RESEARCH

COZOID: Contact Zone Identifier for Protein-Protein Interaction Analysis

Katarína Furmanová[1*], Jan Byška[2], Eduard M. Gröller[3], Ivan Viola[3], Jan J. Paleček[1] and Barbora Kozlíková[1]

Abstract

Background: Studying the patterns of protein-protein interactions (PPIs) is fundamental for understanding the structure and function of biological complexes. However, the exploration of the vast space of possible mutual configurations of interacting proteins and their contact zones is very time consuming and requires domain expert knowledge.

Results: In this paper, we propose a system of visual abstraction techniques for guided exploration of the PPI configuration space. The system follows the workflow of proteomics experts. The first visual abstraction method is based on customized interactive heat maps and provides an overview of all possible residue-residue contacts in PPI configurations. Models containing a particular pair of interacting amino acids can be selectively picked and traversed. The detailed information about individual contact zones and their properties is presented by the contact-zone list-view. These techniques are interactively linked with 3D representations that employ exploded views and open-book views to solve the problem of high overlaps of the configurations.

Conclusions: Using these views, the structural alignment of the best models can be visually confirmed. We verified the usefulness of our system on docking structures covering all three types of PPIs, i.e, coiled-coil, pocket-string, and surface-surface interactions. The results of the evaluation show that our tool helps the domain experts to analyze, filter, and explore large sets of protein configurations in a fraction of time spent when using the previously available techniques.

Keywords: protein-protein interaction; contact zone; visualization

Background

Understanding the constitution and biological function of proteins is essential in many research disciplines, for example in medicine and pharmaceutics. This knowledge is tightly connected with the ability of the protein to interact with other molecules.

Proteins can interact with small molecules, called ligands, which enter the protein. This process, called protein-ligand docking, is widely used in protein engineering. Here the goal is to change specific properties of a given protein by performing a chemical reaction between the protein and a ligand. In drug design, the protein serves as a "factory" for structural changes of a ligand caused again by a mutual chemical reaction. Such a modified ligand can then serve as a basis for a new drug. In drug design, the interactions between proteins are increasingly attracting attention because most of the proteins critical for cellular life act in a cooperative manner, forming multiprotein complexes. It is estimated that about 800 complexes exist in just one yeast cell [1]. Furthermore, all complexes are composed of subunits, which constitute the complex via protein-protein interactions (PPIs). The main goal of the process of studying such PPIs, known as protein-protein docking, is to identify an appropriate spatial **configuration** of interacting proteins. A configuration is represented by the mutual spatial orientation of the interacting proteins. Each configuration contains a **contact zone** consisting of the set of amino acids from both interacting proteins that are in the interaction distance spanning from 3 to 5 Ångströms. These thresholds are commonly used by the existing computational tools, but can be customized. In other words, the contact zone resides on the interface between the proteins and is formed by mutually interacting amino acids.

Structure determination of PPIs in laboratories is very challenging as well as expensive and time consuming. This is due to many problems related to the dynamic nature of the proteins, the difficulties with their purification and the preparation of samples. Therefore,

*Correspondence: furmanova@mail.muni.cz
[1]Masaryk University, Brno, Czech Republic
Full list of author information is available at the end of the article

computational docking is used to study the feasibility of proposed configurations. Many algorithms and tools have appeared in this respect in the last years. A categorization of the existing algorithms along with the description of their basic principles was published recently by Huang [2]. However, these algorithms produce a large number of possible configurations. The domain expert has to explore them and select the biochemically most relevant ones. Even if the computational tools usually provide the users with a score to rank the configurations, the resulting ordering does not correspond to their biochemical relevance. Therefore, the configurations have to be processed manually, which requires a visual support to enhance the exploration process.

Already several algorithms have been published for re-ranking of the configurations according to different criteria. They propose the user those configurations, which should be explored in detail. As a representative of these attempts, Malhotra et al. [3] presented in 2015 DockScore, a webserver for ranking the individual configurations produced by the docking tools. Their idea is based on building a scoring scheme considering several interface parameters, such as surface area, hydrophobicity, spatial clustering, etc. This helps the user to reduce the number of configurations to a smaller set, which still has to be explored manually. For this exploration, a visual support is essential, as it enables us to see the spatial orientation of the contact zones and to compare different configurations.

It is obvious that even for the comparison of two configurations the traditional overlay representation suffers from many occlusion problems and it is hard to perceive the differences between individual solutions. When comparing more configurations, even without a detailed visualization of hot spot amino acids, the problem becomes even more apparent (see Figure 1).

In this paper we propose a solution to this clutter problem. We present a systemic tool comprised of a set of methods for the comparison, selection, and visualization of numerous docking configurations. The combination of our proposed methods removes the problems of the existing solutions and provides the domain experts with an intuitive and user friendly tool for an interactive exploration of PPIs. The PPIs are provided by the computational tools. Our solution is designed for dealing with a large number of configurations, so the user is not required to apply any of the re-ranking algorithms before using our solution.

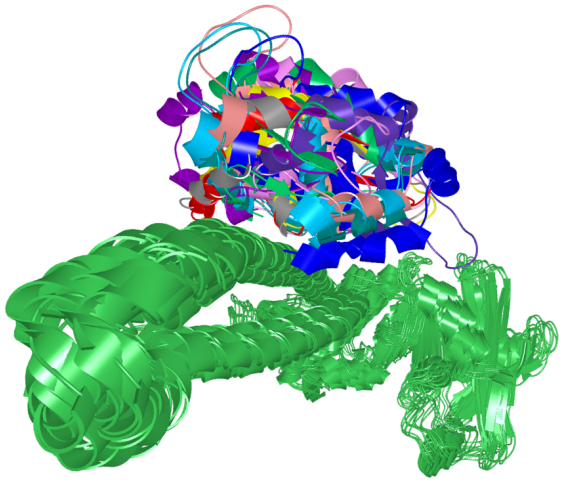

Figure 1 Typical visual representation of configurations used by the domain scientists that suffers from substantial visual clutter. It superposes several possible configurations between two proteins and visualizes them using traditional cartoon rendering. The set of green protein instances corresponds to one of the interacting proteins, the colored components represent the second protein in different configurations.

Related Work

Molecular interfaces have been studied already for decades. Early works, such as a publication by Varshney et al. [4], were focusing on both computation and visualization of these interfaces. The issue of visually representing PPIs can be tackled from different perspectives. One group of existing solutions focuses on the visualization of entire networks of interacting proteins. Because of their complexity, i.e., the number of interacting proteins, the visualizations are mostly graph-based. Jeanquartier et al. [5] presented a survey of databases enabling the visual analysis of protein networks.

The second group consists of techniques visualizing the contact zones and their interacting amino acids. The spatial techniques have to deal with the problem of occlusion and visual clutter caused by the fact that the most interesting parts of the interacting proteins, the contact zones, lie close to each other. Without transformations or visual enhancements (e.g., through transparency) it is impossible to visually explore the contact zones. Jin et al. [6] presented open-book view where the interacting proteins are rotated to orient the contact zones towards the camera. The problem of the presented solution lies mainly in the missing information about the interacting amino acids and the unified coloring of the contact zones. An alternative approach presented by Lee and Varshney [4] computes and visualizes the intermolecular negative volume and the area of the docking site.This way the users can observe the volume between the interacting proteins without a visual display of the contact zones themselves. This can serve the domain experts as an interactive tool for studying possible docking configurations. Ban et al. [7] presented an algorithm to construct the inter-

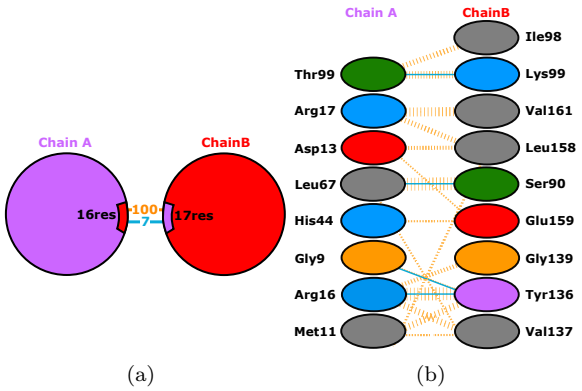

Figure 2 Two abstracted visualizations of the complex NSE1-NSE3 with PDB ID 3NW0 available in the PDBsum database. (a) Overview representation showing the number of amino acids in the contact zones and the types of interactions. (b) Part of the list of all interacting amino acids along with individual interactions and their strength. Images taken from the PDBsum database [9].

face surface between interacting proteins. The surface is visualized as a 3D mesh encoding the information about the core and peripheral regions of the interface.

One of our proposed spatial visualizations adapts the idea of so called exploded views. This technique enables the observation of the parts of objects, which are originally hidden. Bruckner et al. [8] applied this technique to volume data and demonstrated it on the scans of different parts of the human body.

Two-dimensional, abstract representations, are also commonly used for the visualization of contact zones, such as the schematic representation used by the PDBsum database [9] (see Figure 2). In an overview visualization each of the interacting proteins is represented by a sphere equipped with the information about the number of amino acids forming the contact zones and the number of different types of interactions in-between (e.g., salt bridges, disulphide bonds, hydrogen bonds, or non-bonded contacts). Another detailed visualization lists all contact-zone amino acids. The interactions are visualized by lines of different colors and thicknesses, which represent the type and strength of the interactions.

Lex et al. [10] proposed a visual analysis tool serving the exploration of large-scale heterogeneous genomics data for the characterization of cancer subtypes. They use multiple views onto the complex data and one of them is a method for the comparison of different datasets. The abstracted representation shows the similarities in the datasets by connecting the corresponding blocks of data. The thickness of a connection denotes the degree of similarity.

In another of our proposed methods, the Matrix View, we tackle the problem of visualizing many items in a limited 2D space. This may lead to having only little space available for individual items, which are then hard to perceive. Therefore, we also had to adapt one of the interactive lens techniques, which were thoroughly surveyed by Tominski et al. [11].

1 COZOID Overview

Our newly proposed system enables the efficient visual exploration of PPI complexes. A protein P consists of a set of amino acids forming the polypeptidic chain. The order of these amino acids and their properties determine the spatial arrangement of the protein and influence its behavior and function. A complex is represented by a set of mutually interacting proteins. In our case we focus primarily on interactions between two protein structures P_1 and P_2, determining a complex $C(P_1, P_2)$. The spatial orientation of the complex forms a configuration. The i-th configuration of complex $C(P_1, P_2)$, denoted as $CONF_i(C(P_1, P_2))$, represents one of the possible mutual orientations of this complex. Generally, there can be n $(1 \leq i \leq n)$ possible configurations for a given complex and the task is to select the configuration that is the most relevant one from a proteomical point of view. The decision is based on various pieces of knowledge about the geometric arrangement of the configuration as well as other aspects, such as the physico-chemical properties of the amino acids present in the contact zone of the given configuration. The selection of the most relevant configurations cannot be automatic and requires the insight of the domain expert. Therefore, this represents a typical domain-related problem, which has to be supported by specifically designed visualizations.

The visualization methods proposed in this paper allow the user to visually explore a set of possible configurations detected by one of the existing computational tools and to select the proteomically most relevant ones. The users have to iteratively filter out those configurations that do not fulfill the given specific criteria. We propose a workflow integrating a set of specific visualizations, as summarized in Figure 3. The input datasets, consisting of dozens of configurations between two interacting proteins, were computed using the HADDOCK [12] and the PyDock [13] tools. In the following sections, we describe our newly proposed techniques in detail.

The proposed visualizations have been designed in a tight cooperation with the domain experts and support a specific set of tasks defined by them. The techniques are based on the precondition that the users have already an initial knowledge about the interacting proteins. Thus the experts are able to define a pair of amino acids from the proteins that should interact. This information can be provided as an input parameter directly to the selected computational tool, such as

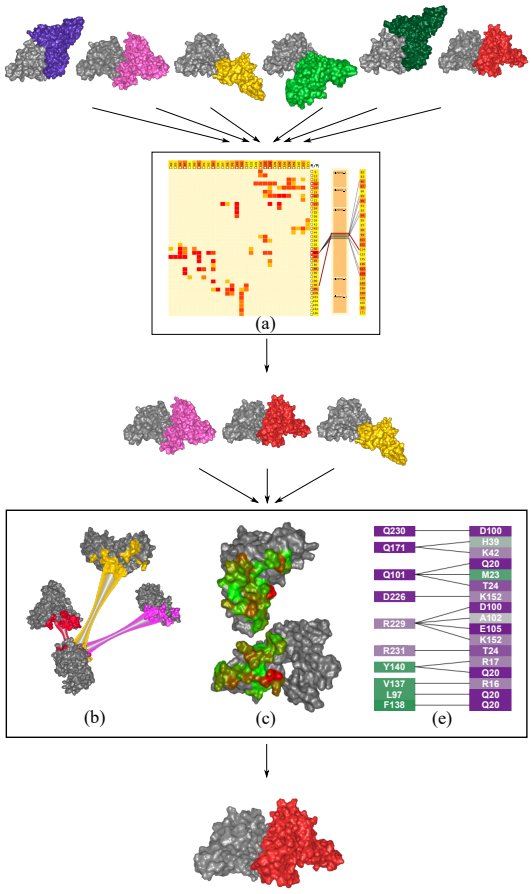

Figure 3 Workflow of the domain experts exploring the input configurations using our newly proposed methods. (a) The Matrix View represents an overview of all configurations. (b) The Exploded View enables us to explore the contact zones and their differences for a set of selected configurations. (c) The Open-Book View animates the opening of a selected configuration. (d) The Contact-Zone List-View supports to compare in detail the constitution of the contact zones of selected configurations.

HADDOCK. The second possibility is that the users do not have such information but are aware of an already explored protein complex, which can serve as a reference complex for further comparison and exploration.

Our methods have been designed in order to help the domain experts to answer the following questions:

- Q1: Which configurations contain a selected interacting pair of amino acids and what is the frequency of occurrence of this pair in all configurations?
- Q2: Which pairs of amino acids are present in a given configuration?
- Q3: How close are the amino acids in the contact zone and which are the closest ones?
- Q4: How similar and different are the contact zones in the configurations?
- Q5: What are the physico-chemical properties of the amino acids in the contact zone?
- Q6: What are the differences between the sets of amino acids in the contact zones of configurations?

To answer these questions, the domain-specific knowledge of the proteomic experts has to be combined with a proper insight into the configurations. This is the reason why it is impossible to fully automate the process of filtering the configurations and a proper visual support is crucial.

Answering these questions helps the proteomic experts to better understand the interactions in the protein-protein complexes. The proposed visualizations enable to find the answers by interactively exploring the configurations. The next step of the expert can lead to the selection of amino acids in the contact zones that could be mutated, i.e., replaced by other amino acids. The ultimate goal of such a mutation can be to strengthen the interactions in the contact zone or, otherwise, completely destroy the interaction between the involved proteins.

2 Matrix View

When using a computational tool for generating possible configurations, the resulting set of configurations $S = \{CONF_i(C(P_1, P_2)); 1 \leq i \leq n\}$ can be very large, where n ranges from dozens to hundreds, and requires some preliminary filtering. This filtering stage is based on answering question Q1. We propose a matrix-based visualization inspired by commonly used heat maps (see Figure 4 a). The rows and columns of the Matrix View correspond to interacting proteins P_1 and P_2. Each row or column represents one amino acid present in a contact zone of some of the configurations $CONF_i(C(P_1, P_2))$. The rows and columns are formed only by those amino acids of the interacting proteins which are in contact in at least one configuration. The contact between the amino acids is based on their Euclidean distance. Two amino acids are considered to be in contact if their distance is between 3 and 5 Ångströms. These thresholds can be interactively changed by the user. The color of each cell in the matrix corresponds to the number of occurrences of the corresponding interacting amino acids in the set S of all configurations. The colored lists of amino acids can be interpreted as histograms encoding the number of their occurrences. The intense red color represents pairs of amino acids that are interacting in most of the configurations. The Matrix View serves directly for filtering out improbable solutions by the interactive user-driven selection of cells. The interaction is performed by clicking on individual cells. Moreover, the matrix allows the expert the selection of a combination of several pairs of amino acids. This is useful if

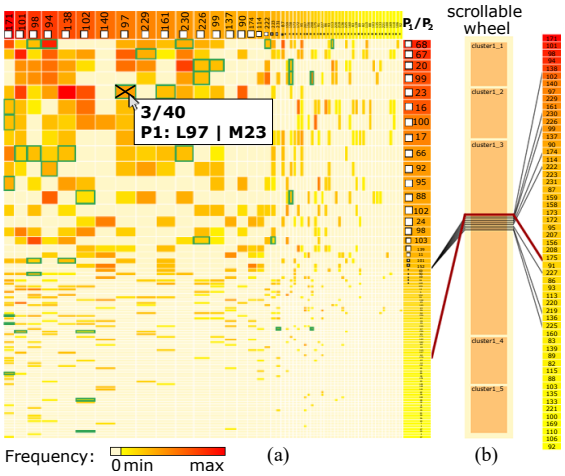

Figure 4 (a) Matrix View showing the aggregated information about the presence of mutually interacting amino acids in all configurations. Horizontal and vertical axes contain lists of amino acids in the contact zones of the interacting proteins P_1 and P_2. (b) The side view shows more detailed information about individual configurations and their interacting amino acids. The Matrix View is sorted according to the frequency of occurrence of the amino acids in all configurations. Green borders of some matrix cells represent the pairs which are present the configuration selected in the side view. Selected cells can be marked with a cross. It is possible to enlarge a selected row and column.

the user wants to further explore only those configurations that contain the interactions between, e.g., the amino acid pair A, B and/or simultaneously also pair C, D.

The rows and columns in the matrix can be also sorted according to the frequencies of the amino acids in the configurations. This results in the concentration of more frequent pairs of amino acids in the top left corner of the matrix (see Figure 4).

A big advantage of the Matrix View is its independence on the size of the input set of possible configurations. The number of rows and columns is limited by the size of interacting proteins, so in the worst case it corresponds to the total number of amino acids in these proteins. Usually, the number of amino acids in the contact zones is much smaller than the total number of amino acids. The number of rows and columns is much smaller as well. Each configuration of the input dataset increases the counters in the respective matrix cells. Therefore, the only parameter influencing the size of the matrix is the size of the interacting proteins. For large proteins the cells of the matrix can become too small. In such situations, the users can employ table lens techniques introduced by Rao and Card [14] that can be applied to both rows and columns of the matrix (see Figure 4).

In order to provide the users with more detailed information about individual configurations, the Matrix View contains an additional side view, which is positioned directly next to the matrix (see Figure 4 b). This helps to answer questions Q2 and Q3 as it enables the iterative search through the list of configurations and exploration of all pairs of interacting amino acids for each configuration. As visual representation, the list is displayed as a scrollable wheel. The user can scroll the view with the mouse wheel and explore individual configurations. The central part consists of rectangles representing individual configurations of a subset of S. The rectangle in the middle, corresponding to the configuration in focus, gets the largest screen space. The vertical list of amino acids is the same list as that one on the horizontal axis of the Matrix View. By default, each configuration in focus contains one polyline connecting those two amino acids from the contact zone which are the closest ones among all the possible pairs. The user can hover the mouse over the lists of amino acids on the left and right side and the corresponding connection lines for a given amino acid appear. By clicking on the rectangle representing a given amino acid, the connection lines remain in the view. The pairs of amino acids forming the configuration in focus can be highlighted in the matrix (with green border rectangles in Figure 4). The user immediately sees the number of configurations in which these pairs are present as well. Vice versa, by interacting with the matrix and selecting given cells, the side view can be automatically filtered to show only those configurations that satisfy the filtering condition.

The Matrix View serves as the first filtration tool for selecting only those configurations, which contain a desired combination of interacting amino acids. This filtering cannot be automated because the frequency of a given pair in configurations does not correlate with the importance of these configurations. The most frequent pair of interacting amino acids can be of the same interest as a pair interacting only in one configuration. Therefore, the insight of the domain expert in combination with the interaction possibilities of Matrix View proved to be a very efficient solution. Selected configurations can be further processed by the following visualization methods.

3 Exploded View

The proteomics experts are already familiar with the manipulation of molecules in a three-dimensional environment, thus 3D space has to be an integral part of the workflow. Moreover, the three-dimensional space helps to find answers for questions Q3-Q5, which are related to the appearance of the contact zone of selected configurations. Exploring and comparing many structures in 3D at once suffer from problems like high overlaps, occlusion, and visual clutter (Figure 5 a).

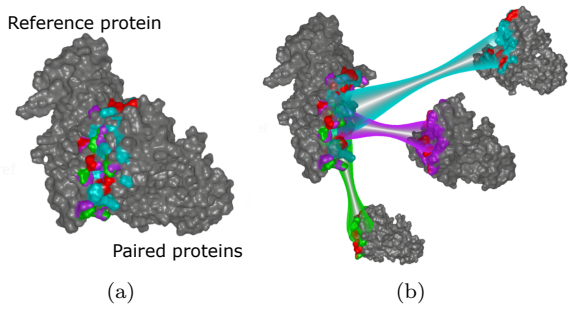

Figure 5 (a) Three aligned configurations represented by surfaces with highlighted contact zones. It is obvious that such a representation is insufficient for the observation and comparison of the contact zones. (b) Exploded View of these configurations. In both views, a different color is used for each configuration. Red spots in the contact zones signify amino acids in collision, i.e., in the distance less than 3 Å.

Traditionally used representations are not sufficient. To overcome the above mentioned limitations, we adapt the commonly used exploded-view technique, so as to enlarge the distance between the interacting proteins. Figure 5 (b) shows the comparison of three configurations using our proposed Exploded View. One of the interacting proteins is selected as a **reference protein**. The reference proteins from all configurations are structurally aligned so that their surface representations overlap. Here it is important to understand that the reference protein shown in the figure (the biggest one) actually consists of three overlapping aligned proteins, each coming from one configuration. The **paired proteins** from the interaction pairs are located around the reference ones, their distance is enlarged, and the corresponding contact zones are connected by colored tubes.

The main principle of the Exploded View is the following. First, all the reference proteins taken from the configurations selected by the Matrix View are aligned using a combinatorial extensions of the structural-alignment algorithm [15] (see the green proteins in Figure 1). The set of paired proteins interacting with the reference proteins is positioned around the aligned reference proteins.

To ensure that the paired proteins in the exploded view will not collide with each other, we employ a simple iterative force-directed placement algorithm, where paired proteins repulse each other [16]. For each reference protein and its paired protein, the Exploded View maintains the information about their interaction. If several configurations are exploded at once the Exploded View contains many paired proteins arranged around the aligned reference proteins and the pairing information must be indicated. We display the connection between corresponding reference proteins (aligned) and paired proteins as a partially transparent tube connecting the centres of their contact zones. The radius of the tube is modulated – it is smaller in the middle of the tube to reduce visual clutter. The tube can be switched on and off as needed.

Figure 5 depicts a set of three configurations before (a) and after (b) using the Exploded View. The exploded view removes the problem of overlapping paired proteins. It also enables the user to see the contact zones which were hidden by the interacting proteins in the original positions. However, this solution does not solve the problem that the contact zones still face each other so the user has to adjust the camera to observe the contact zones of the reference and paired proteins from a perpendicular viewing direction. Such a manipulation does not enable to see both contact zones simultaneously. This problem is solved by the proposed Open-Book View presented in the following section.

4 Open-Book View

The Exploded View does not allow an intuitive spatial navigation to the contact zones of a reference and its paired protein in detail and simultaneously. The Open-Book View is designed specifically to answer questions similar to Q4, Q5, and Q6, which deal with a detailed exploration of one selected contact zone of the complex $C(P_1, P_2)$. This involves the presentation of the information about different properties of individual amino acids forming the contact zone and their pairing.

The Open-Book View is activated if the user selects one of the configurations from the Exploded View. The selection is performed by clicking on the connection tube of the desired configuration $CONF_i(C(P_1, P_2))$ in the Exploded View. The other configurations present in the Exploded View are automatically hidden, the selected configuration returns to its initial position (before applying Exploded View), and an animated transition of the opening of the $CONF_i(C(P_1, P_2))$ is launched. By opening in the animation, the reference and paired proteins are rotated and translated so that they are positioned next to each other and the contact zones are facing towards the observer (see Figure 6).

The algorithm computes the vectors defining the orientation of the contact zones (their normal vectors). From the normal vectors and camera position we compute the rotation angle, which is then applied to the reference and paired protein. To maintain the information about the pairing of amino acids, the user can visualize also individual connections between these pairs by simple lines.

The contact zones represented by their surfaces can be color-coded according to multiple criteria. The color

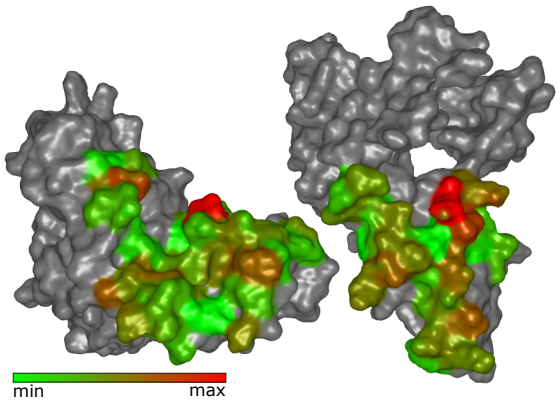

Figure 6 Open-Book View enables us to explore the contact zones between the interacting proteins simultaneously. On the left is the reference protein and on the right is the paired protein. The surface of the contact zones can be color-coded according to different criteria – here the color represents the distance between the pairs of amino acids.

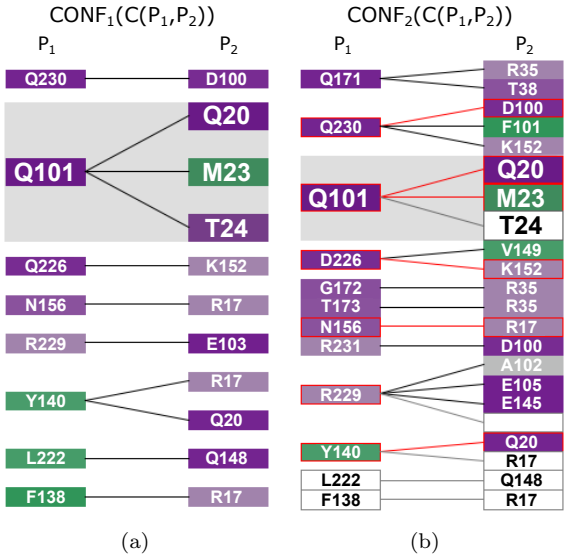

Figure 7 Contact-Zone List-View showing the comparison of one configuration as a primary one (a) with another configuration (b). For better comparison of configurations, the corresponding amino acids are interactively highlighted using zooming. The view is sorted (and colored) according to hydrophobicity of the amino acids in the P_1 protein.

can correspond to the distance between the amino acids, i.e., the closest amino acids have a different color than the most distant ones. Color can also represent different physico-chemical properties of the amino acids or their atoms, such as hydrophobicity or partial charges. The coloring scheme used in Matrix View represents so called conservation of the amino acids in all configurations. It can be used for coloring the contact zone as well. The surfaces can be augmented with labels to inform the users about the type and the identifier of individual amino acids.

In both, the Exploded View and the Open-Book View, a protein can be represented also by other traditionally used visualization styles, such as cartoon, spheres, balls&sticks, sticks, etc. Moreover, these methods can be combined. For example, the proteins can be represented by the cartoon style and the amino acids in the contact zones can be visualized using the sticks representation to see their spatial orientation.

The combination of the Exploded View and the Open-Book View is useful for the exploration of individual pairs. If the task is to compare individual configurations with respect to the pairs of interacting amino acids, a further drill-down is necessary. Therefore, in the next section we propose another abstracted view supporting mainly the comparison of the paired amino acids in individual contact zones of selected configurations.

5 Contact-Zone List-View

The Contact-Zone List-View helps to answer questions related to the comparison of the contact zones on the level of individual amino acids, such as in Q6. It consists of two sets of amino acids in the contact zones, each set coming from one interacting protein (see Figure 7). The left part contains all amino acids coming by default from the reference protein, the right part is formed by their counterparts in the paired protein. However, the order of proteins in the list view can be changed. The order depends on the current task, i.e., if we want to compare the constitution of contact zones of the reference or the paired protein in the given configurations. The view contains all possible connections (wrt. the distance) between amino acids from both contact zones. To avoid intersections of lines representing the connections, some amino acids on the right side are repeated – one instance for each amino acid in the reference protein within a user defined distance. This solution was adopted because without these repetitions there would be many line intersections which substantially decreases the readability of the representation.

For each configuration, one list view is created and all list views are juxtapositioned so the user can see and visually compare the constitution of the contact zones of all selected configurations. The user can interact with this representation by changing the properties of the amino acids mapped onto their corresponding rectangles. The properties are the same as those mapped to the surface of the contact zone in the 3D views. The left part of the list can then be sorted according to these properties (see Figure 8). Moreover, by clicking on individual rectangles representing amino acids, the corresponding amino acids are selected in the 3D view as well.

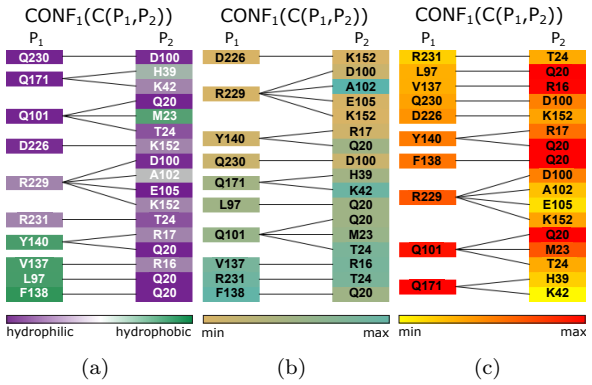

Figure 8 Sorting of the Contact-Zone List-View according to different properties of amino acids – (a) hydrophobicity, (b) mutual distance, (c) frequency of occurrence.

The principle steps for building the Contact-Zone List-View are the following. For all configurations, which should be visualized in the Contact-Zone List-View, we find the interacting pairs of amino acids in their contact zones. Then the list of amino acids present in all reference proteins of the selected configurations is created. Now, for each configuration, we take the interacting amino acids from paired proteins, sort them according to a selected criterion, and add them to the Contact-Zone List-View. The amino acids in the left part of the Contact-Zone List-View are always sorted in the same way for all depicted configurations. To enable comparison, the user can select a primary configuration to which all the remaining configurations are compared (see Figure 7 b). An example of such a primary configuration can be a reference crystal structure downloaded from the PDB database. We propose the following ranking score, which indicates how similar the contact zone of one configuration is to the contact zone of the primary configuration. The score is computed in the following way. For each match of an amino acid in the contact zone of the reference protein with the selected configurations the similarity score is increased by one. For each matching pair of interacting amino acids in contact zones of the reference and paired proteins of the selected configurations the similarity score is increased by four. For each missing pair of interacting amino acids in the contact zones of the selected configurations the similarity score is decreased by one. The Contact-Zone List plots the configurations ordered by this score from the most similar to the least similar ones. The Contact-Zone List-View of the primary configuration is always displayed as first.

The user can select between the compare and the compact visualization modes. In the compare mode, the amino acids from the contact zone of the primary configuration that are not present in the contact zone of any other configuration, are depicted as white rectangles with labels giving the names of the missing amino acids (see Figure 7 b). The compact mode omits the missing amino acids to save space. In both modes, the matches to amino acids in the primary configuration are highlighted with red bounding rectangles and connecting lines. This way, the user can immediately see which amino acids are present in both, the primary configuration as well as other configurations, and which amino acids are missing. To guide the visual comparison, we also introduce interactive highlighting and, if necessary, zooming to corresponding amino acids in different configurations.

6 Demonstration and Results

All presented techniques and their integration into the proposed system were designed in a tight collaboration with domain experts from the field of functional proteomics. Their current main research focus is on structure maintenance of chromosome (SMC) complexes [17]. The SMC complexes are key players in the chromatin organization where they ensure the stability and dynamics of chromosomes. The researchers analyze the architecture and function of such complexes using a variety of experimental approaches. The goal is to uncover the way the subunits of these complexes interact with each other and execute unique functions of these complexes [18]. A visual representation of such an information is highly beneficial because it helps to reveal the spatial relationships between the subunits in an intuitive way.

To demonstrate the usability of our proposed techniques, the proteomic experts selected representatives of three basic types of PPI patterns, according to the shape of the contact zones and interacting protein domains. These types are the coiled-coil interaction, pocket-string interaction, and surface-surface interaction [19]. In the following subsections, we present examples of these three types of interaction.

6.1 Surface-Surface Interaction

The surface-surface interaction between the NSE1 and NSE3 proteins has been analyzed as it represents the dimer of kite proteins, which are critical for the function of the SMC5/6 and the bacterial SMC complexes [20, 17, 21].

The crystal structure of the human NSE1-NSE3 dimer was already examined in detail and the resulting configuration is already published in the PDBsum database under the PDB identifier 3NW0. Therefore, it can serve as a testing reference complex for both computational tools as well as our proposed visualizations, which have been integrated in our COZOID system. The proteomic expert conducting this study

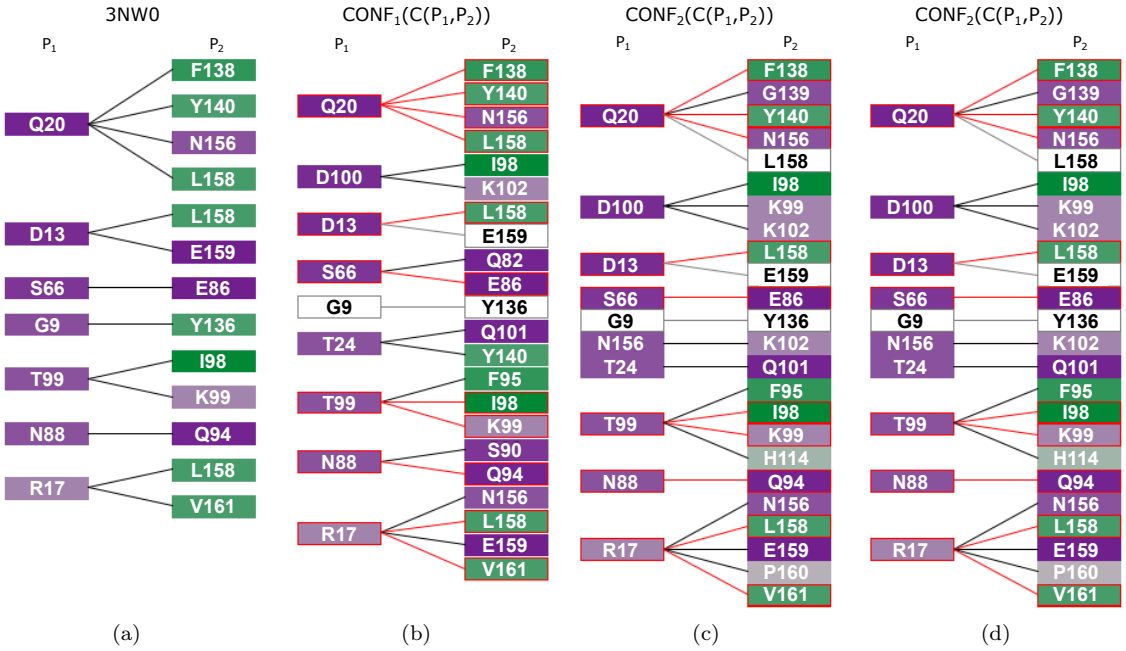

Figure 9 Example of four configurations represented by juxtapositioned Contact-Zone List-View. (a) Reference 3NW0 crystal structure, (b), (c), (d) three selected HADDOCK models. Lists are colored and sorted according to hydrophobicity of the amino acids in the reference protein in each configuration.

provided docking results and prior knowledge about interacting amino acids. To restrict the set of possible docking configurations, he selected the web version of the HADDOCK tool and a pair of interacting amino acids, i.e., methionine with ID 23 from the reference protein and leucine with ID 97 from the paired protein. This selection was based on experimental data from previous work [22, 23, 24]. The tool resulted in 40 configurations. HADDOCK groups the configurations into clusters, according to a similarity, which is defined internally by HADDOCK. In our case it led to 10 clusters, each containing 4 configurations.

All computed configurations were loaded into our visualization system, which includes all proposed visualizations. From these configurations the Matrix View was computed, which contains the frequencies of all pairs of amino acids within the interaction distance within these 40 configurations. The matrix identified configurations containing the pairs of interacting amino acids with the interaction distances of less than 4 Å. In our particular case, the leucine 97 and methionine 23 amino acids were in contact distance in only three configurations.

In the next step, the proteomic expert switched to the Contact-Zone List-View and compared the list of amino acids of the 3NW0 crystal structure with the lists of all three selected configurations. Figure 9 shows the comparison between the 3NW0 structure and three HADDOCK configurations.

Even from the given part of the Contact-Zone List-View, the similarities and differences between the crystal 3NW0 and the three selected HADDOCK configurations on the level of individual amino acids are clearly visible. In addition, pairs of the interacting amino acids identical to the 3NW0 crystal structure are highlighted (red lines in Figure 9). The order of the modeled configurations in Figure 9 reflects their similarity to the crystal based on the number of identical pairs of amino acids (the best model is next to the crystal).

Finally, the 3NW0 crystal and three selected configurations were explored in 3D space. In 3NW0, the first interacting protein was selected as the reference protein and all three configurations were aligned with respect to paired proteins. The paired proteins were positioned around the reference one. Figure 5 (a) shows the situation where the three configurations are visualized using the commonly available method. Configurations are represented as surfaces and the contact zones are highlighted using different colors. However, the most interesting parts, i.e., the contact zones, are hidden.

Our Exploded View overcame this limitation (Figure 5 b), so the individual contact zones on all paired proteins are clearly visible. Moreover, if the user is pointing the camera towards the aligned reference proteins, the differences between the positions of the contact zones in the reference proteins can be observed as

well. The Exploded View representation gave the proteomic expert the information about the mutual positioning of the individual configurations with respect to the positions of the contact zones.

The investigation has to go more deeply into the level where individual contact zones can be explored in detail. In this case, each configuration can be explored individually using the Open-Book View. By animating the opening of the contact zone the user was able to look inside the contact zone and perceive individual amino acids. The enhancements of the Open-Book View, i.e., labeling the surface of the contact zones with the names of the corresponding amino acids, and coloring according to different criteria, were highly appreciated by the domain expert.

Our tool can be used also for selecting an alternative input pair of interacting amino acids, which then serves as the input for the computational tools. These amino acids might be selected based on a COZOID analysis of the 3NW0 crystal – using the Matrix View or Exploded View (searching for the most central and the closest amino acids).

Altogether, our tool helped to quickly select the best docking configuration using several visualization approaches. First, Matrix View allowed the biochemist to pick models containing a particular pair of interacting amino acids. Next, with the Contact-Zone List the biochemist sorted these models based on the similarity of their contacts with the original crystal structure. Using the 3D Exploded View, the best model was determined and confirmed. While an exploded view is available in some of the current 3D visualization tools, the power of our other approaches lies in the speed and elegance of the selection mechanism. In addition, a similar workflow can be applied to the selection of docking models of homologous proteins not available in the PDB database but very often used when different model organisms are employed in proteomic studies.

6.2 Coiled-Coil Interaction

For the second type of interaction, the biochemist picked the SMC3 coiled-coil arm of the SMC complex [18]. The interaction site is formed by two helical fragments of the SMC3 protein. The reference structure is published under the PDB identifier 4UX3.

Using this structure, the results of both HADDOCK and PyDock tools were tested. The HADDOCK results contained 40 output configurations. Using the Matrix View, the biochemist set the interaction distance threshold between 3 and 5 Å and selected the methionine 186 and isoleucine 1030 pair of interacting amino acids (Figure 10). These amino acids were used as the input restraints for the HADDOCK computation. These restraints configurations were selected for further exploration.

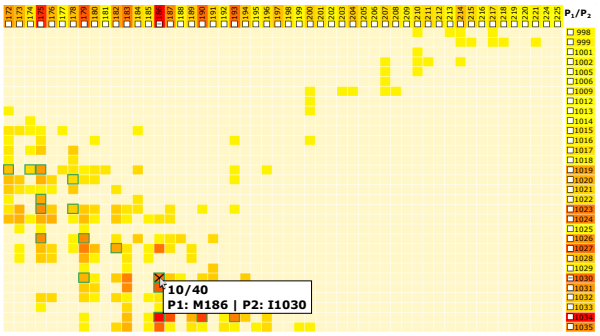

Figure 10 The Matrix View of interacting amino acids in all HADDOCK models. The Matrix View indicates that the selected pair of M186 and I1030 amino acids is present in 10 out of 40 loaded models.

Next, the selected configurations were structurally aligned to the reference 4UX3 protein. Afterwards, the biochemist selected the first amino acid (A172) within the respective helices and visually compared the positions in the 3D view. In this case it was not necessary to use other views to see that the preselected HADDOCK configurations exhibited a wrong orientation of two helices. In all 40 output models these A172 amino acids were located on the opposite side of the chain in comparison with the reference crystal 4UX3 (see Figure 11).

The 3D view of COZOID helped to reveal this misorientation intuitively and quickly, without a detailed exploration of all 40 configurations one by one.

As for the PyDock results, 28 out of 100 output PyDock models were selected using Matrix View and residues M186 and I1030 have been used. The next selection provided the biochemist with 14 models with correct orientation (see Figure 12), which was a significant improvement in comparison to the HADDOCK results.

In the final step, the biochemist compared Contact-Zone Lists of our selected models with the original crystal structure (4UX3). Figure 13 shows similarities (highlighted in red) of one of our models to the crys-

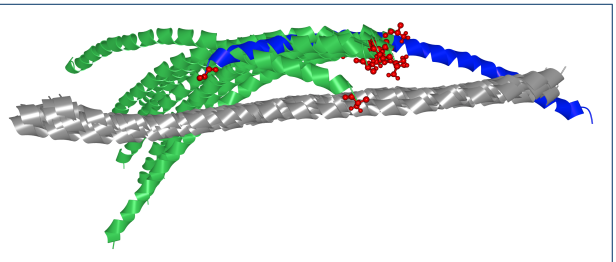

Figure 11 4UX3 crystal (blue) and selected configurations computed by HADDOCK (green). The A172 amino acid (red) is highlighted in all loaded structures. The opposite orientation of 4UX3 and HADDOCK models is clearly visible.

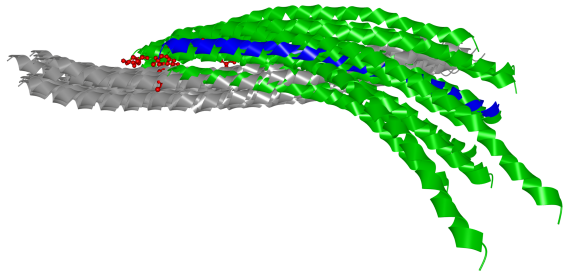

Figure 12 14 PyDock models (green) in the correct orientation with the 4UX3 crystal (blue). This can be seen by the A172 amino acid visualized using balls and sticks (red).

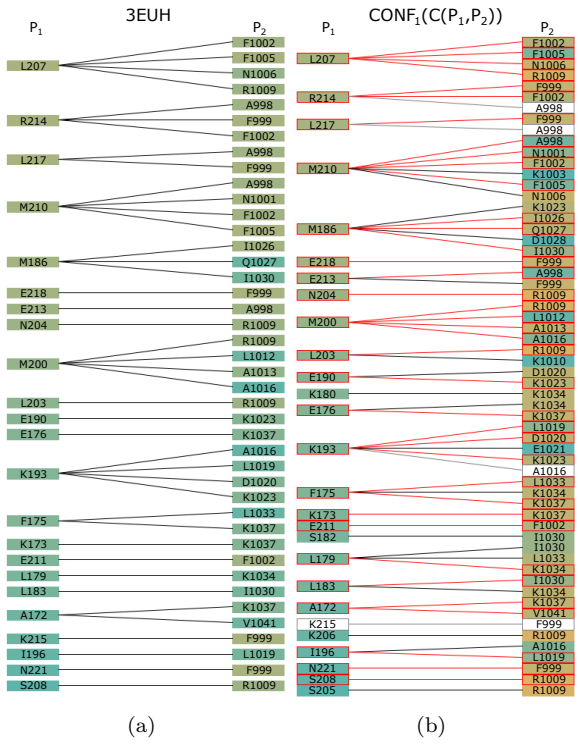

Figure 13 Contact-Zone List comparing the 4UX3 crystal with one best fit PyDock model with respect to the distances of amino acids.

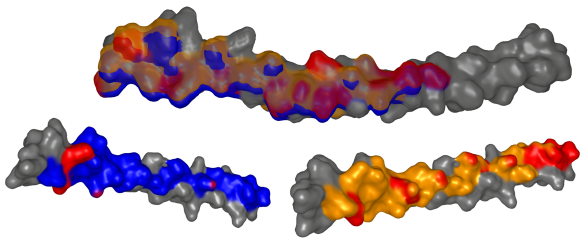

Figure 14 Open-Book View showing contact zone of the best fitting PyDock model (orange) and 4UX3 crystal (blue).

tal. It is the best model, which fits the crystal structure very well. The comparison of the contact zone of crystal structure and the selected model can be seen in Figure 14.

6.3 Pocket-String Interaction

In this case, the biochemist selected the interaction present in the crystal structure of the MukE-MukF Complex – proteins involved in chromosome partitioning in Escherichia coli. The crystal structure is published under the PDB identifier 3EUH [25]. The pocket-string interaction is present between two chains of this structure – the pocket is formed by the winged helix domain of the MukE protein, while one of the MukF helical fragments is sitting inside the MukE pocket (Figure 15 a). This time, the biochemists picked valin 200 and arginine 300 as the pair of amino acids for the docking restraint. These were the closest contact amino acids of the structure, as can be seen from the Contact-Zone List ordered by distance of interacting amino acids (see Figure 16), as well as from an Open-Book View of the crystal structure (Figure 15 b).

The docking models were generated with both HADDOCK and PyDock docking tools. The HADDOCK run resulted in 32 output configurations, which were first scrutinized using Matrix View and the V200-R300 amino acid pair. This first selection step filtered away only 8 models leaving 24 models for further analysis. Then, the biochemist repeated a Matrix View filtering using the second tightest amino acid contact in the crystal – tyrosin 110 and arginin 302. This filtration resulted in 6 docking models. Contact-Zone Lists of these models were compared with the original crystal structure (3EUH) resulting in an ordered list of the best models. Visual exploration confirmed, that the first model from Contact-Zone List fits best to the original structure (Figure 17).

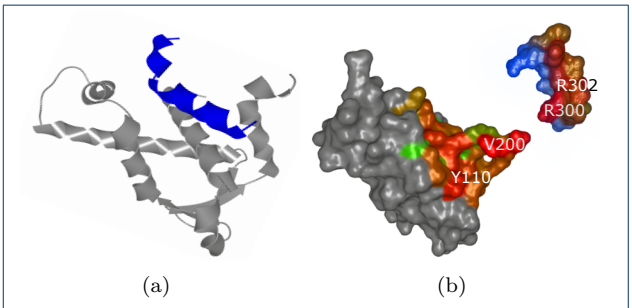

Figure 15 (a) 3EUH crystal structure consisting of the protein containing the pocket (grey) and the helical fragment of the second protein (blue) shown with traditional cartoon rendering. (b) The same structure shown with the Open-Book View. The contact zones are colored according to the distance between interacting amino acids and the labels for the two closest pairs are shown.

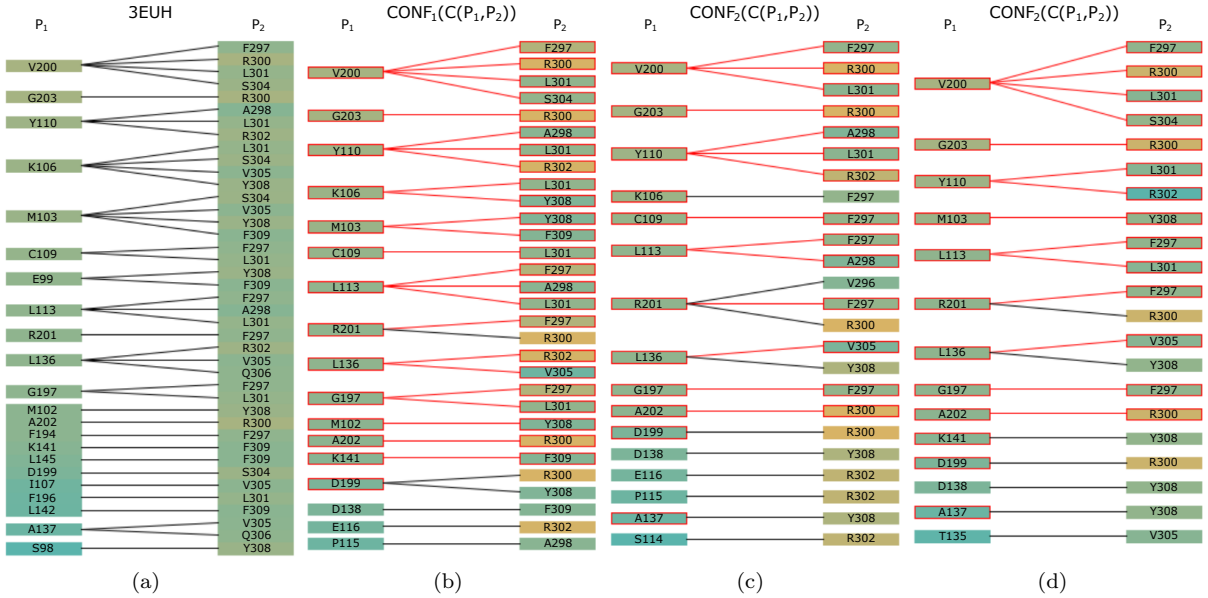

Figure 16 Contact-Zone List sorted according to the distance of amino acids shows that the V200-R300 pair is one of the closest ones in the crystal in all selected models.

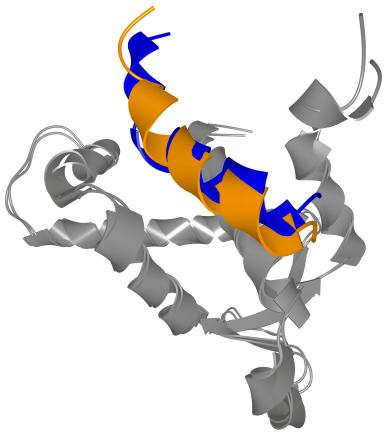

Figure 17 The best fitting HADDOCK model (orange) aligned with the crystal (blue).

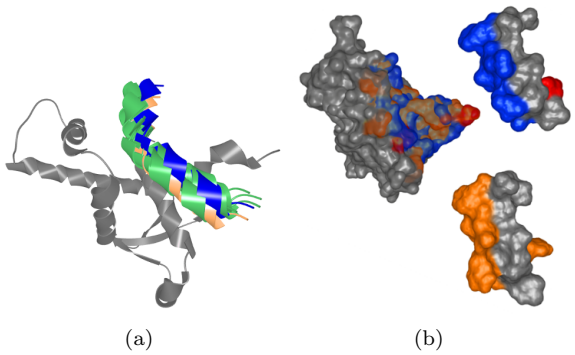

Figure 18 (a) Best 5 PyDock models with the distance parameter 4 Å. The crystal is blue. The model, which exhibited the most similar contacts with the crystal, is orange and the remaining models are green. (b) Exploded View showing comparison of contact zones of best PyDock model (orange) with crystal (blue).

PyDock docking provided 100 models, which were analyzed in the same way as the HADDOCK models. Selection steps with Matrix View resulted in 32 and then 19 models, respectively. Contact-Zone Lists of these models were then compared with the original crystal structure. The models with Contact-Zone Lists matching most closely to crystal structure were then visually explored in 3D using Exploded View and Open-Book View. This step revealed that the best five models from the list are very close to the original crystal, but none of them fits precisely to the crystal structure.

Here the biochemist took advantage of our test set-up (using the tightest contacts between interacting amino acids) and altered the distance parameter in Matrix View for the selection procedure. All PyDOCK models were reevaluated with the distance parameter 4 Å (compared to the 5 Å default parameter). The Matrix View selection steps resulted in 21 and 13 models, respectively. Based on the Contact-Zone List and a 3D view analysis using Exploded View and Open-Book View, five models most similar to the crystal were selected again (Figure 18 a). Four out of these five models overlapped with five best models detected with

previous system set-up, however a new model with the closest match was identified (Figure 18 b).

This test indicates the robustness of our tool at different parameter settings and its potential for proteomic experimental use. For example, our Contact-Zone List can be used in the experimental design of mutants by disturbing key contact residues.

7 Conclusion

In this paper we have presented COZOID, a novel system for the visual exploration of configurations of two interacting proteins. It integrates and modifies a set of visualization methods for the exploration and evaluation of the biochemical relevance of large sets of configurations detected by existing computational tools. Our proposed methods were designed to follow the workflow of the proteomic experts. We described the design rationale and principles of the methods as well as interaction possibilities. The methods were tested by the proteomic experts on real datasets for structure maintenance of chromosome complexes and we demonstrated the usability on one of the executed case studies. The domain experts confirmed that our proposed solution provides them with information, which was very hard or even impossible to get using the previously available methods. They confirmed that using our solution their exploration process can lead to a satisfying conclusion about the biochemical relevance of individual configurations much faster. The system enables the iterative filtering of the configurations that do not satisfy the given conditions in individual stages of the workflow.

In the future we will focus on the extension of our proposed techniques to cases where the user has no a priori knowledge about the protein complex. We also plan to extend the methods to complexes consisting of more than two interacting proteins. Additionally, the proteomic experts are already using the COZOID tool for training students of proteomics to analyze and understand the crystal structures as well as computed models. Their feedback will lead to further improvements of our tool.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

KF and JB participated on the design and implementation of the tool. JJP and BK were responsible for the design and evaluation by the domain experts. With KF they were also major contributors in writing the manuscript. EMG and IV contributed to the design and shaped the manuscript content.

Acknowledgements

This work was supported through grants from the Vienna Science and Technology Fund (WWTF) through project VRG11-010, the OeAD ICM and MSMT-1492/2015-1 through project CZ 17/2015, the PhysioIllustration research project 218023 funded by the Norwegian Research Council, Czech Science Foundation grant GA13-00774S and the Ministry of Education, Youth and Sports of the Czech Republic project CEITEC 2020 (LQ1601), and an Internal Masaryk University grant (MU/0822/2015).

Author details

[1]Masaryk University, Brno, Czech Republic. [2]University of Bergen, Bergen, Norway. [3]TU Wien, Wien, Austria.

References

1. Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B., Superti-Furga, G.: Proteome survey reveals modularity of the yeast cell machinery. Nature **440**(7084), 631–636 (2006)
2. Huang, S.-Y.: Search strategies and evaluation in protein-protein docking: principles, advances and challenges. Drug Discovery Today **19**(8), 1081–1096 (2014)
3. Malhotra, S., Mathew, O.K., Sowdhamini, R.: DOCKSCORE: a webserver for ranking protein-protein docked poses. BMC Bioinformatics **16**(1), 1–6 (2015)
4. Lee, C., Varshney, A.: Computing and Displaying Intermolecular Negative Volume for Docking. In: Scientific Visualization: The Visual Extraction of Knowledge from Data, pp. 49–64. Springer, Berlin, Heidelberg (2006)
5. Jeanquartier, F., Jean-Quartier, C., Holzinger, A.: Integrated web visualizations for protein-protein interaction databases. BMC Bioinformatics **16**(1), 1–16 (2015)
6. Jin, L., Wang, W., Fang, G.: Targeting Protein-Protein Interaction by Small Molecules. Annual Review of Pharmacology and Toxicology **54**, 435–456 (2014)
7. Ban, Y.-E.A., Edelsbrunner, H., Rudolph, J.: Interface Surfaces for Protein-Protein Complexes. Journal of the ACM (JACM) **53**(3), 361–378 (2006)
8. Bruckner, S., Gröller, M.E.: Exploded Views for Volume Data. IEEE Transactions on Visualization and Computer Graphics **12**(5), 1077–1084 (2006)
9. Laskowski, R.A., Hutchinson, E.G., Michie, A.D., Wallace, A.C., Jones, M.L., Thornton, J.M.: PDBsum: a Web-based database of summaries and analyses of all PDB structures. Trends in Biochemical Sciences **22**(12), 488–490 (1997)
10. Lex, A., Streit, M., Schulz, H., Partl, C., Schmalstieg, D., Park, P.J., Gehlenborg, N.: StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. Computer Graphics Forum **31**(3), 1175–1184 (2012)
11. Tominski, C., Gladisch, S., Kister, U., Dachselt, R., Schumann, H.: A Survey on Interactive Lenses in Visualization. In: Borgo, R., Maciejewski, R., Viola, I. (eds.) Eurographics Conference on Visualization - STARs. The Eurographics Association, Swansea (2014)
12. Dominguez, C., Boelens, R., Bonvin, A.M.J.J.: HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information. Journal of the American Chemical Society **125**(7), 1731–1737 (2003)
13. Jimenez-Garcia, B., Pons, C., Fernandez-Recio, J.: pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. Bioinformatics **29**(13), 1698–1699 (2013)
14. Rao, R., Card, S.K.: The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '94, pp. 318–322. ACM, Boston (1994)
15. Shindyalov, I.N., Bourne, P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering **11**(9), 739–747 (1998)
16. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. Software: Practice and Experience **21**(11), 1129–1164 (1991)
17. Paleček, J.J., Gruber, S.: Kite Proteins: a Superfamily of SMC/Kleisin Partners Conserved Across Bacteria, Archaea, and Eukaryotes.

Structure **23**(12), 2183–2190 (2015)
18. Gligoris, T., Lowe, J.: Structural Insights into Ring Formation of Cohesin and Related Smc Complexes. Trends Cell Biology **26**(9), 680–693 (2016)
19. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P.: Molecular Biology of the Cell, Fourth Edition, 4th edn. Garland Science, New York (2002)
20. Zabrady, K., Adamus, M., Vondrová, L., Liao, C., Skoupilová, H., Nováková, M., Jurcisinová, L., Alt, A., Oliver, A.W., Lehmann, A.R., Paleček, J.J.: Chromatin association of the SMC5/6 complex is dependent on binding of its NSE3 subunit to DNA. Nucleic Acids Research **44**(3), 1064–1079 (2016)
21. Doyle, J.M., Gao, J., Wang, J., Yang, M., Potts, P.R.: MAGE-RING Protein Complexes Comprise a Family of E3 Ubiquitin Ligases. Molecular Cell **39**(6), 963–974 (2010)
22. Hudson, J.J., Bednářová, K., Kozáková, L., Liao, C., Guerineau, M., Colnaghi, R., Vidot, S., Marek, J., Bathula, S.R., Lehmann, A.R., Paleček, J.: Interactions between the Nse3 and Nse4 components of the SMC5-6 complex identify evolutionarily conserved interactions between MAGE and EID Families. PLoS ONE **6**(2), 17270 (2011)
23. Kozáková, L., Vondrová, L., Stejskal, K., Charalabous, P., Kolesár, P., Lehmann, A.R., Uldrijan, S., Sanderson, C.M., Zdráhal, Z., Paleček, J.J.: The melanoma-associated antigen 1 (MAGEA1) protein stimulates the E3 ubiquitin-ligase activity of TRIM31 within a TRIM31-MAGEA1-NSE4 complex. Cell Cycle **14**(6), 920–930 (2015)
24. van der Crabben, S.N., Hennus, M.P., McGregor, G.A., Ritter, D.I., Nagamani, S.C.S., Wells, O.S., Harakalová, M., Chinn, I.K., Alt, A., Vondrová, L., Hochstenbach, R., van Montfrans, J.M., Terheggen-Lagro, S.W., van Lieshout, S., van Roosmalen, M.J., Renkens, I., Duran, K., Nijman, I.J., Kloosterman, W.P., Hennekam, E., Orange, J.S., van Hasselt, P.M., Wheeler, D.A., Paleček, J.J., Lehmann, A.R., Oliver, A.W., Pearl, L.H., Plon, S.E., Murray, J.M., van Haaften, G.: Destabilized SMC5/6 complex leads to chromosome breakage syndrome with severe lung disease. The Journal of Clinical Investigation **126**(8), 2881–2892 (2016)
25. Woo, J.S., Lim, J.H., Shin, H.C., Suh, M.K., Ku, B., Lee, K.H., Joo, K., Robinson, H., Lee, J., Park, S.Y., Ha, N.C., Oh, B.H.: Structural Studies of a Bacterial Condensin Complex Reveal ATP-Dependent Disruption of Intersubunit Interactions. Cell **136**(1), 85–96 (2009)