Eurographics Workshop on Visual Computing for Biology and Medicine (2016)
S. Bruckner, B. Preim, and A. Vilanova (Editors)

Unfolding and Interactive Exploration of Protein Tunnels and their Dynamics

Ivan Kolesár[1], Jan Byška[1,2], Julius Parulek[1], Helwig Hauser[1], and Barbora Kozlíková[2]

[1]University of Bergen, Norway
[2]Masaryk University, Brno, Czech Republic

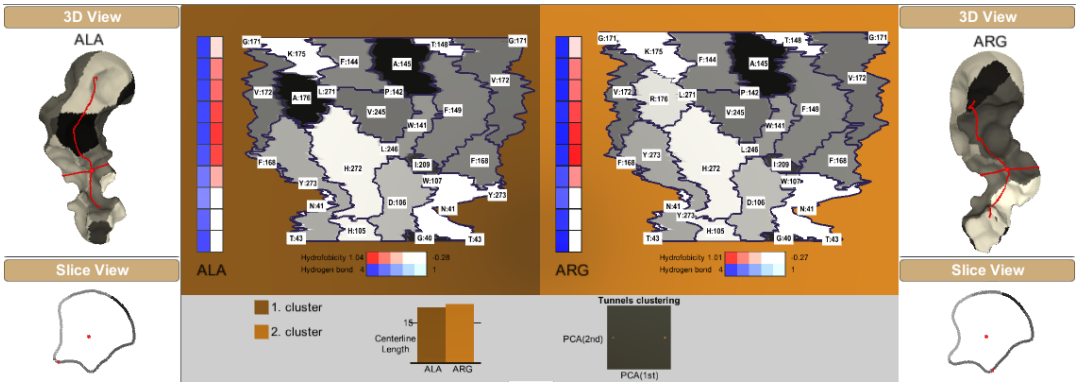

Figure 1: *Our proposed approach to exploring an ensemble of protein tunnels. It consists of three main parts: a scatterplot and a bar chart representation of tunnel properties on the bottom and the unfolded view of selected tunnels on the top. Additionally, there are also two linked views – a traditional 3D view and a slice view showing the contour of the horizontal cut through the tunnel.*

Abstract

The presence of tunnels in protein structures substantially influences their reactivity with other molecules. Therefore, studying their properties and changes over time has been in the scope of biochemists for decades. In this paper we introduce a novel approach for the comparative visualization and exploration of ensembles of tunnels. Our goal is to overcome occlusion problems with traditional tunnel representations while providing users a quick way to navigate through the input dataset and to identify potentially interesting tunnels. First, we unfold the input tunnels to a 2D representation enabling to observe the mutual position of amino acids forming the tunnel surface and the amount of surface they influence. These 2D images are subsequently described by image moments commonly used in image processing. This way we are able to detect similarities and outliers in the dataset, which are visualized as clusters in a scatterplot graph. The same coloring scheme is used in the linked bar chart enabling to detect the position of the cluster members over time. These views provide a way to select a subset of potentially interesting tunnels that can be further explored in detail using the 2D unfolded view and also traditional 3D representation. The usability of our approach is demonstrated by case studies conducted by domain experts.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation

1. Introduction

Studying molecular structures and their reactivity plays a crucial role in many research disciplines, including drug design, agriculture, cosmetics, industry, and others. The product of such a reaction can form a basis of new chemical substances, including medication. The reactivity of a molecule is highly influenced by the presence of a void space inside its structure. More densely packed molecules do not contain too much empty space between their atoms. The reaction site, which is often located deeply inside the molecule, is therefore less accessible, even for small ligands. Therefore, by exploring the amount and properties of void space inside molecules we

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

DOI: 10.2312/vcbm.20161265

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

are able to assess the possibility of a ligand with specific properties to enter these molecules. Such void space can be categorized according to different criteria. The most common approach takes into account the accessibility of the void space from the outer molecular surface. Then we can distinguish between inner closed cavities and paths. Closed cavities are inaccessible from the surface and are buried in the molecular structure. Accessible paths can be further divided into those which connect a specific site (here the so called active site) inside the molecule with its surface and those that pass through the molecule and connect two points on the molecular surface. The first type of paths is denoted as tunnels and the second as channels or pores.

In our research we are mainly interested in tunnels in protein molecules because they enable the transport of small ligands to the active site where a chemical reaction can occur. This process is crucial namely in drug design and protein engineering. From the latter field also come our user studies which we present at the end of the paper. In our research we focus on protein structures, but the proposed method is applicable to other kinds of molecules containing tunnels, as well.

Molecules are not static structures; their atoms are in permanent movement. This also impacts the void space, when detected tunnels become wider or narrower or even disappear. Recently it was revealed that the function of proteins is determined not only by their constitution but also their molecular dynamics [HMH*12]. Therefore, from the biochemical point of view, studying protein tunnels in one static time step is not too relevant because a given tunnel can be opened only for a fraction of time, for example. Biochemists are rather interested in the stability and changes of the tunnel shape over time. This exploration process has to be supported by special visualization techniques.

The majority of existing tunnel representations aim to visualize the 3D surface depicting the tunnel shape. However, such representations fail in cases when it is required to compare multiple tunnels at once. This can happen in situations when we want to explore changes of a tunnel in molecular dynamics, or several mutations of amino acids around a given tunnel.

When performing these tasks using traditional 3D visualization, the first case typically relies on an animation of the dynamic tunnel and superposition of the tunnels under different mutations has been used in the second case. Both cases suffer from several drawbacks. When dealing with large molecular dynamics simulations, consisting of hundreds of thousands of time steps, the animation of such dynamics is very hard to explore. The user is overwhelmed by movements and the differences in the tunnel shape can be easily overlooked. Moreover, using the 3D tunnel representation only a portion of the tunnel surface is visible from the viewpoint. When aiming to explore a tunnel influenced by several mutations of its surrounding amino acids, the user wants to observe and evaluate the impact of these mutations on the tunnel surface. The superposition of multiple mutated tunnels suffers from visual clutter caused by surface overlaps (see Figure 2 left). These overlaps can be removed by using juxtaposition of the tunnel representations (see Figure 2 right). But such a case is applicable only to a small set of objects because it can be hard to perceive the differences between them, in particular for those being further away from each other. This problem is even more exacerbated when using a 3D tunnel representation.

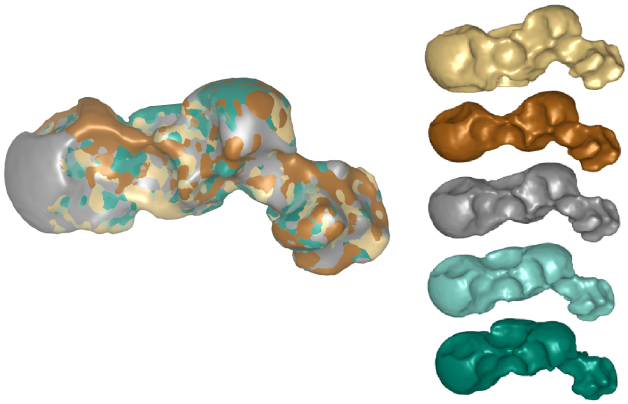

Figure 2: *Left: superposed representations of five conformations of a tunnel, differing by one mutated amino acid each. Right: a juxtaposed view of these conformations.*

The proper 3D visualization of the surface itself is not sufficient for exploration and the complete understanding of differences between protein tunnels, because essential information also relates to the amino acids forming the tunnel boundary, namely their physicochemical properties and the extent of their influence on the tunnel. A possible solution is to map this information onto the tunnel surface. Nevertheless, due to the aforementioned problems present in 3D tunnel visualization, it is hard to compare and explore these properties as well.

In this paper we aim at solving these problems by introducing a novel technique for the intuitive comparison and exploration of entire ensembles of protein tunnels. It consists of linked views, enabling to:

- get an overview of all tunnels in the input ensemble and spot possibly interesting tunnels
- localize interesting tunnel configurations in time
- visualize selected tunnels using a 2D unfolded view and compare their amino acid constitution and influence
- combine the unfolded view with the traditional 3D view to see the original shape

The benefit of our approach is demonstrated by two case studies, conducted by biochemists.

2. Related Work

The problem of visual exploration of spatial structures of tubular shape has been already addressed in other domains as well, e.g., in medical visualization [KFW*02, MVB*12]. Here the methods based on 3D data projections and reformations are used in order to reveal specific anatomical features. Nevertheless, due to the nature of the data, these methods mostly focus on the exploration rather than on the comparison. On the other hand, the comparative visualization in connection with surface reformation was applied many times when studying flow dynamics.

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

In the following sections we touch upon related existing approaches in more detail. First we provide the reader with an overview of the related work regarding the protein void detection and their visualization. Then we mention techniques which have been proposed for 3D data projections and reformation in medical visualization, flow dynamics, and other fields.

2.1. Protein Void Detection and Analysis

The earliest approaches to protein tunnel detection were based on grids [POB*06, VG10]. Even though these techniques allow to describe the void space of nearly any shape, the precision and performance of grid-based methods highly depend on the grid resolution. These drawbacks were later overcome by Voronoi diagram-based methods [MBS07, YFW*08, LBH11, CPB*12]. These methods first compute a Voronoi diagram of the protein atoms and use the Dijkstra algorithm on top of it. Such an approach enables to save a considerable amount of resources. The resulting tunnel is then represented by a set of spheres positioned onto the Voronoi edges forming the tunnel centerline. A more precise tunnel boundary is however determined by the surrounding atoms, so it is far from being spherical. Therefore, this traditional spherical representation may lead to a misinterpretation of the analyzed tunnel. As a consequence, some alternative methods combining the grid and Voronoi-based techniques appeared [JBSK15]. These approaches detect the tunnel centerline using the Voronoi diagram and then rasterize the space around the centerline to obtain more precise representation of the tunnel surface.

Nowadays, biochemists focus more on the analysis of simulations of molecular dynamics rather than on static molecules, as these techniques yield biochemically more relevant results. The analysis and subsequent exploration of tunnel behavior in long molecular dynamics simulations is, however, still one of the biggest challenges. Nevertheless, there are already several tools which are able to process such simulations [CPB*12, BCG*13]. In consequence, a few approaches to their visual exploration appeared, as well [LBBH13, KKL*16]. These methods focus on the interactive 3D exploration of inner protein voids in real time. However, they also suffer from problems related to the 3D representation itself. Moreover, by using these techniques it is hard to capture the significant changes of tunnel shape or constitution over time.

For the thorough analysis of protein behavior, the physico-chemical properties of amino acids in tunnel vicinity are at least equally important as the geometrical aspects. Therefore, according techniques were proposed by Parulek et al. [PTRV13], by Byska et al. [BJG*15, BLMG*15], and by Masood et al. [MSCN15]. These techniques are using a high level of abstraction in order to depict the evolution of protein voids including the nearby amino acids in a single image. However, when using their proposed 2D views, the information about the spatial influence of individual tunnel-lining amino acids and their mutual positions is hard to understand. This problem is addressed by our proposed approach.

2.2. Projections, Reformations and Comparative Visualization

The core idea of our technique is based on unfolding the 3D tunnel surface to a 2D image in order to prevent occlusion problems. Additionally, such a reformation allows us to dedicate one of the dimensions to the requirements of comparative visualization [GAW*11]. A similar approach has been used in other domains as well. For example, Kretschmer et al. [KST*14] adapted a variant of the as-rigid-as-possible reformation, proposed by Liu et al. [LZX*08], for medical applications. The resulting method defines a reformation of volumetric data driven by a particular geometry – in this case the anatomy of a specific patient. Another extension of the work of Liu et al. was presented recently by Brambilla et al. [BAAH16] who focus on so called time surfaces which are frequently used for the investigation of fluid flows. The proposed method is using a reformed 2D space in order to observe an evolution of complex surfaces over time. Unfortunately, since all these methods are relatively general, they produce an inevitable distortion in the distance. This is unsatisfactory when comparing protein tunnels since here the main focus is placed on the tunnel length, bottleneck, the overall surface shape and the surrounding amino acids.

In order to minimize the distortion as much as possible we can exploit the unique tubular shape of protein tunnels. A similar approach was used, for instance, by Gurijala et al. [GSZ*13]. They presented a method that produces a rectangular flattening of the colon. This method handles topological noise very well and preserves the colon wall shape, which allows to easily observe polyps (abnormal growth of tissue) inside the colon. The proposed normalized rectangle view, however, does not depict the radius of the colon. This can be omitted in this particular use case of colon polyp detection, but it is essential when designing a comparative visualization of protein tunnels.

Another example of surface reformation of tubular structures comes from the field of vessel visualization. The Curved Planar Reformation (CPR) [KFW*02] is a widely known technique used for the diagnosis of vascular diseases of peripheral arteries with a single dominant direction. In order to investigate the vessels of an arbitrary orientation, Mistelbauer et al. [MVB*12] proposed so called Centerline Reformations (CR). The CR is an extension of the CPR method but it uses a wavefront propagation of the vessel centerline for the automatic setting of the vector of interest. Due to the nature of this particular reformation there is a trade-off between the visibility of the vessels and errors when depicting the surrounding tissue as context. In order to overcome these problems, another variant of the CPR method was proposed by Auzinger et al. [AMB*13]. This method avoids the visibility issues of CPR and CR caused by the projection by computing the vessel lumen (empty space inside of a vessel) fully in 3D. This is done by finding a cut through the vessel and surrounding tissue orthogonal to the vessel centerline and view direction. The view-dependency would, however, cause some issues in our case, since we are aiming for comparative visualization and hence we need to find the same cut for each tunnel in order to obtain a common space. Moreover, the proposed cutaway discards a part of the vessel surface while we want to communicate the shape and physico-chemical properties of the whole protein tunnel.

There are several other examples in the literature regarding the comparative visualization of unfolded, or abstracted, tubular structures. For instance, Angelelli et al. [AH11] presented a method for the investigation of blood flow through tubular structures (the

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

aorta). Lampe at al. [LCMH09] on the other hand provided a solution for a curve driven analysis of general flow. Similarly to CPR, both methods are based on straightening of the volumetric data in the direction of the flow. Such an alignment allows to employ comparative visualization (e.g., juxtaposition view), because one axis becomes shared between multiple instances (in this case between various snapshots of the flow simulation over time). The method proposed by Angelelli et al. computes the reformation directly in the 3D space and hence it suffers from occlusion. On the other hand, the method designed by Lampe at al. provides a great overall view of individual flow features, but it does not preserve the shape nor angles of external objects. In both cases it means that at least some essential information about the protein tunnel surface and its surrounding amino acids would be lost. We are looking for a similar solution, but none of the mentioned methods can be used directly.

3. Unfolding-Based Exploration

As already stated, the main goal of our work is to allow an interactive exploration of molecular tunnels ensembles coming from molecular dynamics simulations or protein mutations. In both cases the tunnel shape and the physico-chemical properties can develop and change dramatically.

We derive tunnels from the raw data as a set of 3D objects representing the tunnel shape. In general, there is no trivial or direct way to orient the obtained set of tunnels such that the user could intuitively explore similarities and dissimilarities of individual tunnels, i.e., to perform co-registration. There are several issues preventing the direct comparison of such structures in 3D, e.g., the self-occlusion of tunnel features, occlusion between multiple tunnels, etc. Moreover, there is always a limit of how many 3D objects a user can explore at the same time in one window. To overcome the self-occlusion issues we unfold a 3D tunnel representation to 2D as suggested by various related work.

As already mentioned, molecular dynamics simulations can consist of many thousands of time steps. Therefore, the direct comparison of the shape of all tunnels, using either 3D or 2D representation, is a challenging task. To tackle this problem we utilize a semi-automated method for the evaluation of tunnel similarities. To do this, it is necessary to define a set of descriptors that would properly characterize the important properties of each tunnel. Several shape descriptors for whole proteins or binding sites were already suggested [WPS07], but these take into account only the shape of protein surfaces while in our case we emphasize also the physico-chemical properties of tunnel-lining amino acids.

Since we deal with 2D representation of tunnels as a solution for occlusion issues, we decided to employ a set of descriptors based on image moments [FSZ09]. The advantage of this approach is that the image moment is a well known descriptor for image analysis and classification. We can easily utilize the color channel for information about amino acids and hence directly add it to the formed image. Additionally, it gives us an explicit description of the representation that we will ultimately provide to the user.

In order to communicate the obtained results we need to employ a dimension reduction technique due to the large amount of both tunnels and possible descriptors. In our technique we exploit principal component analysis (PCA) due its simplicity, stability, and applicability to dimension reduction problems in data analysis; nevertheless, any other dimension reduction method could be applied as well. The results of PCA are then presented in a scatterplot view where the similarities/dissimilarities between tunnels are communicated by the distance between points.

Finally, having multiple views, the interaction allows us to explore the space of all tunnels in 2D using scatterplot and at the same time to compare spatial characteristics of the individual unfolded tunnels. The individual steps of our algorithm are described in the following sections in more detail.

3.1. Unfolding Process

The input data consists of a triangle mesh representing the tunnel shape, a graph of connected points representing the tunnel centerline and information about spatial and physico-chemical properties of tunnel-lining amino acids. We need to clearly communicate three features ordered from the most to the least important. First is the position and distribution of amino acids that form the tunnel boundary. The second feature is the notion of the tunnel width meaning that narrower and wider areas of the tunnel should be communicated adequately to the user. The narrowest part of the tunnel, called bottleneck, is usually the most important one because it determines the size of a ligand molecule which can follow this tunnel. The third feature is the spatial information about the tunnel length.

As already mentioned, in order to avoid occlusion issues we represent the tunnel as a 2D image. The general idea of our unfolding method is based on the centerline parametrization [KFW*02]. We define the centerline of each tunnel as curve $\mathbf{c}(t) : [t_0, t_1]$, where t_0 is positioned in the vicinity of the reaction site inside the molecule and t_1 is located at the molecular surface. Using this definition any point on the tunnel surface can be described by two parameters: distance t along the centerline and angle a around it. Therefore, in order to obtain the unfolded view, we firstly cut the tunnel along its centerline in uniform steps, which produces a set of contours. Each contour is computed as the intersection of a plane perpendicular to the centerline at position $\mathbf{c}(t)$ with the tunnel mesh. Due to the centerline curvature two or more neighboring cuts may intersect. To avoid this unwanted situation we employ a simple smoothing method proposed for similar purposes in the MoleCollar [BJG*15] technique. In this approach the normal of the cutting plane is computed as an average vector of directions from ten neighboring points on the centerline.

With the second parameter a we uniformly sample points on each contour every 10°. Hence, the samples at positions $a = 0$ and $a = 36$ are the same while the value $a = 18$ corresponds to the sample on the opposite side of the contour (see Figure 3, middle). However, the results depend on the precise location of the point at the contour from where we start the sampling. In our case the starting point is defined by so called cutting vector $\mathbf{v}$. The cutting vector $\mathbf{v}$ is a vector perpendicular to the centerline such that it points in the direction between two most stable amino acids in the first ten slices, i.e., those that do not change their relative position significantly.

However, having a global cutting vector $\mathbf{v}$ might introduce un-

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

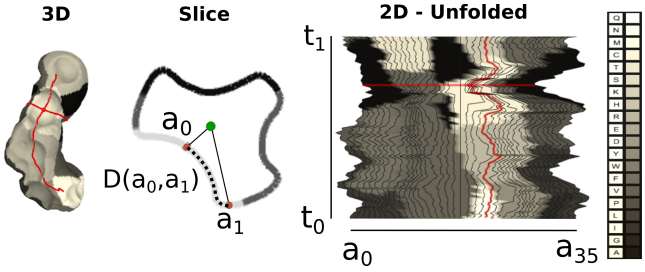

Figure 3: *We cut the tunnel along its centerline. Then for each point on the centerline we create one slice. Each slice is uniformly sampled $(a_0, a_1, \ldots a_{35})$ around centerline point (green). Additionally, distances between neighboring samples $(D(a_0,a_1),\ldots)$ are stored as well. By straightening the cuts and putting them on top of each other we unfold the tunnel. The unfolded representation resembles a map where each region corresponds to one amino acid. The density of the horizontal curves (each representing the evolution of uniform sampling $a_0, \ldots a_{35}$) helps to understand the tunnel width.*

wanted artifacts due to the curvature of the centerline. To overcome this problem, we compute the cutting vector only for the first contour at position $\mathbf{c}(t_0)$. Then for every subsequent contour we simply project the previous cutting vector onto the current cutting plane.

Another reason for computing the cutting vector only at the beginning of the tunnel is that the most stable amino acids wrt. their spatial position can be found close to the deeply buried protein active site while the amino acids closer to the protein surface are usually much more flexible. This assumption also allows us to create one of the common dimensions for the comparative visualization, as the amino acids used for the cutting vector are supposed to be stable in all time steps of the molecular dynamics simulation. There are several other ways to define the cutting vector (including manual selection) but for our current needs this setting is sufficient.

In order to reconstruct the final 2D view from the computed samples we need to transfer the obtained information (including the information about the closest amino acids) from 3D to the reduced 2D coordinate system. Here we simply utilize the two already described parameters t and a. Parameter t can be used directly since its value ranges from 0 to 1 and corresponds to the length of the centerline. On the other hand, if we use parameter a without any transformation we lose the information about the tunnel width. Therefore, for each contour we also compute the distance $D(a_i, a_{i+1})$ between every two neighboring samples a_i and a_{i+1} (see Figure 3, middle). This distance is then used in order to transform the position of the sample in the final 2D space. We are able to reconstruct the circumference of the contour as sum $\sum_{l=0}^{35} D(a_l, a_{l+1})$.

Thus, our unfolding method allows us to encode the tunnel surface values in 2D. Moreover, the size of the tunnel along the centerline is preserved as well as the size and position of the amino acids around the surface and their respective neighborhood.

By connecting the samples with the same value of the parameter a we get a shape description that communicates the changes of the tunnel shape. For example, at the place where the lines are dense the tunnel is narrower. On the other hand, more scattered lines signify that the tunnel is wider (see Figure 3).

Nevertheless, the core benefit of the unfolded tunnel representation lies in the ability to perform a comparative visualization. Juxtaposition or superposition of the unfolded tunnels give an overview of the spatial and content-wise differences between different tunnels.

Note that in case of a smaller number of tunnels a side by side comparative visualization represents a natural choice. However, when analyzing molecular dynamics simulations one can obtain hundreds of time frames that are worthwhile. As the tunnel changes over time, together with the whole molecule, also its shape and physico-chemical properties are changing. To show the similarity/dissimilarity of tunnels one needs to have a suitable descriptor for all instances. Since the unfolded representation contains all the relevant information in 2D, i.e. forming an image, we opted for image moments as the main set of descriptors.

3.2. Image Moments

Image moments [FSZ09] provide a practical way to characterize individual tunnels since the acquired values capture the similarities and dissimilarities between tunnel images, i.e., they produce similar values for similar images and more distinct for different ones. A gray scale image with pixel intensities $I(x,y)$ can be described by the following set of moments (of order $i+j$):

$$M_{ij} = \sum_x \sum_y x^i y^j I(x,y) \tag{1}$$

There are various versions of image moments, each of them is invariant to different transformations, such as rotation, scaling, or translation. For example, centralized moments are invariant to translation, which is a valuable property in our case since some amino acid areas might change their longitudinal position with respect to the length of the centerline. Centralized image moments for an image $I(x,y)$ are defined as:

$$\mu_{ij} = \sum_x \sum_y (x - x_c)^i (y - y_c)^j I(x,y) \tag{2}$$

where x_c and y_c are components of the centroid:

$$x_c = \frac{M_{10}}{M_{00}} \qquad y_c = \frac{M_{01}}{M_{00}}$$

Adapting to the grayscale requirement for the image moments, we have encoded the amino acids using grayscale values (from white to black). Because the unfolding process preserves the centerline length and the area of amino acids, no scaling is necessary as a preprocessing step before computing the image moments. This way we are able to characterize each tunnel by the vector of values with the central moments of order up to three to condense the information. After several empirical tests with different combinations of image moments we opted for the following set I_m of 10 image moments that was found to be sufficient to describe the differences between individual tunnels:

$$I_m = (\mu_{00}, 0, 0, \mu_{11}, \mu_{02}, \mu_{20}, \mu_{12}, \mu_{21}, \mu_{03}, \mu_{30}) \tag{3}$$

Having such a vector for each tunnel, we can use dedicated approaches from information visualization, or interactive visual analysis [ODH*07], to get an overview about the similarity of those

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

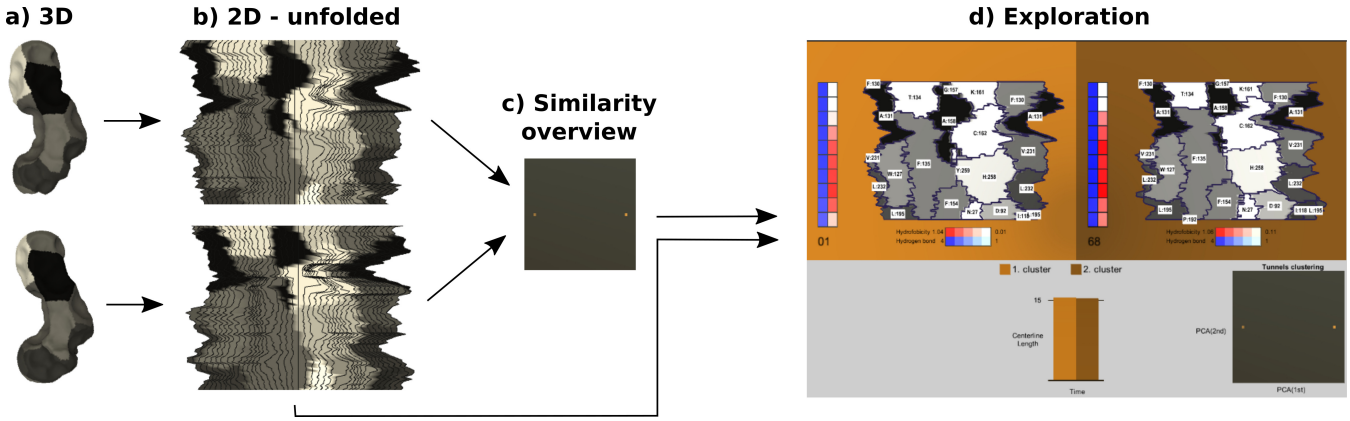

Figure 4: *Processing pipeline. Each tunnel (a) is unfolded (b), described with image moments, and processed to a 2D point of similarity/dissimilarity information (c). Eventually, the tunnels are shown in multiple views (d). The color scale on the left side of the unfolded representation shows the summary of additional information in the sparser sections of the tunnel. Here the properties are hydrophobicity and hydrogen bonds.*

tunnels. In our case, emphasizing outliers is very helpful. We suggest to use a distance matrix containing the distance of 2D PCA points, i.e., each matrix entry i, j represents the distance between tunnel i and j. Here the bar chart height would stand for the row-wise sum of these distances indicating how isolated the tunnel is with respect to other tunnels.

3.3. Similarity Overview

We are able to set up an m-by-n tunnels matrix IM, whose rows, m, consist of the centralized image moment vectors I_m for each tunnel.

When dealing with a tunnel in molecular dynamics, the matrix IM contains hundreds of rows. To make such an ensemble understandable, we project the IM matrix of tunnel image moments to 2D, where each point represents one tunnel. Here, the important requirement is that the similarity between the tunnels must be preserved as much as possible.

First, the matrix dimension needs to be reduced. There are several approaches for dimensionality reduction, such as Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA) [KFS05], Independent Component Analysis (ICA) [OF97], t-Distributed Stochastic Neighbor Embedding (t-SNE) [vdMH08] and others. The important property of the approach is the ability to reduce the ten floating point values tunnel descriptor to one point in 2D coordinates and to visualize the statistical results for the whole matrix. Visualizing statistical results can help to identify clusters and outliers as well as to analyze deviation, distribution, and correlation. Because most of the approaches produce similar results, and for its simplicity, we have adopted PCA for dimension reduction. PCA is a statistical tool that transforms a set of observations, in our case a set of image moment vectors, to the set of linearly uncorrelated variables. Each of the resulting components is orthogonal to each other and the number of components is lower than or equal to the number of the original variables. The components are ordered by the variance value meaning that the first component has the largest possible variance and each succeeding component has a lower variance. In our overview, we showcase the first two PCA components for the tunnel, which are encoded in the coordinates. The first component value encodes the x coordinate, and the second encodes the y coordinate (see Figure 4c).

Our 2D overview is just a simple demonstration, however, more complex visualization method can be used instead, for example Combined design [FTB*13] to show multiple descriptors at the same time. This is one of the possible future extensions of our method.

3.3.1. Clustering

Having the visual representation of the similarity between tunnels can be essential for visual navigation through possible tunnel arrangements, which was also shown and discussed in the work of van den Elzen et al. [vdEHBvW16] on the example of dynamic networks. With points spread in 2D space it is easy to spot densely populated areas (clusters), and also points being further away from the others (outliers). For a successful linking with other views, one needs to establish a proper color assignment of clusters and outliers. Here, an additional explicit clustering allows us to color groups of points to determine patterns in the overview.

There are different clustering algorithms that could be applied, for example hierarchical clustering, or the k-means approach [Mac67]. In k-means clustering the number of clusters is fixed to k and needs to be specified in advance, which is considered to be one of the main drawbacks of this algorithm. In our case, we would like to show an unfolded representative from each cluster. As this representation requires available space in window, the number of unfolded tunnels is limited. We have empirically found out that showing five tunnels at the same time is the limit that also gives us a suitable amount of clusters for k-means clustering, which transforms a k-means drawback to our advantage. Moreover k-means was selected because of its simplicity and for its conceptual closeness to the nearest neighbor classification. However, a well known prob-

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

lem relates to the borders of clusters, since the algorithm optimizes cluster centers, not cluster borders. This represents another direction of our future work, where we will conduct experiments with other families of cluster algorithms such as hierarchical clustering, density clustering, and other methods from machine learning.

The image moment descriptors, dimension reduction, and clustering should all bring the same notion of tunnels similarity/dissimilarity. As can be seen in Figure 5, we tested the generation of the overview scatterplot with four different ways of generating the value, which are: standard, tunnels rotated, tunnels scaled, and tunnels with amino acids colored by the different color scheme. Coloring by a different coloring scheme has a similar effect as cutting by different cutting vector. The overall shape of the amino acid areas is the same, only the color distribution is different. We can see that even though the positional distribution of the points differs, the notion of cluster and three outliers is preserved in all explored ways of generating image moment values.

Moreover, the important aspect here is that for the overview scatterplot we have a shared space which allows us to see the neighborhood of the tunnels, which is our requirement for employing comparative visualization as a part of the exploration process.

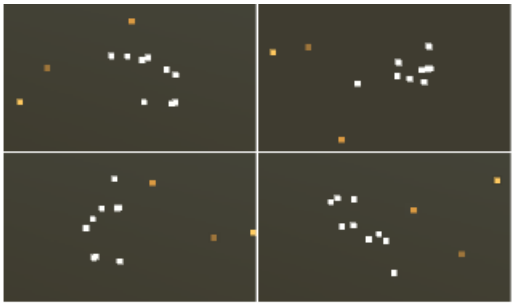

Figure 5: *The tunnel similarity overview for four different ways of generating image moments: standard (upper-left), with amino acids coloring permutation (upper-right), rotated (lower-left), and with normalized size of the centerline (lower-right). Points are colored according to the detected clusters.*

3.4. Exploration

So far we have described individual components of our technique. All of them play a crucial role in the exploration process. Therefore, our final exploration solution consists out of the following views (see Figure 4d). First, we introduce a view showing the set of unfolded tunnels, which allows us to make a comparison of spatial and structural aspects of the tunnel. Second, we exploit scatterplot showing the similarity/dissimilarity overview that gives the notion of the differences between tunnels. Finally, we employ a bar chart view where each tunnel is represented by one bar and the height of the bar corresponds to the tunnel length.

Using the unfolding tunnel view the user is able to intuitively compare the width of the tunnels, the position of their bottleneck, and the set of tunnel-lining amino acids along with their extent of influence. In order to communicate various physico-chemical properties of individual amino acids we use color mapping. However, using this approach we are able to visualize only a single property at once. Hence we provide also more aggregated view situated next to the unfolded tunnel image. This view communicates the average value for each observed physico-chemical property computed for the tunnel-lining amino acids.

The unfolded tunnels are positioned side by side, or in other words by using the juxtaposition composition. Positioning tunnels from different time frames allows the user to see the behavior of the bottleneck as well as the amino acid areas of influence. However, the amino acid areas on the unfolded tunnel might be too small. For a proper examination of the tunnel composition of amino acids, it is necessary to limit the downscale of the unfolded tunnel representation as well as the number of visualized tunnels. If more tunnels should be visualized, they need to be additionally downscaled, which means that with more tunnels more spatial details are discarded and the visual representation has to be scaled.

To show the dataset of hundreds of tunnels from the molecular dynamics simulation, it is unfeasible to use the unfolding view to get the overview information about tunnel changes. Easily shown in an example – with the resolution width 1900 and more than 1900 tunnels we would have more unfolded tunnels than available pixels. Therefore, we have constructed a similarity overview which allows us to code one tunnel as a point. Having here too many tunnels is still feasible as we are here more interested in tunnels similarity/dissimilarity represented by a groups/outliers.

The last missing link is to have an overview of the tunnels with respect to time. This requirement led us naturally to the bar chart representation, where the height of bars corresponds to the tunnel length. This way we have a view which allows us to have a unique and well distinguishable representation for each tunnel. Moreover, each tunnel is consistently colored by the group colors in the scatterplot and barchart and the same color is used for the background in the unfolded tunnel view.

The real power of the exploration lies in the possibility of brushing between the similarity overview (scatterplot) and tunnel length view (bar chart). Selecting one point in the scatterplot highlights the corresponding tunnel in the bar chart and shows the unfolded content of the tunnel. Naturally, this selection works also in the opposite direction when selecting a bar highlights the corresponding point in the scatterplot in order to easily recognize whether the tunnel has a stable shape or if it is an outlier.

Overall, the user has the power to explore the spatial and content-wise aspects of the tunnels by selecting the desired tunnels in the bar chart and the scatterplot. Thanks to the bar chart it is possible to visualize the tunnel development over time and scatterplot helps to explore the similarity of the tunnel and to spot the outliers.

4. Evaluation

To demonstrate the usability of our approach we conducted two case studies following the research of our cooperating researchers from the field of protein engineering. The first study deals with the exploration of simulations of molecular dynamics and the second study focuses on scrutinizing the tunnel shape modified by mutations of a selected tunnel-lining amino acid.

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

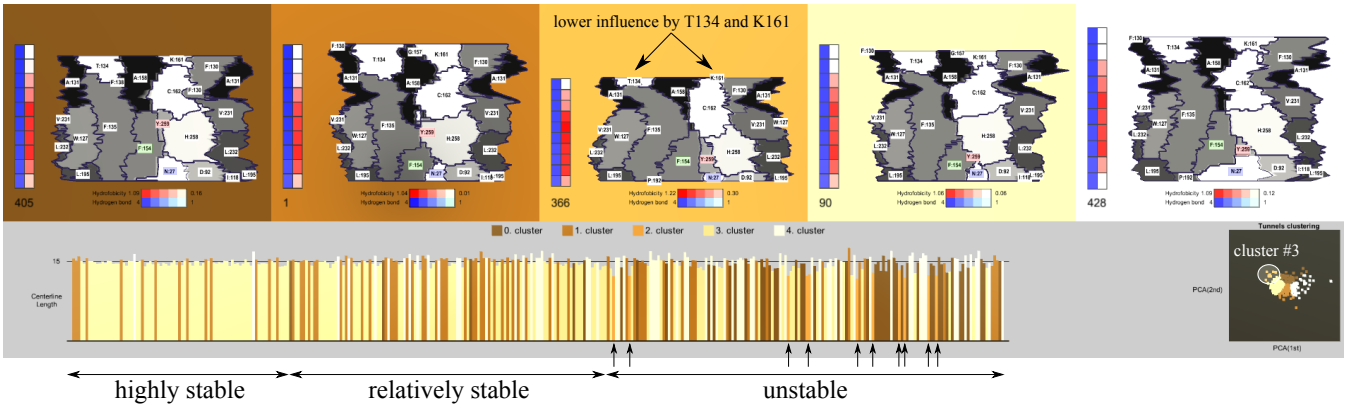

Figure 6: *Visualization of sequence of 350 time steps of the DhaA haloalkane dehalogenase protein. The bar chart shows the tunnel stability while the scatterplot provides an information about different outliers. The unfolded representation then depicts the influence of individual amino acids.*

4.1. Exploration of Molecular Dynamics

In this case the task is to explore and evaluate a single tunnel in each time step (or a selected subset) of a molecular dynamics simulation. We first derive the unfolded 2D representation and show the cluster representatives, which are the closest samples from the individual clusters centroids. These images are subsequently scrutinized to reveal similarities between them. This produces clusters of tunnels according to their similarity which are visualized using the scatterplot. Additionally, the coloring of individual clusters helps us to find the correspondence between the scatterplot and the bar chart representation where we can see the positions of tunnels belonging to these clusters spread over time. The bar chart immediately shows the time portions in which the tunnel was stable or unstable (depending of the interest of the researcher). Figure 6 shows the resulting representation for a sequence of 350 time steps of the DhaA haloalkane dehalogenase protein. From the bar chart view we can clearly see that the tunnels shape is stable in the first part of the simulation while it becomes more and more unstable towards the end. On the other hand, from the unfolded representation it is obvious that during the stable period (clusters #2 and #4) there was almost no influence of Tyrosine (Y:259 highlighted by red). This changes later when also the influence of the neighboring amino acids, such as Phenylalanine F:154 and Asparagine N:27 is starting to change significantly. Moreover, using our tool, the biochemists were immediately able to spot several outliers belonging to the cluster #3 (see vertical arrows in bottom of Figure 6). This cluster represents tunnels which are significantly shorter then others. The same cluster can be also easily identified in the similarity overview where this whole cluster is clearly separated from the rest (see white circle in Figure 6).

By clicking on the individual bars in the bar chart the user can select a subset of tunnels to be explored in more detail. These selected time steps are immediately visualized using the unfolded representation and the tunnels are juxtaposed to enable the comparison of the width and constitution of the tunnel.

The biochemists testing our approach confirmed that the first main benefit is in the navigation phase where they were able to immediately spot the time steps with significant tunnel changes which are usually the best candidates for further exploration. In the subsequent exploration phase using the unfolded view they appreciated the information about the area of influence of the amino acids and their mutual position. Some of them were slightly confused about the fact that some amino acids are divided by the cut into two parts which makes the comparison task more complicated. However, this can be solved by interactive manipulation with the cut position or in the future also by introducing more sophisticated method for defining the cut. But in summary they confirmed that this tool has a high potential for this type of tasks.

4.2. Exploration of Tunnel Mutations

One of the research topics conducted by the cooperation group of protein engineers is dealing with engineering protein stability and resistance to organic cosolvents [KCB*13]. This is accomplished by mutating the amino acids around the main tunnel of the DhaA haloalkane dehalogenase protein molecule. The method first presents the confirmation of importance of the amino acids surrounding the tunnel by presenting the results of several mutations of amino acids on different places in the molecule. It was revealed that by mutating only the amino acids in the close vicinity of the main tunnel the researchers were able to change the stability and resistance of the protein substantially. The biochemists aimed to find the best possible mutation around the main tunnel which would increase the stability of the dehalogenase and its resistance to organic cosolvents. The researchers revealed that the DhaA80 mutation in which the amino acid on position 176, which in the wild type form contains Cysteine, was replaced by Phenylalanine. This mutation closed the main tunnel which caused the increase in melting temperature of the protein by 19 degrees and the resistance to cosolvent dimethyl sulfoxide was increased 4000 times. This situation is clearly visible in our representation as well. Figure 7 demonstrates the situation when we performed multiple mutations of a single amino acid in position 176 in the protein chain. It is obvious that if the original amino acid (Figure 7 – CYS) is replaced by similar or even a smaller one, such as Glutamic Acid or Glycine

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

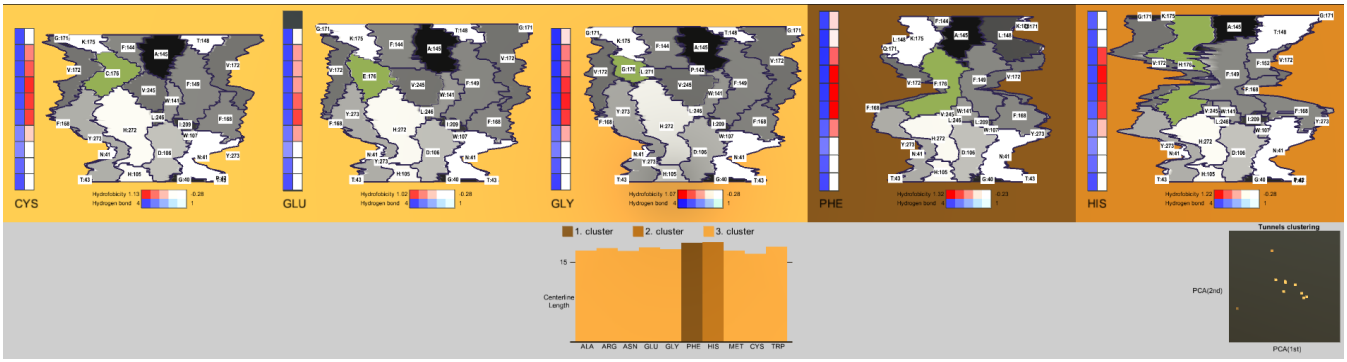

Figure 7: *Visualization of different mutations of the amino acid with ID 176 (highlighted in green) surrounding the main tunnel in DhaA haloalkane dehalogenase. The scatterplot and bar chart show the information about all tunnels in the dataset whereas the unfolded representation shows only one selected representative of each detected cluster. The amino acid is highlighted for better comparison.*

(Figure 7 – GLU and GLY respectively) the tunnel becomes wider. On the other hand, using bigger amino acids, such as Phenylalanine or Histidine (Figure 7 – PHE and GLY respectively) decreases the bottleneck of the tunnel significantly. This corresponds exactly to the outcomes of the real in-vitro experiments which were conducted by the biochemists.

When evaluating the suitability of our approach to this type of tasks the biochemists appreciated the unfolded view much more than in the case of molecular dynamics simulations. Here the information about the changes to the tunnel surface and its width caused by the mutations is crucial and it is clearly visible from our representation. Moreover, in this particular case when the user is interested in a specific amino acid, our tool enables to highlight it in order to even enhance the comparison. The biochemists confirmed that similar information is very hard to understand from 3D tunnel representation, namely when aiming to compare more than two or three mutated tunnels at once. Therefore, they appreciated our representation and suggested its future extension by calculating and highlighting areas with the most significant changes.

5. Conclusion & Future work

In this paper, we present and demonstrate the power of a method comprised of tunnel unfolding, similarity overview, and connected selection in order to support the actual application of exploring the shape and content of entire ensembles of tunnels in protein molecules. The resulting system can interactively show the overall development of the tunnel centerline, the tunnel similarities and dissimilarities, and the shape and content of a selected set of tunnels.

Even with simple methods used in some steps (such as PCA and k-means), we were able to achieve a solution that proved to be helpful for the domain experts. We demonstrated the usage of the method in two case studies from the field of biochemistry, which were proposed and evaluated by the domain experts. This confirmed that our method can be used for time-dependent data as well as for datasets containing tunnel mutations.

Most of the running time is spent on the generation of the unfolded representations. In our CPU-based implementation it takes $\approx$ 2-3 seconds for each tunnel. This is required as a precomputation in order to generate the image moments for all tunnels, and to generate points for the similarity overview. For hundreds tunnels it takes minutes to compute all of them and to prepare all the views. Therefore, one of the future directions can be the parallelization of the unfolding process.

Certainly, it would be interesting to investigate more closely whether more advanced, alternative techniques for individual steps (non-linear embedding, bottom-up clustering, ...) could further improve the results. Regarding the unfolding the domain experts proposed the possibility to have a more sophisticated cutting mechanism which will enable more semantically-driven cut that would follow the borders of the amino acids in the tunnel. For the similarity overview, it is worthwhile to try the combination of image moment, reduction, and clustering, that is invariant to amino acids color permutation, rotation, and scale. From the obtained feedback also new ideas about the final visualization composition emerged. For example, one suggestion is to incorporate the unfolded overview for the whole cluster. It means to visualize the clusters side by side, where each cluster would not be seen by one representative, but rather by the variance visualization showing the stability of the cluster.

6. Acknowledgement

This work was supported through grants from the Norway grants project NF-CZ07-MOP-2-086-2014, and the PhysioIllustration research project 218023 funded by the Norwegian Research Council.

References

[AH11] ANGELELLI P., HAUSER H.: Straightening tubular flow for side-by-side visualization. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 2063–2070. 3

[AMB*13] AUZINGER T., MISTELBAUER G., BACLIJA I., SCHERNTHANER R., KOCHL A., WIMMER M., GRÖLLER M. E., BRUCKNER S.: Vessel visualization using curved surface reformation. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 2858–2867. 3

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

[BAAH16] BRAMBILLA A., ANGELELLI P., ANDREASSEN Ø., HAUSER H.: Comparative visualization of multiple time surfaces by planar surface reformation. In *2016 IEEE Pacific Visualization Symposium (PacificVis)* (2016), IEEE, pp. 88–95. 3

[BCG*13] BREZOVSKÝ J., CHOVANCOVÁ E., GORA A., PAVELKA A., BIEDERMANNOVÁ L., DAMBORSKÝ J.: Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnology Advances 31*, 1 (2013), 38–49. 3

[BJG*15] BYŠKA J., JURČÍK A., GRÖLLER M. E., VIOLA I., KOZLÍKOVÁ B.: MoleCollar and Tunnel Heat Map visualizations for conveying spatio–temporo–chemical properties across and along protein voids. *Computer Graphics Forum 34*, 3 (2015), 1–10. 3, 4

[BLMG*15] BYŠKA J., LE MUZIC M., GRÖLLER M. E., VIOLA I., KOZLÍKOVÁ B.: AnimoAminoMiner: Exploration of protein tunnels and their properties in molecular dynamics. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 747–756. 3

[CPB*12] CHOVANCOVÁ E., PAVELKA A., BENEŠ P., STRNAD O., BREZOVSKÝ J., KOZLÍKOVÁ B., GORA A. W., ŠUSTR V., KLVAŇA M., MEDEK P., BIEDERMANNOVÁ L., SOCHOR J., DAMBORSKÝ J.: CAVER 3.0: A tool for the analysis of transport pathways in dynamic protein structures. *PLOS Computational Biology 8*, 10 (2012). 3

[FSZ09] FLUSSER J., SUK T., ZITOVA B.: *Moments and Moment Invariants in Pattern Recognition*. John Wiley & Sons, Ltd, 2009. 4, 5

[FTB*13] FANG H., TAM G. K. L., BORGO R., AUBREY A. J., GRANT P. W., ROSIN P. L., WALLRAVEN C., CUNNINGHAM D., MARSHALL D., CHEN M.: Visualizing natural image statistics. *IEEE Transactions on Visualization and Computer Graphics 19*, 7 (2013), 1228–1241. 6

[GAW*11] GLEICHER M., ALBERS D., WALKER R., JUSUFI I., HANSEN C. D., ROBERTS J. C.: Visual comparison for information visualization. *Information Visualization 10*, 4 (2011), 289–309. 3

[GSZ*13] GURIJALA K. C., SHI R., ZENG W., GU X., KAUFMAN A.: Colon flattening using heat diffusion riemannian metric. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 2848–2857. 3

[HMH*12] HENSEN U., MEYER T., HAAS J., REX R., VRIEND G., GRUBMUELLER H.: Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. *PLoS ONE 7*, 5 (2012), e33931. 2

[JBSK15] JURČÍK A., BYŠKA J., SOCHOR J., KOZLÍKOVÁ B.: Visibility–based approach to surface detection of tunnels in proteins. In *31th Proceedings of Spring Conference on Computer Graphics* (Bratislava, Slovakia, 2015), Jorge J., Santos L. P., Ďurikovič R., (Eds.), Comenius University, pp. 85–92. 3

[KCB*13] KOUDELÁKOVÁ T., CHALOUPKOVÁ R., BREZOVSKÝ J., PROKOP Z., ŠEBESTOVÁ E., HESSELER M., KHABIRI M., PLEVAKA M., KULIK D., KUTÁ SMATANOVÁ I. K., ŘEZÁČOVÁ P., ETTRICH R., BORNSCHEUER U. T., DAMBORSKÝ 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), 1959–1963. 8

[KFS05] KIM K. I., FRANZ M. O., SCHOLKOPF B.: Iterative kernel principal component analysis for image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 9 (2005), 1351–1366. 6

[KFW*02] KANITSAR A., FLEISCHMANN D., WEGENKITTL R., FELKEL P., GRÖLLER M.: CPR - curved planar reformation. In *Visualization, 2002. VIS 2002. IEEE* (Nov 2002), pp. 37–44. 2, 3, 4

[KKL*16] KRONE M., KOZLÍKOVA B., LINDOW N., BAADEN M., BAUM D., PARULEK J., HEGE H.-C., VIOLA I.: Visual analysis of biomolecular cavities: State of the art. *Computer Graphics Forum 35*, 3 (2016). 3

[KST*14] KRETSCHMER J., SOZA G., TIETJEN C., SUEHLING M., PREIM B., STAMMINGER M.: ADR-anatomy-driven reformation. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (2014), 2496–2505. 3

[LBBH13] LINDOW N., BAUM D., BONDAR A.-N., HEGE H.-C.: Exploring cavity dynamics in biomolecular systems. *BMC Bioinformatics 14*, Suppl 19 (2013), S5. 3

[LBH11] LINDOW N., BAUM D., HEGE H.-C.: Voronoi–based extraction and visualization of molecular paths. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2025–2034. 3

[LCMH09] LAMPE O. D., CORREA C., MA K.-L., HAUSER H.: Curve-centric volume reformation for comparative visualization. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 1235–1242. 4

[LZX*08] LIU L., ZHANG L., XU Y., GOTSMAN C., GORTLER S. J.: A local/global approach to mesh parameterization. In *Computer Graphics Forum* (2008), vol. 27, Wiley Online Library, pp. 1495–1504. 3

[Mac67] MACQUEEN J.: Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (1967), University of California Press, pp. 281–297. 6

[MBS07] MEDEK P., BENEŠ P., SOCHOR J.: Computation of tunnels in protein molecules using Delaunay triangulation. In *Journal of WSCG* (2007), pp. 107–114. 3

[MSCN15] MASOOD T. B., SANDHYA S., CHANDRA N., NATARAJAN V.: CHEXVIS: a tool for molecular channel extraction and visualization. *BMC Bioinformatics 16*, 1 (2015), 1–19. 3

[MVB*12] MISTELBAUER G., VARCHOLA A., BOUZARI H., STARINSKY J., KÖCHL A., SCHERNTHANER R., FLEISCHMANN D., GRÖLLER M. E., ŠRÁMEK M.: Centerline reformations of complex vascular structures. In *Visualization Symposium (PacificVis), 2012 IEEE Pacific* (2012), IEEE, pp. 233–240. 2, 3

[ODH*07] OELTZE S., DOLEISCH H., HAUSER H., MUIGG P., PREIM B.: Interactive visual analysis of perfusion data. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1392–1399. 5

[OF97] OLSHAUSEN B. A., FIELD D. J.: Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research 37*, 23 (1997), 3311 – 3325. 6

[POB*06] PETŘEK M., OTYEPKA M., BANÁŠ P., KOŠINOVÁ P., KOČA J., DAMBORSKÝ J.: CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics 7*, 1 (2006), 316. 3

[PTRV13] PARULEK J., TURKAY C., REUTER N., VIOLA I.: Visual cavity analysis in molecular simulations. *BMC Bioinformatics 14*, Suppl 19 (2013), S4. 3

[vdEHBvW16] VAN DEN ELZEN S., HOLTEN D., BLAAS J., VAN WIJK J. J.: Reducing snapshots to points: A visual analytics approach to dynamic network exploration. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 1–10. 6

[vdMH08] VAN DER MAATEN L., HINTON G.: Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research 9* (2008), 2579–2605. 6

[VG10] VOSS N. R., GERSTEIN M.: 3V: cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Research 38*, Web Server issue (Jul 2010), W555–W562. 3

[WPS07] WEISEL M., PROSCHAK E., SCHNEIDER G.: Pocketpicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central Journal 1*, 1 (2007), 1. 4

[YFW*08] YAFFE E., FISHELOVITCH D., WOLFSON H., HALPERIN D., NUSSINOV R.: MolAxis: Efficient and accurate identification of channels in macromolecules. *Proteins 73*, 1 (2008), 72–86. 3

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.