AnimoAminoMiner: Exploration of Protein Tunnels and their Properties in Molecular Dynamics

Jan Byška, Mathieu Le Muzic, M. Eduard Gröller, Ivan Viola, and Barbora Kozlíková

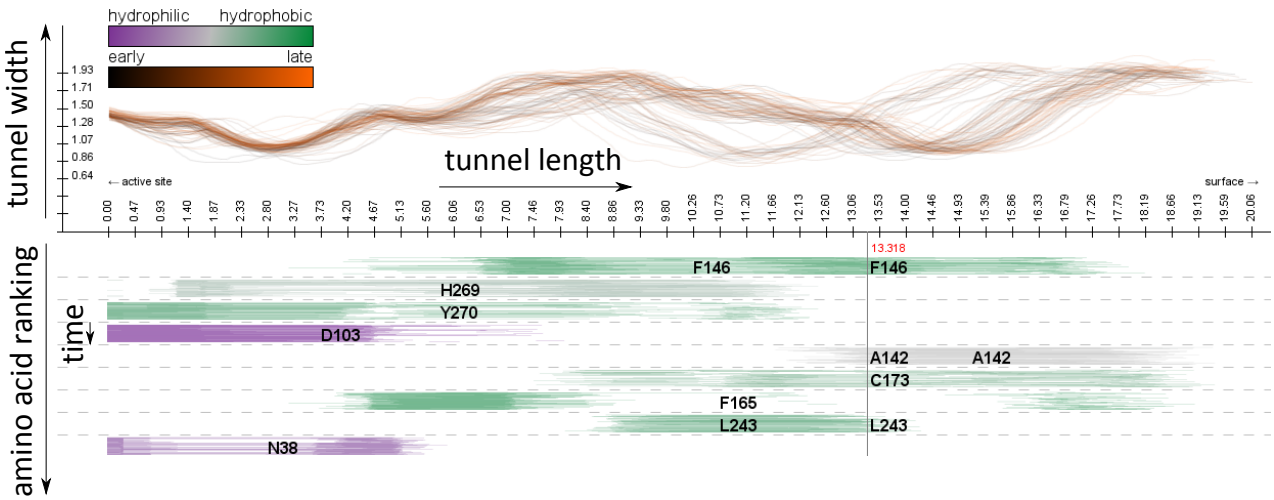

Fig. 1. Proposed AnimoAminoMiner representation for the abstracted overview of the protein-tunnel evolution over time. The top part, TunExplorer, shows the changes in tunnel width along its centerline, starting from the active site on the left side. The bottom part, AAExplorer, represents all amino acids surrounding the tunnel at a particular centerline location in molecular dynamics. The vertical arrangement of strips represents individual amino acids and can be sorted with respect to preferred ranking strategies.

Abstract— In this paper we propose a novel method for the interactive exploration of protein tunnels. The basic principle of our approach is that we entirely abstract from the 3D/4D space the simulated phenomenon is embedded in. A complex 3D structure and its curvature information is represented only by a straightened tunnel centerline and its width profile. This representation focuses on a key aspect of the studied geometry and frees up graphical estate to key chemical and physical properties represented by surrounding amino acids. The method shows the detailed tunnel profile and its temporal aggregation. The profile is interactively linked with a visual overview of all amino acids which are lining the tunnel over time. In this overview, each amino acid is represented by a set of colored lines depicting the spatial and temporal impact of the amino acid on the corresponding tunnel. This representation clearly shows the importance of amino acids with respect to selected criteria. It helps the biochemists to select the candidate amino acids for mutation which changes the protein function in a desired way. The AnimoAminoMiner was designed in close cooperation with domain experts. Its usefulness is documented by their feedback and a case study, which are included.

Index Terms—Protein, tunnel, molecular dynamics, aggregation, interaction

1 Introduction

Proteins are the molecules responsible for driving the machinery of life. Understanding their function helps to reveal the fundamentals of biochemical processes that are taking place in living cells. Based on this knowledge, biologists are able to conduct mutations on proteins, in order to affect their reactivity and thereby alter their functions. These mutations change the constitution of the protein. Mutated proteins play a crucial role in many research areas, such as protein engineering or drug design, and may help to find new chemical matters for different industry fields, e.g., new inhibitors or improved biocatalysts, and cures for various diseases.

The protein function is highly influenced by its structure. Proteins consist of one or more chains of amino acids. The chain often folds into an energetically favorable configuration forming a hydrophobic core and produces the three-dimensional arrangement of the protein. This spatial arrangement has a deep impact on the protein reactivity, which is another important factor influencing the protein function. Proteins interact with other molecules via chemical reactions which happen directly on their surface or in the hollow parts inside their own structure. In the case of a protein reacting with smaller molecules, also known as ligands, the reaction often takes place deep inside the protein inner void. The path which connects the molecular surface with the reaction site forms a so called tunnel. The tunnel represents an accessible path for a ligand. As the reactivity of a protein with a given ligand strongly depends on this accessibility, the characteristics of such tunnels are carefully studied by biochemists.

- *Jan Byška is with Masaryk University, Czech Republic. E-mail: xbyska@fi.muni.cz.*
- *Mathieu Le Muzic is with TU Wien, Austria. Email: mathieu@cg.tuwien.ac.at.*
- *M. Eduard Gröller is with TU Wien, Austria and University of Bergen, Norway. Email: groeller@cg.tuwien.ac.at.*
- *Ivan Viola is with TU Wien, Austria and University of Bergen, Norway. Email: viola@cg.tuwien.ac.at.*
- *Barbora Kozlíková is with Masaryk University, Czech Republic and TU Wien, Austria. E-mail: kozlikova@fi.muni.cz.*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication 20 Aug. 2015; date of current version 25 Oct. 2015. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org. Digital Object Identifier no. 10.1109/TVCG.2015.2467434

Recent studies [4, 22] revealed that protein function is not solely influenced by its static structure. Molecular dynamics play an important role as well. The movements of the protein highly influence also the shape and openness of tunnels. In consequence, studying the ligand accessibility of a protein tunnel over time offers a relevant insight into the functioning of a protein [8]. In order to explore the tunnels over time, we need to simulate molecular dynamics. The continuous advances in hardware capabilities are providing the possibility to capture longer molecular dynamics simulations, currently reaching hundreds of thousands of time steps. These simulations are represented by a set of consecutive snapshots in a given file format which capture the molecular dynamics in discrete time steps. The time interval between two time steps is usually in femtoseconds. Several existing methods are available which analyze such simulations with respect to the detection of protein tunnels [3, 21, 25].

Tunnel characteristics, such as length or width, are highly influenced by the conformation of the surrounding amino acids and their physico-chemical properties. When designing a new protein with differing properties, biochemists are searching for changes (mutations) of its structure which influence the desired property the most. Among such properties are, e.g., the thermal stability [9] or the activity [18] of a protein. It was found out that mutations of amino acids in the vicinity of a given tunnel have the largest impact on these properties. It is tightly related to altering the shape of a given tunnel, i.e., either enlarging or closing it. Therefore, information related to the surrounding amino acids is particularly relevant since it can ease the decision process of choosing a candidate for a targeted mutation of amino acids. Currently available visual representations of tunnels in molecular dynamics showcase the changes in tunnel shape and conformation of lining amino acids by animating their movements (see Figure 2).

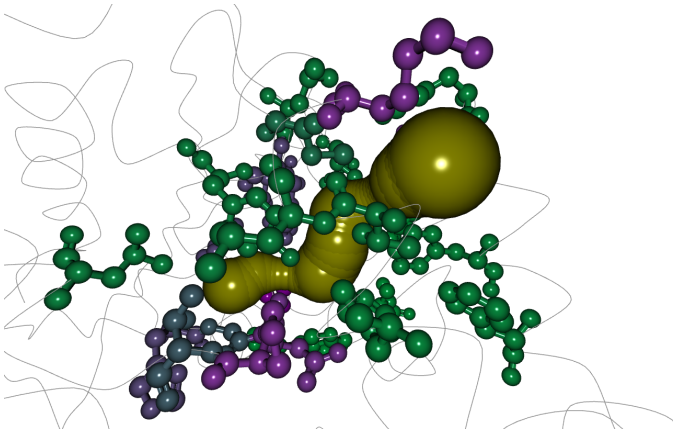

Fig. 2. 3D visualization of a tunnel (in green) and its surrounding amino acids. The amino acids are colored with respect to their hydrophobicity.

The main problem of these methods is that they do not enable fast and intuitive exploration of a given tunnel and its surrounding amino acids. The user has to follow all snapshots in the molecular dynamics, which becomes impracticable for large simulations. In other words, a general overview of the tunnel behavior over time is completely missing.

In this paper we overcome this limitation by introducing a novel aggregate visual representation of a tunnel over time. It focuses on the interactive exploration of various properties (e.g., hydrophobicity or partial charge) of the lining amino acids and their influence on geometric tunnel properties, such as its width and length. The design of the new representation was tailored to fit the most important needs of the domain experts. Among these crucial aspects is the possibility to study the behavior of the whole tunnel over long-time simulations at once. Here observing the changes of the tunnel profile is essential. The next requirement is to observe the amino acids lining the tunnel. In this case the most important questions to answer are how often, how much, and which part do the amino acids influence the tunnel over time.

These input requirements, given by the domain experts, formed the initial design. Then in close cooperation with the biochemists the design was iteratively refined until its final appearance which we present in this paper.

The proposed representation consists of two main parts, the top part shows the width profiles of a tunnel over time and the bottom part reveals the individual amino acids around the tunnel (see Figure 1). To leverage the capability of the proposed method, we enable the user to explore the tunnel surroundings for individual time steps in detail. This is achieved by using a new table representation containing various statistics about a given tunnel.

The novel contributions of this work are:

- A new quantitative visualization for molecular void spaces, which is abstracted directly from the 3D molecular dynamics simulations.
- A visual analytics framework for identification of highly-ranked mutation candidates among amino acids.
- Demonstration of applicability on an example from the target domain of structural biology.

The domain experts involved in the design and evaluation of the proposed representation are coming from the research field of protein and metabolic engineering. They aim to understand the structure-function relationships of haloalkane dehalogenase enzymes and to improve their functionalities. The group involved in this project consisted of three researchers with long-term experience in protein analysis.

2 Related Work

In this section we look at the related work from several points of view. Firstly we will present related work in the domain of protein void analysis. While the literature is saturated with inner void computation we will only provide a superficial overview since this computation is not the primary focus of our work. Although protein tunnels are our main topic of interest, most presented techniques can be also applied to other types of protein voids, such as cavities, channels, pockets, and pores. Secondly we mention the analogy between our work and other application domains of scientific visualization.

2.1 Tunnel Computation and Analysis

Tunnels form the core input structures of the newly proposed representation, so first we focus on the existing approaches to their detection. The early proposed solutions for the detection of tunnels were using a grid-based approach [20]. The atomic data of the protein were used to fill-in a uniform grid. Then, using the Dijkstra algorithm, the widest void path leading from the active site to the outer space was detected. Like many other grid-based approaches, the precision and the performance of this technique was highly dependent on the grid resolution. This approach was later improved by subdividing the protein space using the Voronoi diagram [16, 19, 21, 14]. Rather than using uniformly shaped subdivision cells, the shape of each cell was defined according to the position of individual atoms. Consequently, the accuracy in terms of the geometry of the computed tunnels was greatly improved.

Since molecular dynamics plays an important role in the protein reactivity, efforts have also been made in order to compute cavities in real-time and thereby save computation time [11, 17].

When studying protein inner voids, not only the 3D geometry is relevant. Providing the tools for an interactive exploration and analysis is equally important for biochemists. A set of software solutions has also been developed in order to enable the visual analysis of protein inner voids [10, 11, 13].

Additionally, several approaches have been developed to identify and track a single tunnel throughout the molecular dynamics sequence. In order to track the path of a ligand through the inner void, Lindow et al. [13] analyze the evolution of a single cavity for all time steps and aggregate the results. Alternatively, CAVER [3] and MOLE [21] are both capable of clustering the tunnels and are able to compute the

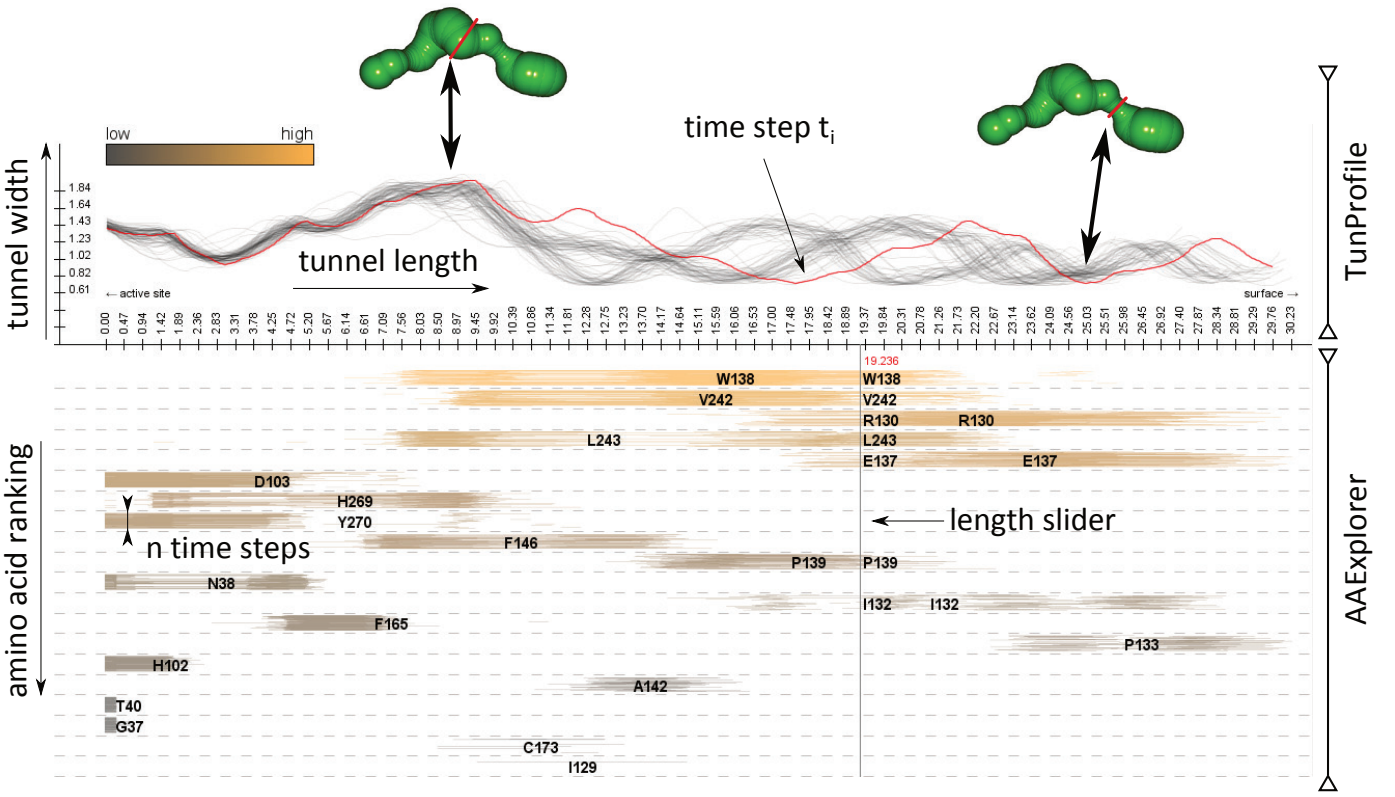

Fig. 3. The AnimoAminoMiner representation consists of two main parts. The top part, TunProfile, shows the tunnel width and length and their changes over time. The bottom part, AAExplorer, enables to explore the properties of all amino acids surrounding the tunnel.

evolution of individual tunnels for the entire molecular dynamics simulation. However, the existing visual representation of such an evolution is focusing on a 3D view which suffers from the visual clutter caused by large overlaps and the loss of understandability when analyzing large simulations.

2.2 Reformations and Time-Series Data

To reduce visual clutter due to occlusion, we unfold the 3D shape of the tunnel into the 2D domain. Unfolding has been the basis for several medical visualization applications. One relevant example is Curved Planar Reformation [7] where tubular structures such as blood vessels are mapped onto a 2D plane in order to deal with occlusion issues. A similar approach is also used for virtual colon unfolding [23, 6]. Here the situation is more complex than in our case as the authors aim to visualize the unfolded surface. While popular in the medical domain, 2D flattening of 3D data has also been used in the domain of seismic data visualization [12]. In this work the authors presented a method for reforming a curved 3D volume along a linear axis in order to facilitate visual comparison between two different datasets. Their streamline-centric visualization is based on a similar principle as our visualization of tunnel profiles. In our work we straighten the tunnel void in molecular structures for many time steps. Furthermore we abstract its spatial characteristics into schematic representations that quantitatively convey properties of the structure. Another example of visual abstraction was presented by Bidmon et al. [1]. The authors analyzed solvent pathways and their behavior in molecular dynamics. Similar pathways were then clustered to reveal the principal paths.

An important characteristic of a tunnel is the set of *lining* amino acids that surround and delineate the tunnel. It is possible to display the information about the lining amino acids for a given time step with current existing solutions. To the best of our knowledge, there is no technique available which can comprehensively show all lining amino acids throughout the whole dynamics and even explore their properties. This is due to the fact that a simple extension of the existing solutions for the dynamics is impracticable due to the changes of the amino acids lining the tunnel over time. This work is an attempt to fill this gap and to provide at a glance an overview of the amino acids' conformation for an entire tunnel and throughout the whole dynamics.

One of the most recent attempts to create an abstracted representation of tunnels in proteins [2] used two types of representations to explore the tunnels in molecular dynamics. Heat maps were used for conveying the information about tunnel width over time. The tunnel width is represented by a distinct color of the corresponding rectangle in the heat map. However, according to perception studies [24], interpreting and comparing length is more effective than reading out the difference between varying colors. Thus our newly proposed approach uses a curve for depicting the relationship between the width and length of the tunnel. It uses an aggregated representation showing the curve of the tunnel in each time step. This conveys also information about the tunnel stability with respect to the fluctuation of tunnel parts. Another abstraction introduced by Byska et al. [2] focuses on a contour representation of the most critical, i.e., narrowest, part of a tunnel, namely its bottleneck. For each time step, the corresponding bottleneck contour is plotted along with its surrounding amino acids. This representation focuses on the exploration of one bottleneck and various physico-chemical properties of its amino acids over time. But more importantly, in cases when the tunnel contains several bottlenecks and the task is to select the most critical one, this representation is not suitable as it enables to explore the bottlenecks only separately. Here our new method helps to identify and compare several tunnel bottlenecks much better than using colored heat maps.

Despite the fact the abscissa axis in our curved tunnel representation does not correspond to time but to tunnel length, one can see a certain analogy here with time-series graphs. Due to the importance of such type of data, numerous tools and techniques have been developed to facilitate interactive exploration [5]. We took inspiration from these techniques in order to improve filtering and selection of the most interesting time steps from the whole data set.

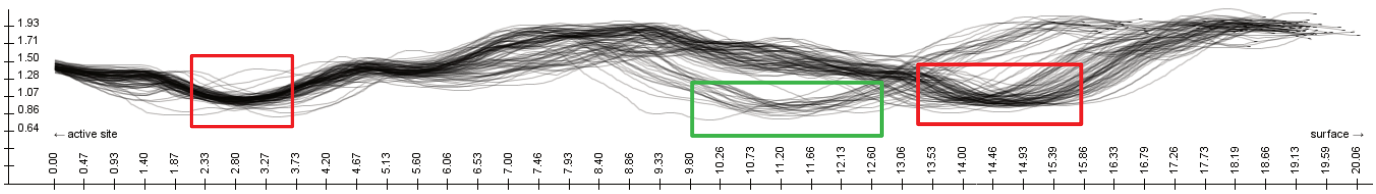

Fig. 4. The TunProfile part of our proposed representation. The curves depict tunnel profiles for individual time steps.

3 AnimoAminoMiner Overview

Interactive exploration of amino acids surrounding the tunnel in molecular dynamics faces two main challenges. The first challenge is related to the data complexity. Long simulations of molecular dynamics, which have to be explored, can contain hundreds of thousands of time steps and even more, taking into account the current hardware possibilities. Each tunnel present in this data is surrounded by dozens of amino acids, depending on the tunnel length. In consequence, we obtain a complex space with many time steps, tunnels, and surrounding amino acids, where biochemists are searching for one or few amino acids with specific properties. Secondly, the data possess a very complex spatial arrangement which is hard to perceive and understand. This is mainly due to the fact that tunnels are usually highly curved structures with non-obvious relationships to surrounding amino acids.

The AnimoAminoMiner focuses on these challenges. Moreover, it helps the biochemists in the process of detecting the best candidate amino acids for mutations which aim to change the protein function. A thorough study of the amino acids surrounding the tunnel, and their properties and positions with respect to the tunnel, leads to the proposition of such candidates.

Our visualization design is an abstraction of the complex multi-dimensional scenario into simpler two-dimensional schematic representations of the tunnel straightened by unfolding. We depict a tunnel's width and length in one time step in a function plot. This allows us to plot all time steps in an aggregated view to convey the tunnel's overall shape and stability over time. Moreover, the two-dimensional layout gives the possibility to plot the areas of influence for each amino acid which is lining the tunnel. The areas of influence of amino acids are associated with the horizontal direction along a tunnel's centerline. Where exactly the amino acid is located around the centerline is abstracted away. The visual estate is used to display the rank of amino acids according to user specified descriptors. By this the analyst can have an unobstructed view on which amino acid affects which part of the tunnel and what are its specific chemical and physical properties.

The AnimoAminoMiner consists of two main views (see Figure 3). The top view, denoted as TunProfile, focuses on the visual representation of tunnel widths and their changes over time. The bottom view of our method, named AAExplorer, provides the users with the representation of all amino acids surrounding the tunnel. In the following sections we describe the two main views of the AnimoAminoMiner in detail.

4 AnimoAminoMiner – TunProfile

The first part of our new method focuses on the representation of tunnel width and length over time (see Figure 4). The profile of the tunnel is represented by a function graph which describes the tunnel width along the tunnel centerline. For the tunnel computation we use the CAVER 3.0 algorithm [3] which detects all tunnels satisfying the input parameters for each time step and then performs a clustering in order to find the correspondence between tunnels over time. The result of such a computation is a set of tunnels containing the information about their temporal changes.

The AnimoAminoMiner enables to explore these tunnels individually and focuses on communicating the changes of a tunnel and its surroundings over time.

In the TunProfile part, a tunnel is represented by a smoothed curve in a 2D graph where the abscissa axis shows the tunnel length and the ordinate axis represents the tunnel width. In each time step, the tunnel shape is computed and its corresponding curve is plotted, i.e., if there are n time steps, TunProfile will include n curves. All curves start at the active site which is located on the left side of the 2D graph. Because the length of the tunnel over time can differ significantly, the end points of the curves are scattered on the abscissa axis. The width of the 2D graph then corresponds to the longest detected curve. In cases where the tunnel length changes significantly over time (e.g., see Figure 11), this representation exhibits a progressive misalignment of individual time steps from left to right and is not sufficient for the exploration of changes of the tunnel shape. For such situations we enable to change the TunProfile appearance to a relative distance visualization. Here the abscissa represents the relative length of the tunnel in percentage. This leads to a representation where the tunnel in each time step is scaled to the maximal tunnel length during the entire simulation period. This change better highlights the tunnel shape and its endpoint over time. In both cases, the resulting aggregated view reveals the most stable and unstable parts of the tunnel. The stable parts are represented by a dense packing of curves.

With the TunProfile representation, the user can easily detect tunnel bottlenecks and their positions along the tunnel centerline. Figure 4 shows an example of a tunnel with two bottlenecks (marked by red rectangles). The first bottleneck is very stable, i.e., its width is changing minimally over time. It is located close to the protein active site, i.e., closer to the left side of the graph. The second bottleneck is less stable and is located much closer to the protein surface. Moreover, its position varies over time and the graph view is clearly able to show the shifting in position towards the active site for a certain period of time (green rectangle).

Individual curves can be colored according to time within the molecular dynamics simulation. Curves representing the tunnel in earlier phases of the simulation have different colors than curves depicting the tunnel shape at the end of the simulation. The coloring helps to reveal interesting temporal information. For example, Figure 1 shows the same tunnel profile as Figure 4 but the individual curves are colored. For the less stable bottleneck, the coloring indicates that in earlier phases of the simulation the bottleneck is closer to the active site. In later time steps, the bottleneck shifts its position closer to the surface. This information is very useful when biochemists are studying a given tunnel with respect to its suitability for a ligand transportation.

5 AnimoAminoMiner – AAExplorer

The second part of the proposed method, called AAExplorer, enables the exploration of the lining amino acids of the tunnel in detail throughout the whole dynamics.

Figure 2 shows a representation of amino acids surrounding a tunnel at one specific time step, which suffers from visual clutter caused by overlaps. Moreover, the amino acids are captured only for one time step. In order to see their behavior in the dynamics, one option is to animate their movements. As already mentioned, this is impractical for very long molecular dynamics simulations. AAExplorer overcomes this limitation by using a highly abstracted 2D representation for lining amino acids rather than a 3D view. It provides an aggregated view in 2D which conveys the desired characteristics. Moreover, it leaves out many unnecessary spatio-temporal details present in a 3D view at the benefit of occlusion avoidance and clarity. The following description of AAExplorer is divided into four subsections. These subsections

discuss the sequence of steps in our approach. The last subsection explains the details of the table view which the AAExplorer offers for individual amino acids.

5.1 Amino Acid Detection

The first task is to calculate a set of amino acids which are surrounding the tunnel. These amino acids are often denoted as the lining amino acids. There are many different methods for the computation of the lining amino acids. Almost every solution for tunnel computation uses its own method for the detection of these amino acids. For example, MOLE 2.0 [21] samples the tunnel centerline where the distance between two samples is set at 0.1 Ångström. In each sample, the five closest atoms are detected, and their amino acids are marked as the lining ones. In our case we use the approach of CAVER 3.0 [3]. Here the amino acids are defined as tunnel lining if their distance to the tunnel centerline is smaller than a user defined value (by default experimentally set to 3 Ångströms).

It is possible to use any other algorithm for the detection of tunnel-lining amino acids. Currently none of the existing algorithms for this problem can be taken as a standard giving the most relevant results. Our method can be easily adjusted to any of the existing solutions. It requires as an input only the list of spheres forming the tunnel body and the list of surrounding amino acids with additional information, such as their spatial orientation, type, or physico-chemical properties. Tunnel spheres are spheres of maximal radii with respect to the surrounding atoms and are positioned on the tunnel centerline. The union of the tunnel spheres forms the rough shape of the tunnel (see Figure 2).

5.2 Amino Acid Mapping

Our proposed abstracted view represents the tunnel 3D geometry by unfolding the tunnel in 2D along its centerline. As the information about the spatial influence of the lining amino acids on the tunnel has to be preserved in the 2D projection, we apply the following special mapping. First of all, for each amino acid and for each single time step, the influence is computed. We employ a tunnel representation which consists of a set of intersecting spheres positioned on the tunnel centerline (see Figure 5).

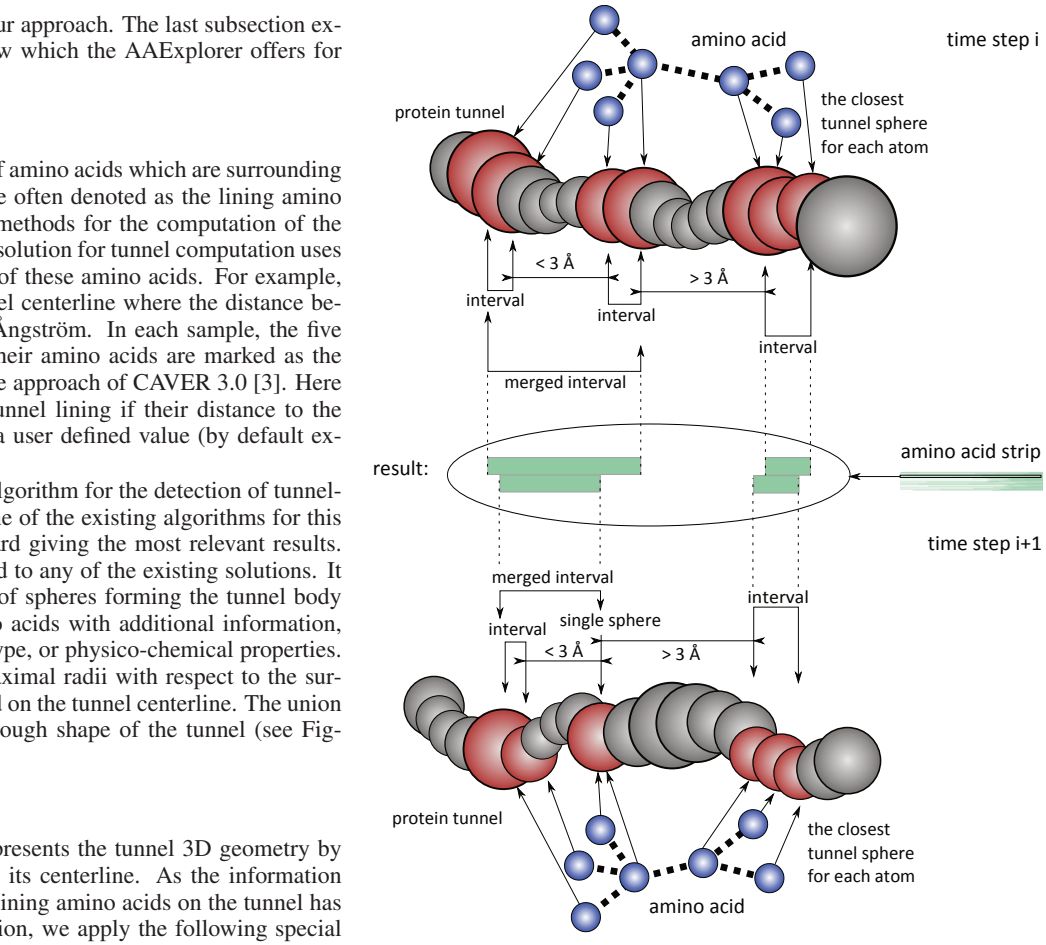

Fig. 5. The part of a tunnel that is affected by an amino acid is identified using atom-tunnel distances.

Then for each atom of the lining amino acid we find the nearest tunnel sphere. This is achieved by computing the distance between the atom center to the closest point on the surface of the tunnel sphere. As a result, we obtain discrete intervals of tunnel spheres (red spheres in Figure 5) that are affected by each amino acid.

If these individual intervals do not cover the affected area completely, we employ the following post processing procedure. For neighboring intervals we compute the distance between centers of tunnel spheres along the centerline and merge those intervals that are closer to each other than 3 Ångströms (see Figure 5). This threshold value corresponds to the experimentally specified value which is used by the algorithm for the computation of surrounding amino acids in CAVER 3.0.

If the distance between intervals is bigger than 3 Ångströms, it corresponds to the situation where the amino acid influences distinct, more distant parts of the tunnel. In AAExplorer, this results in the discontinuous representation of one amino acid. Figure 5 shows the mapping in two different time steps. The size of the detected intervals is mapped onto the length of lines for the corresponding amino acid. Each time step contributes with one, possibly disconnected line to the final amino acid strip.

5.3 Amino Acid Timeline

Once the influence region is computed for all amino acids and in all snapshots the results are displayed below the TunProfile view. A single line corresponds to the influence region of one amino acid for a single time step. For each amino acid the influence line is displayed for all the snapshots in chronological order, from top to bottom, as shown in Figure 6. As the line density increases with the number of snapshots, one can immediately see the importance of an amino acid over the corresponding portion of the tunnel.

The color of the lines corresponds to physico-chemical properties of amino acids, such as hydrophobicity or partial charge. These properties do not change over time so the color of lines representing one amino acid will always remain the same. To prevent visual clutter we also offer a more aggregated view of the temporal contributions of each amino acid in a single strip (see Figure 7). This view is helpful in three main cases. First, it helps to differentiate between amino acids with close spatial position and very similar or the same color (see Figure 8). Second, it provides the user with an overview of the overall contribution of a given amino acid. This is useful mainly if the influence regions of the amino acid are too much scattered. The third case is relevant for bigger numbers of time steps. If processing more than approximately 120 time steps, the width of each line in the

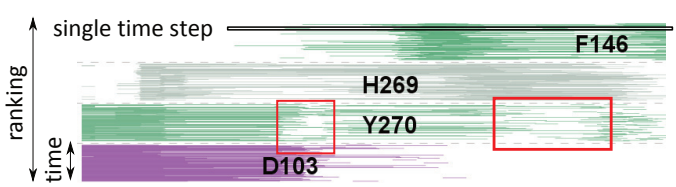

Fig. 6. AAExplorer showing an example of four amino acids lining a tunnel. One line of each amino acid corresponds to one time step. If the amino acid influences several distant parts of the tunnel, its representation is discontinuous (see red rectangles for the Y270 amino acid).

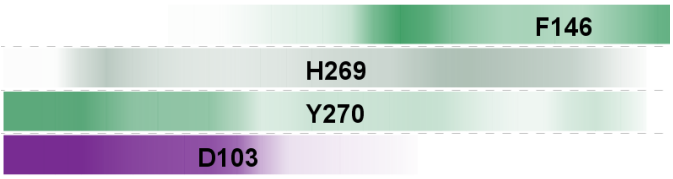

Fig. 7. Line blending of spatial and temporal contributions of amino acids which removes the visual clutter.

amino acid strip starts to cover less than one pixel. This leads to a lossy representation and there are two basic strategies solving this problem. We can interpolate the length between several neighboring time steps and create a joint representative, or we can select each n-th time step and skip the rest. The second strategy may remove some significant values which can be highly undesirable. In both cases we are losing some information. So by using our aggregated view based on blending, another loss of information is negligible and the result is more comprehensible.

5.4 Amino Acid Ranking

The importance of an amino acid for the tunnel shape may vary depending on the property of the amino acid. Therefore the vertical order of the amino acids in the AAExplorer can be adjusted with respect to user-defined criteria. Biochemists can change the ordering by giving each amino acid a score which influences its vertical positioning.

A higher score means that the amino acid will be depicted at a higher position in the AAExplorer. The score of the amino acids is computed from:

- Hydrophobicity – more hydrophobic amino acids have a higher score.
- Partial charge – more positively charged amino acids get a higher score.
- Donors and acceptors – hydrogen acceptors get a higher score than hydrogen donors (both playing important roles in hydrogen bonding).
- Extent of influence – this value combines the length of the tunnel centerline which is lined by a given amino acid (length of the amino acid strip in the AAExplorer) and its stability over time (number of lines for each amino acid). The amino acids having a larger influence on the tunnel, both in space and time, get a higher score.

These criteria were carefully selected in cooperation with the biochemists. They correspond to the most important properties of amino acids, which are commonly studied when searching for most suitable mutations. All these criteria can be expressed by numbers from a predefined interval. Thus, the user can define a threshold for each criterion and filter out amino acids below this value. Furthermore, the filtering can be applied according to several criteria simultaneously, e.g., the user can select amino acids which are hydrophilic and have a high extent of influence.

The vertical axis of the AAExplorer consists of several sections divided by dashed lines (see Figure 8). Each section contains amino acids with the same score computed according to the given criteria. This is very unlikely to happen for the extent of influence criterion but, e.g., sorting according to donors and acceptors, the amino acids may get the same score.

The vertical axis has two meanings (see Figure 6). The global vertical axis corresponds to the ranking of amino acids with respect to user-defined criteria. Additionally, each strip defining one amino acid has its own local vertical axis which shows individual time steps of the dynamics simulation.

5.5 AAExplorer – Table View

The AAExplorer enables the user to explore the lining amino acids on different levels of detail. The user can utilize horizontal and vertical zooming in order to explore the spatial and temporal contributions of lining amino acids in more detail. Each amino acid can be selected individually so that the AAExplorer shows only the selected one. Moreover, further information on such an amino acid can be shown in a detailed table view which contains specific statistical information (see Figure 9).

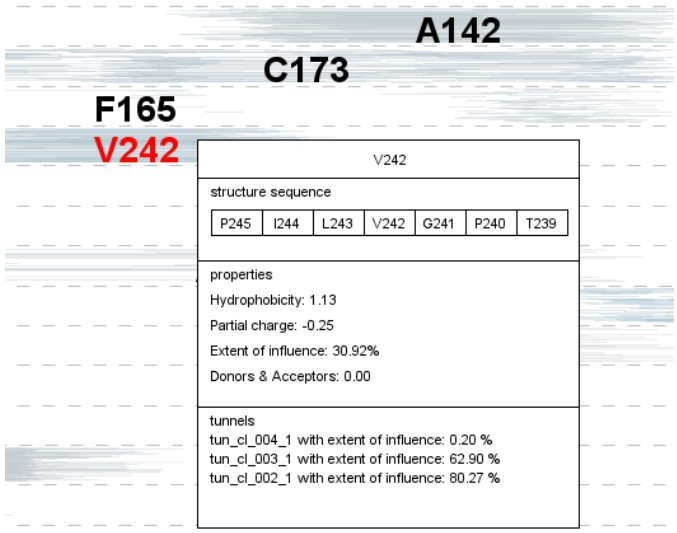

Fig. 9. Table view containing detailed information about a single amino acid.

The goal of the table view is to provide the users with the information about individual amino acids which is missing in the aggregated AAExplorer. It appears when the user selects one amino acid by clicking on its strip in the AAExplorer. The table then contains the information about the amino acids neighboring with the selected amino acid in the protein chain. Then the detailed values of several properties of a given amino acid are given, such as its hydrophobicity, partial charge, extent of influence, and if the amino acid is a hydrogen donor or acceptor. Most importantly, the table informs about the contribution of the amino acid to other tunnels present in the protein. This information is crucial to check for side effects in cases where the biochemist aims to mutate this amino acid in order to change the width of the scrutinized tunnel. In consequence, such a mutation can influence also other tunnels in the protein which can be undesirable. To avoid such effects to be overseen, we included this information in the detailed table view.

6 User Interaction

The strength of our proposed AnimoAminoMiner highly depends on the interactive exploration of both its parts, i.e., the TunProfile and the AAExplorer where a tight linking and brushing is realized. The TunProfile initially shows the profiles of the tunnel for all time steps. The profiles and the tunnel-lining amino acids can be explored using a Length slider. It is positioned vertically and the user can do interactive manipulations. While moving the slider, the names of all affected amino acids are interactively positioned next to it. This gives the information about all amino acids influencing a given position on the tunnel.

Furthermore, the biochemists may want to focus on a smaller time span, e.g., when the scrutinized bottleneck is not present in the whole simulation. For such cases, TunProfile enables brushing where the user can specify a rectangle to select a desired set of tunnel curves, i.e., time steps of interest. The remaining curves are automatically filtered out and the graph shows only the selected time steps (see Figure 10). Through linking, the AAExplorer is adapted to the filtering as well.

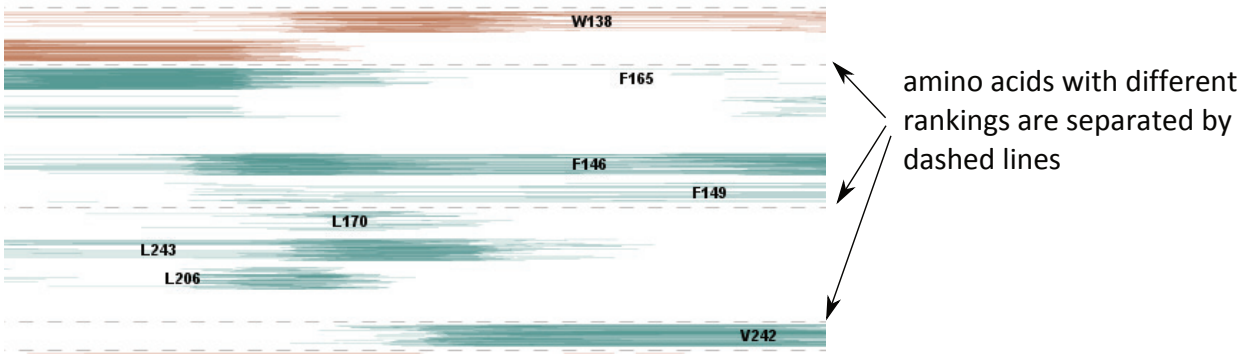

Fig. 8. Amino acids with the same ranking score fall into one section.

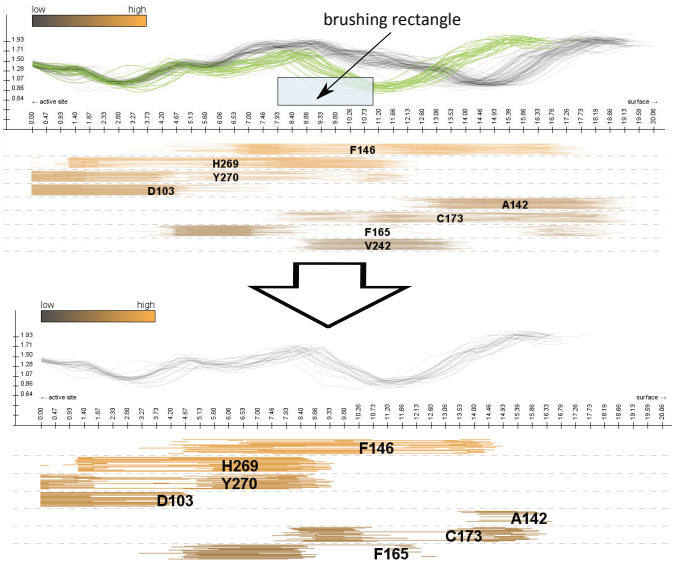

Fig. 10. Rectangular brushing enables to select a subset of tunnel curves which corresponds to a subset of time steps.

The AAExplorer further enables to sort the amino acids with respect to selected criteria, i.e., hydrophobicity or partial charge. These criteria can be easily and intuitively combined. The user can get a set of amino acids fulfilling only these criteria and their specific values.

7 Demonstration and Results

To evaluate the usability of the newly proposed representations, several case studies led by the domain experts were conducted. Their aim was to evaluate the correctness of our approach. This concerns the capability of our method to reveal mutation candidates which were already previously determined using a combination of complex analyses and lab experiments. These case studies are derived from the already published results of our cooperating biochemical group. All of them focus on the design of mutations changing specific protein properties. The candidate amino acids for mutations are always located around a tunnel. Thus there was the requirement that these amino acids and their importance should be clearly detected by the AnimoAminoMiner. To confirm this assumption, the following studies were performed.

7.1 Case Study 1

First we present a recently published case study [15] which describes the process of balancing the trade-off between activity and stability of DhaA haloalkane dehalogenase. The aim of this study was to create a stabilized protein with superior catalytic activity by proposing mutations around the main tunnel of the protein. The experiments started with the DhaA80 mutant which was very stable but its activity was very low. By complex analyses it was revealed that the most important tunnel of this mutant contains a bottleneck which is located close to the protein surface and prevents the respective ligand to enter the protein. By another set of explorations and experiments, the biochemists revealed and proposed candidates for mutation which would remove this bottleneck and thus increase the protein activity. After series of lab experiments it was determined that the most appropriate mutation replaces the phenylalanine amino acid F176 by glycine. The newly created mutant, marked as DhaA106, eliminated the bottleneck and the activity was increased. The study confirmed that a delicate balance between activity and stability of enzymes can be achieved by fine-tuning the diameter and the dynamics of the main tunnel as well of the other tunnels in a protein.

This study was taken as ground truth and the AnimoAminoMiner was tested with respect to its usability on these data. First the biochemists analyzed the initial DhaA80 protein mutant and tested if the AnimoAminoMiner helps to reveal the best candidates for mutation without a deep knowledge about the protein. The testing was performed repeatedly on several time steps taken in regular intervals from the simulation of the DhaA80 molecular dynamics comprising 200.000 time steps in total. These intervals were decreased and the last test was performed on all time steps. Figure 11 shows an example with 100 time steps.

First of all, the biochemists highly appreciated the TunProfile representation which gives them an immediate overview of the tunnel width and length over time. Figure 11 clearly shows that even if the length of the tunnel is changing substantially, the tunnel bottleneck is always located in the vicinity of the protein surface and almost closes the tunnel. The next step is to propose one or several candidates for mutation which will remove this bottleneck. Here the AAExplorer plays a crucial role as it depicts the spatial and temporal influence of lining amino acids on the tunnel. By sorting the amino acids according to user-defined criteria, i.e., hydrophobicity, partial charge, extent of influence, and donors and acceptors, the mutation candidates can be detected easily. When testing the AAExplorer with the DhaA80 molecular dynamics simulation, the biochemists firstly sorted the amino acids with respect to their extent of influence. Figure 11 shows that the top ranked amino acid is F176 (phenylalanine) which corresponds exactly to the amino acid mutated in the case study. The top position of F176 signifies that this amino acid had the highest ranking with respect to its stability over time and its influence on the tunnel. So the AAExplorer is able to reveal the best candidate immediately.

The procedure of searching for the best candidate often ends with a set of possible amino acids which are further explored. In such a case the biochemists will use the extent of influence sorting in the AAExplorer to define several top ranked amino acids. These amino acids can be further explored by other ranking possibilities which focus on physico-chemical properties. The information about the properties of candidate amino acids is crucial. It helps the biochemist to select even

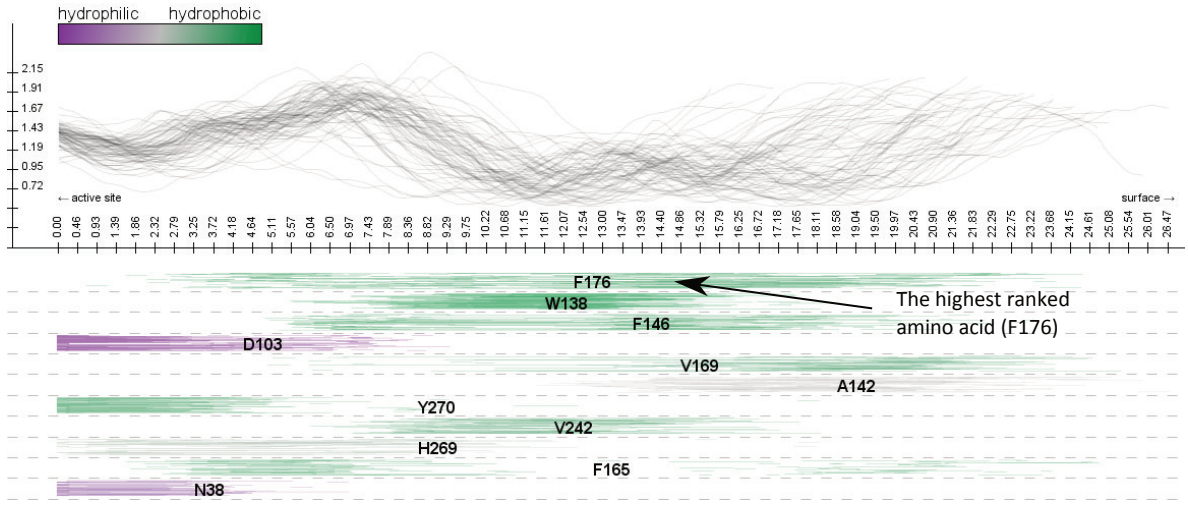

Fig. 11. AnimoAminoMiner applied to the DhaA80 protein shows the position of the tunnel bottleneck along its centerline. The ranking immediately reveals the best amino acid candidate for mutation. It is the highest ranked F176 amino acid which influences a substantial part of the tunnel and which has the largest impact on the width of the bottleneck.

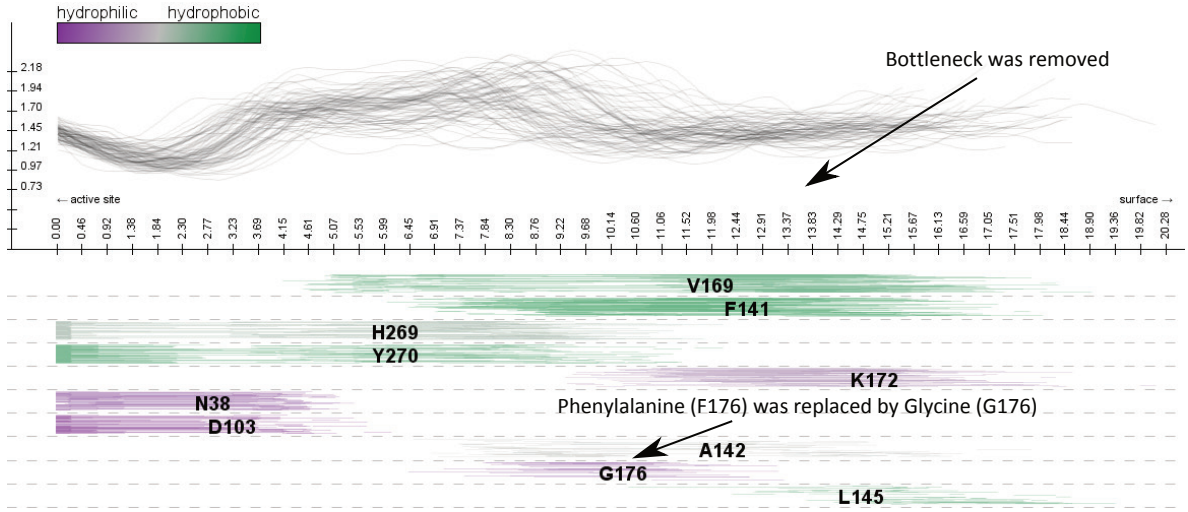

Fig. 12. AnimoAminoMiner applied to the DhaA106 mutant reveals that by mutating the F176 position the bottleneck has been removed and the extent of influence of the replacement G176 amino acid has been decreased.

more precisely the new amino acid which will replace the mutation candidate.

The conclusion of the first part of the testing was that the AnimoAminoMiner helps to reveal the most critical parts of a tunnel as well as the candidate amino acids for mutation. Using our approach, the set of selected candidates can be much smaller as compared to the biochemists following their traditional process of exploration. This process involves many phases, such as sequential alignment, computation of energy, or exploration of tunnels in all time steps. If these phases have to be applied on many possible candidates, the whole process of finding the most appropriate mutation takes months. AnimoAminoMiner helps to reduce the set of possible candidates by easily detecting those amino acids which do not have a substantial effect on a given tunnel over time. In consequence, also the time for finding the correct mutation is shortened.

The second part of the testing focused on the exploration of the DhaA106 mutant where phenylalanine at position 176 is replaced by glycine. Figure 12 depicts the mutated access tunnel.

The TunProfile shows that the tunnel has been slightly shortened. More importantly, the bottleneck present in DhaA80 was completely removed and now the part of the tunnel close to the protein surface is more stable over time. Also the AAExplorer indicates that the extent of influence of the new G176 amino acid decreased in comparison with the original F176 amino acid. Furthermore G176 has been shifted to a less important position.

7.2 Case Study 2

The second study follows the design of protein mutations described by Pavlova et al. [18] which also led to the increase of the enzyme activity. In this case redesigning the access tunnels of the haloalkane dehalogenase DhaA provided a 32-fold increase in activity with toxic pollutant 1,2,3-trichloropropane (TCP). The structure contains two important access tunnels, the main and the side tunnel. In contrast to the previous case study, the idea behind this proposed mutation was to close the main tunnel of the DhaA. This tunnel was too wide, so not only TCP was able to pass through to the active site but also many water molecules used the tunnel to penetrate the protein. The presence of a large amount of water in the interior prevented the reaction of the protein with TCP. The increase of DhaA activity towards TCP was achieved by closing the main tunnel and leaving only the side tunnel open. The TCP was still able to follow the side tunnel to the active site. The water molecules reached the active site via the side tunnel

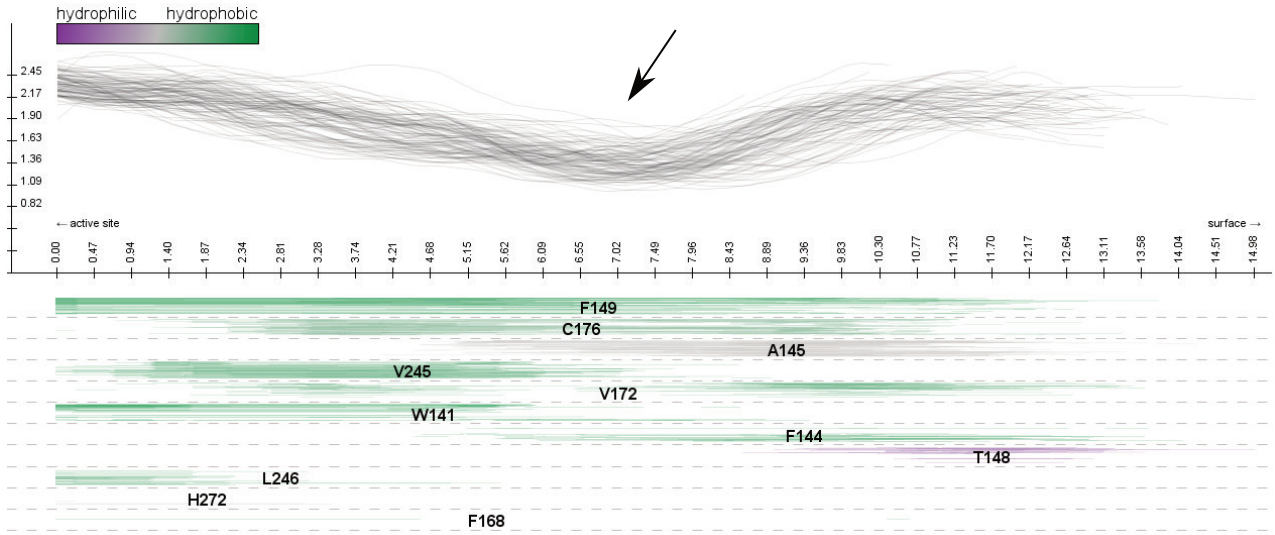

Fig. 13. AnimoAminoMiner applied to haloalkane dehalogenase DhaA reveals that the potential candidates for mutation are the F149 and the C176 amino acids. After further exploration the C176 was selected as the best candidate.

as well but in this case their number decreased dramatically. So the water barrier between the active site and the TCP was removed and the reaction proceeded much easier.

Studying this situation with the AnimoAminoMiner, we search for the amino acids in the surrounding of the main tunnel which significantly influence its width. These can be replaced by a larger amino acid which will close the main tunnel. The TunProfile helps to reveal the tunnel bottleneck close to the central part (see arrow in Figure 13) where the proposed mutations can have an immediate desired impact.

The amino acids forming the bottleneck are explored in detail using the AAExplorer (see Figure 13). The two highest ranked amino acids, F149 and C176, influence the bottleneck substantially. Both amino acids are potential mutation candidates. As the aim of the mutation is to close the tunnel, the selected candidate has to be small enough to be replaced by a larger amino acid. In this case the first amino acid F149 (phenylalanine) is one of the largest amino acids so it cannot be easily replaced by even a larger one. On the other hand, the C176 (cysteine) amino acid is small enough so the biochemists can easily find a larger amino acid for its replacement. After consideration of other properties, the biochemists selected tyrosine as the most suitable amino acid. The replacement of cysteine causes the closing of the main tunnel.

7.3 Evaluation Summary

An integral part of the evaluation by the domain experts is related to the correspondence between the initial requirements and the final result. The verification of the correctness of the results provided by the AnimoAminoMiner was performed by the above described studies on real cases. The biochemists also considered if the final design followed their initial demands. The first requirement was to study an entire tunnel for long-time simulations at once. This was defined as the main objective of the AnimoAminoMiner. After conducting the case studies the biochemists confirmed that the necessary information is conveyed in a very comprehensible manner. They highly appreciated the possibility to observe changes in the tunnel shape using the TunProfile. The AAExplorer then enables to reveal a set of potentially interesting tunnel-lining amino acids within seconds. The biochemists confirmed that the initial requirements were fulfilled.

In summary, the biochemists concluded that the novel AnimoAminoMiner gives valuable information about the tunnel surroundings and its evolution over time. Such information is impossible to get easily with any of the currently available methods. They stated that our method intuitively and comprehensibly communicates the highly complex protein environment and its properties. Using AnimoAminoMiner in biochemical research will hopefully lead to substantially faster and even more focused propositions of mutation candidates, which will change the protein function with respect to particular needs.

8 Conclusion

In this paper we proposed the novel AnimoAminoMiner representation. It provides biochemists with an abstracted overview on the main characteristics of a protein tunnel, namely its width, length, surrounding amino acids and their properties. This also includes changes over time represented by molecular dynamics simulations. Among other numerous characteristics of protein tunnels, these were selected as the most crucial ones when searching for the best mutation candidates influencing protein reactivity and its function. The AnimoAminoMiner focuses also on the interactive exploration of the proposed representations which leads to a fast and intuitive selection of the most appropriate candidate amino acids. This was also confirmed by several case studies and two of them have been presented in detail.

The results of the testing phase performed by biochemists lead to two interesting conclusions. First, it was confirmed that focusing on an abstracted representation of a complex protein environment containing tunnels due to molecular dynamics is a meaningful step forward in molecular visualization.

Second, we expected that the AnimoAminoMiner will give relevant information to biochemists and will serve as a tile in the mosaic of steps which they are following in the process of searching for the best mutation candidates. The results of the testing phase exceeded our expectations as they revealed that the AnimoAminoMiner ranking system very well detects promising candidates. This indicates that the new representation can be a powerful tool taking an important role in biochemical research.

The AnimoAminoMiner currently focuses on studying a single tunnel over time. Proteins contain usually several potentially relevant tunnels. So there is a demand to study them together. The most interesting combinations of tunnels would be those sharing some lining amino acids. Then the AAExplorer could contain information about all amino acids lining the desired tunnels with new ranking possibilities. This opens an interesting direction for the future extension of our method.

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

[1] K. Bidmon, S. Grottel, F. Bös, J. Pleiss, and T. Ertl. Visual abstractions of solvent pathlines near protein cavities. *Computer Graphics Forum*, 27(3):935–942, 2008.

[2] J. Byska, A. Jurcik, M. E. Gröller, I. Viola, and B. Kozlikova. MoleCollar and Tunnel Heat Map Visualizations for Conveying Spatio-Temporo-Chemical Properties Across and Along Protein Voids. *Computer Graphics Forum*, 34(3):1–10, 2015.

[3] E. Chovancova, A. Pavelka, P. Benes, O. Strnad, J. Brezovsky, B. Kozlikova, A. Gora, V. Sustr, M. Klvana, P. Medek, L. Biedermannova, J. Sochor, and J. Damborsky. CAVER 3.0: A tool for the analysis of transport pathways in dynamic protein structures. *PLoS Computational Biology*, 8(10), 2012.

[4] U. Hensen, T. Meyer, J. Haas, R. Rex, G. Vriend, and H. Grubmueller. Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. *PLoS ONE*, 7(5):e33931, 2012.

[5] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets, timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, 2004.

[6] W. Hong, X. Gu, F. Qiu, M. Jin, and A. Kaufman. Conformal virtual colon flattening. In *Proceedings of the 2006 ACM Symposium on Solid and Physical Modeling*, SPM '06, pages 85–93, New York, NY, USA, 2006. ACM.

[7] A. Kanitsar, D. Fleischmann, R. Wegenkittl, P. Felkel, and M. Gröller. CPR - curved planar reformation. In *Visualization, 2002. VIS 2002. IEEE*, pages 37–44, Nov 2002.

[8] M. Klvana, M. Pavlova, T. Koudelakova, R. Chaloupkova, P. Dvorak, Z. Prokop, A. Stsiapanava, M. Kuty, I. Kuta-Smatanova, J. Dohnalek, P. Kulhanek, R. C. Wade, and J. Damborsky. 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):1339–1356, 2009.

[9] T. Koudelakova, R. Chaloupkova, J. Brezovsky, Z. Prokop, E. Sebestova, M. Hesseler, M. Khabiri, M. Plevaka, D. Kulik, I. Kuta Smatanova, P. Rezacova, R. Ettrich, U. T. Bornscheuer, and J. Damborsky. Engineering enzyme stability and resistance to an organic cosolvent by modification of residues in the access tunnel. *Angewandte Chemie International Edition*, 52(7), 2013.

[10] B. Kozlikova, E. Sebestova, V. Sustr, J. Brezovsky, O. Strnad, L. Daniel, D. Bednar, A. Pavelka, M. Manak, M. Bezdeka, P. Benes, M. Kotry, A. W. Gora, J. Damborsky, and J. Sochor. CAVER Analyst 1.0: Graphic tool for interactive visualization and analysis of tunnels and channels in protein structures. *Bioinformatics*, 30(18):2684–5, 2014.

[11] M. Krone, M. Falk, S. Rehm, J. Pleiss, and T. Ertl. Interactive exploration of protein cavities. *Computer Graphics Forum*, 30(3):673–682, 2011.

[12] O. D. Lampe, C. Correa, K.-L. Ma, and H. Hauser. Curve-centric volume reformation for comparative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1235–1242, 2009.

[13] N. Lindow, D. Baum, A.-N. Bondar, and H.-C. Hege. Exploring cavity dynamics in biomolecular systems. *BMC Bioinformatics*, 14(S-19):S5, 2013.

[14] N. Lindow, D. Baum, and H.-C. Hege. Voronoi-based extraction and visualization of molecular paths. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2025–2034, Dec 2011.

[15] V. Liskova, D. Bednar, T. Prudnikova, P. Rezacova, T. Koudelakova, E. Sebestova, I. K. Smatanova, J. Brezovsky, R. Chaloupkova, and J. Damborsky. Balancing the stability-activity trade-off by fine-tuning dehalogenase access tunnels. *ChemCatChem*, 7(4):648–659, 2015.

[16] P. Medek, P. Benes, and J. Sochor. Computation of tunnels in protein molecules using Delaunay triangulation. *Journal of WSCG*, 15(1-3):107–114, 2007.

[17] J. Parulek, C. Turkay, N. Reuter, and I. Viola. Visual cavity analysis in molecular simulations. *BMC Bioinformatics*, 14(Suppl 19):S4, 2013.

[18] M. Pavlova, M. Klvana, Z. Prokop, R. Chaloupkova, P. Banas, M. Otyepka, R. C. Wade, M. Tsuda, Y. Nagata, and J. Damborsky. Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nature Chemical Biology*, 5:727 – 733, 2009.

[19] M. Petrek, P. Kosinova, J. Koca, and M. Otyepka. MOLE: A Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure*, 15(11):1357–1363, 2007.

[20] M. Petrek, M. Otyepka, P. Banas, P. Kosinova, J. Koca, and J. Damborsky. CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics*, 7:316, 2006.

[21] D. Sehnal, R. Svobodova Varekova, K. Berka, L. Pravda, V. Navratilova, P. Banas, C.-M. Ionescu, M. Otyepka, and J. Koca. MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *Journal of Cheminformatics*, 5(1), 2013.

[22] J. Sykora, J. Brezovsky, T. Koudelakova, M. Lahoda, A. Fortova, T. Chernovets, R. Chaloupkova, V. Stepankova, Z. Prokop, I. K. Smatanova, M. Hof, and J. Damborsky. Dynamics and hydration explain failed functional transformation in dehalogenase design. *Nat. Chem. Biol.*, 10(6):428–430, Jun 2014.

[23] A. Vilanova Bartroli, R. Wegenkittl, A. König, and E. Gröller. Nonlinear virtual colon unfolding. In *Visualization, 2001. VIS 2001. IEEE*, pages 411–579, Oct 2001.

[24] C. Ware. *Information Visualization, Third Edition: Perception for Design*. Elsevier, 2013.

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