- [PTRV13] PARULEK J., TURKAY C., REUTER N., VIOLA I.: Virtual cavity analysis in molecular simulations. *BMC Bioinformatics* 14, 19 (2013), 1–15. 3
- [SSVB*13] SEHNAL D., SVOBODOVA VAREKOVA R., BERKA K., PRAVDA L., NAVRATILOVA V., BANAS P., IONESCU C.-M., OTYEPKA M., KOCA J.: MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *Journal of Cheminformatics* 5, 1 (2013). 2, 4
- [TBB*07] TERMEER M., BESCOS J., BREEUWER M., VILANOVA A., GERRITSEN F., GROELLER M.: CoViCAD: Comprehensive visualization of coronary artery disease. *Visualization and Computer Graphics, IEEE Transactions on* 13, 6 (Nov 2007), 1632–1639. 4
- [YBWK01] VILANOVA BARTROLI A., WEGENKITTL R., KONIG A., GROELLER E.: Nonlinear virtual colon unfolding. In *Visualization, 2001. VIS '01. Proceedings* (Oct 2001), pp. 411–579. 4
- [WZQ*14] WU W., ZHENG Y., QU H., CHEN W., GROELLER E., NI L.: BoundarySeer: Visual analysis of 2D boundary changes. In *IEEE VAST* (2014). 4
- [YFW*08] YAFFE E., FISHELOVITCH D., WOLFSON H. J., HALPERIN D., NUSSINOV R.: MolAxis: Efficient and accurate identification of channels in macromolecules. *Proteins: Structure, Function, and Bioinformatics* 73, 1 (2008). 3, 4