Visualization of Solution Sets from Automated Docking of Molecular Structures

Johannes Jansen van Vuuren [*] University of Cape Town
Michelle Kuttel [†] University of Cape Town
James Gain [‡] University of Cape Town

Abstract

Aligning structures, often referred to as docking or registration, is frequently required in fields such as computer science, robotics and structural biology. The task of aligning the structures is usually automated, but due to noise and imprecision, the user often needs to evaluate the results before a final decision can be made. The solutions involved are of a multidimensional nature and normally densely populated. Therefore, some form of visualization is necessary, especially if users want to achieve higher level understanding, such as solution symmetry or clustering, from the data.

We have developed a system that provides two views of the data. One view places focus on the orientation of the solutions and the other focuses on translations. Solutions within the views are cross-linked using various visual cues. Users are also able to apply various filters, intelligently reducing the solution set. We applied the visualization to data generated by the automated cryo-EM process of docking molecular structures into electron density maps. Current systems in this field only allow for visual representation of a single solution or a numerical list of the data. We evaluated the system through a multi-phase user study and found that the users were able to gain a better high-level understanding of the data, even in cases of relatively small solution sets.

CR Categories: I.3.6 [Computer Graphics]: Methodology and Techniques; I.6.8 [Simulation and Modeling]: Types of Simulation

Keywords: multi-dimensional visualization, molecular docking, glyphs

1 Introduction

Registration, or docking, involves bringing two or more structures into as close an alignment as possible within a single coordinate system. It is a problem common to a number of disciplines, including medicine, computer vision, biology and computer science [Zitov and Flusser 2003]. The nature of the data varies with domain. For instance, volumetric images constituted from multiple slices are typically produced in medical imaging, while dense point clouds are generated from laser range scans by mobile robots. All forms of registration do, however, have in common that the final solution is simply a rigid-body transformation (in the case of non-deforming registration) and an associated fitness score. Automated algorithms exist for registration but have difficulties in coping with noise or resolution mismatches. For instance, laser range scanning may produce significant outliers, while medical imaging has to deal with aligning different modalities with differing resolutions, such as MRI, X-ray and PET for the purposes of diagnosis [Pluim et al. 2000]. This often leads to a proliferation of solutions in which the optimal registration may not achieve the highest fitness.

As a consequence, human intervention is required, but users cannot be expected to analyze a numerical solution list with thousands of potential entries. Each solution can be decomposed into a translation (3-vector), rotation (3-vector) and fitness correlation (single value), thereby producing a 7-dimensional space. Without some form of visualization it is next to impossible for users to derive an understanding of high-level properties within these solution sets, in particular clustering, symmetry and mirroring that point to the structure of the true solution.

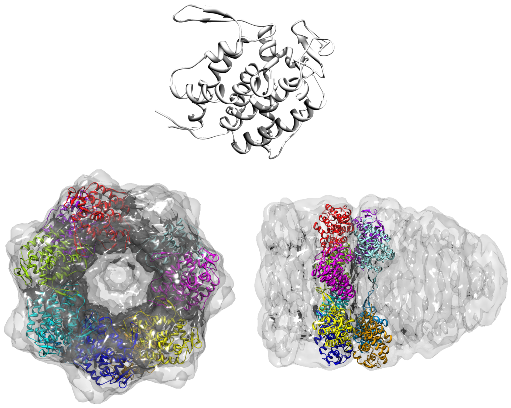


Figure 1: *Molecular Docking. [Top] A GroEL molecular sub-unit, derived from X-Ray Crystallography and represented using a ribbon visualization. [Bottom] The GroEL molecule docked in several configurations into an electron density map, depicted from a front and side view.*

In this paper, we present a visualization tool to display the solution sets generated by registration algorithms. We focus on a specific application of registration in structural biology (see Figure 1): docking of high resolution molecular structures produced by X-Ray Crystallography (XRC) experiments into electron density maps produced by Cryo-Electron Microscopy (Cryo-EM) experiments. XRC structures are typically high resolution, but reflect a molecule in a solid-state regular crystal environment. This experimental technique is limited to molecules that can be crystalized, which usually precludes complex molecular assemblies. Conversely, Cryo-EM techniques are able to produce three-dimensional images of large molecular assemblies in solution, which much more closely matches the biological environment. However, current Cryo-EM techniques produce fairly low-resolution images. It is important to model molecular structures in as much detail as possible, since a molecule's structure is expected to be intimately related to its function. The process of fitting a high-resolution X-Ray Crystallography (XRC) structure into a low-resolution electron density map generates a pseudo-structure of a large assembly at atomic resolution. The docking procedure must be accurate, otherwise supportable conclusions cannot be inferred.

Automated approaches to this type of molecular docking are supported by software packages, such as Situs [Wriggers et al. 1999],

[*] e-mail:johannes.jvv@gmail.com
[†] e-mail:mkuttel@cs.uct.ac.za
[‡] e-mail:jgain@cs.uct.ac.za

Afrigraph 2010, Franschhoek, South Africa, June 21 – 23, 2010.
© 2010 ACM 978-1-4503-0118-3/10/0006 $10.00

DockEM [Roseman 2000] and EMatch [Lasker et al. 2007], but they currently only allow the graphical inspection of single solutions or evaluations of numerical list data.

We provide a multi-view visualization of large solution sets which splits the focus between translation and rotation. Care is taken to ensure visual correlation between these two views. Furthermore, our system supports dynamic filtering, single solution selection, and associated numerical details on demand. We evaluated the visualization through a user experiment that compares the current numerical list of solutions against our visualization and this shows with statistical significance, that identification, comparison and high-level understanding of solutions are improved. In the final part of this paper, we present a case study of the visualization applied to the docking of a GroEL molecular sub-unit [Fayet et al. 1989] into an electron density map.

2 Related Work

There are many visualization schemes that present higher dimensional data in two- or three-dimensions. The most appropriate choice will naturally be one that respects the characteristics of registration solution space. This space is discretely sampled, so visualizations that rely on continuous input data tend to be less appropriate. Furthermore, there are two natural subspaces – rotation and translation – inherent in the 7-dimensional data, so the choice of visualization should follow from this natural decomposition. The problem of multidimensional and multivariate visualization is well studied. For general coverage of the field we recommend the surveys of Wong and Bergeron [1997] and Ellis and Dix [2007]. Within the field of molecular analysis the work of Pettersen et al. [Pettersen et al. 2004] and Moll et al. [Moll et al. 2006] have relevance to our system, although they address more general visualization. We focus more narrowly here on techniques with potential relevance to our problem domain.

2.1 Two dimensional visualization techniques

There are a number of effective two-dimensional methods for visualizing high-dimensional datasets centered around creating relationships between data items. For example, Assa et al. [1997] evaluate each data item according to defined relations, and assign a correlation factor to indicate how well each data item and relation match. The multidimensional data is displayed using relevance maps, in which each relation is converted into a gravitation node. Each node is mapped onto a two-dimensional plane and serves as an attractor for the data items, which are represented by points on the plane. A point is placed close to a node if it has a high relevance to that node, and between two nodes if it has relevance to both; the relative distance indicating the relevance. A related method is the Parallel Coordinates approach [Inselburg 1985]. This constructs relations, as follows: points in the data sets are mapped to lines between variables, which represent the relationship between two variables. This enables the user to draw conclusions based on visual cues gained from the parallel coordinate representation. The user is also able to perform simple and complex queries on the visualized data, which will eliminate and highlight certain data items. Both of these techniques can be powerful tools for data mining. However, they are most useful when working with relationships between independent variables. In our case, the three variables used to represent a single translation or orientation are tightly coupled and cannot meaningfully be separated.

Another approach, as embodied by the VisBio package [Hibbard 2003], is to take a two-dimensional slice through the solution space as is often done in medical applications. For instance, a single slice might represent a constant z-translation and set rotation and show various x-y translations and corresponding fitnesses using a colour scale. The slice representation is somewhat limited, in that the system does not allow for viewing the entire solution space at once. The user must scroll through each slice. Another drawback is that the visualization does not intuitively reflect the changes that the information undergoes as a particular variable changes. In other words, each slice exists separately and, thus, one has to switch between slices to see differences. Particularly problematic is that this scheme fragments the natural three-dimensional subspaces.

2.2 Three-dimensional visualization

Feiner et al. [1999] advocate a three-dimensional approach to multi-dimensional visualization, termed “worlds within worlds”. In this system, the higher data dimensions are removed and then added back in a controlled fashion. Removed dimensions are returned as three-dimensional subsets of the primary three dimensions. Each of these 3D sets is represented as a height map, where higher coordinate values are displayed as higher regions in the height map. This visualization is able to successfully represent abstract multidimensional worlds. Worlds within worlds is difficult to apply to our solution since it relies on continuous data, whereas registration is discretely sampled.

Dos Santos et al. [2002] propose another method of 2D or 3D visualization. Their HyperCell concept breaks down n-dimensional data by enabling the user to specify which dimensions should serve as the axes of the cells that are generated. If the user selects a single variable, that variable will become the x-axis and the y-axis will be assigned values as a dependent function. When two variables are selected as axes, the rest of the variables are displayed as contour maps of those two functions. Finally, if the user selects three different variables to represent the axes, the remaining variables will be converted into a 3D function. The process is able to represent all seven dimensions, but does not enable users to fully understand the spatial dimensionality of the data because it separates the dependency of the underlying variables.

Stump et al. [2003] discuss a third way of constructing three-dimensional representations for multidimensional data structures, using glyphs. In their treatment, glyphs are data points in either two-dimensional or three-dimensional space, with tails that represent relations or variables. For instance, the length of the tail might represent the magnitude of a variable and its direction might represent the orientation of the variable. Glyphs effectively visualize multidimensional data by both presenting the correlation between variables within a single data item, and the correlation between multiple data items in the data set. This has direct applicability in the context of registration visualization.

Principal Component Analysis (PCA) is another widely used multidimensional data processing algorithm [Schölkopf et al. 1999]. The PCA algorithm is used to project a multidimensional covariant matrix onto the orthogonal Eigenvectors, principal components that best represent the data. The principal components produce a summary of the data due to this mapping. The components are sorted in a descending manner according to their variance [Yeung and Ruzzo 2001] and lower valued components are discarded. The PCA method serves as a successful filtering technique since it categorizes which variables have the greatest variation and therefore the greatest effect on the solution set. However, since it summarizes the data in question, it is not suitable for the visualization of each solution contained in the data set.

Unfortunately, none of these visualizations (with the exception of Glyphs) adequately respects the structure of the registration solution space. This necessitates the development of a new visualization scheme.

3 Design

Using three-dimensional display without proper justification is frowned on by the visualization community [Cockburn 2004] since it smacks of "pretty" pictures. In many cases two-dimensional methods are more effective.

However, considering the properties of the solution set being visualized, three-dimensional visualization is appropriate. The solution set contains various transformations, which represent the rigid-body rotations and translations necessary to align (or dock) one structure relative to another. By sub-dividing the data, there exists a one-to-one mapping in three dimensions. In this case, a two-dimensional mapping is counterintuitive. Furthermore, Glyphs are applicable in three-dimensions and can be used to represent fitness (and other variables) effectively.

Each solution is portrayed by two spherical Glyphs, in separate 3D views – one for translation and one for rotation (as shown in Figure 2). The fitness is encoded using both colour and deformation. A standard cold-to-hot colouring strategy is adopted; solutions with higher fitness are more red and worse solutions are more blue. The colour scale is normalized according to the range of fitnesses in the solution space. Thus, the best solution will always be red while the worst solution will always be blue.

We follow Shneiderman's mantra: "overview first, zoom and filter, then details on demand" [Shneiderman 1996] by presenting the user with a complete overview of the solution space. The user is able to filter out certain solutions, using one of the predesigned filtering mechanisms (described in Section 3.3), and can explore the space by zooming or rotating, and gain detailed information about a solution by selecting it.

Providing a single mapping to three-dimensional space introduces a bias towards one of the two transformation subsets. Instead, two side-by-side views are implemented. One view represents the translation of solutions (with a subsidiary visual cue indicating rotation); the other view represents the rotations alone. A third view provides the standard numerical list of solutions. These views are integrated in a variety of ways (as will be discussed in Section 3.3) to ensure that the user obtains a complete rather than fragmented understanding.

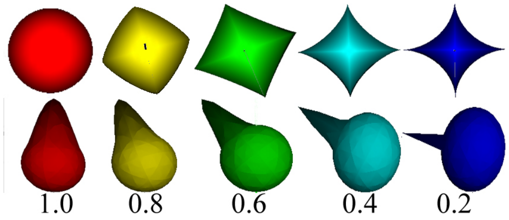


Figure 2: *Colour and shape of glyphs as indicators of fitness. [Top] Orientation glyphs use superquadrics to augment colour, with lower fitness indicated by narrower more stellated shapes and higher fitness by more rounded bulbous shapes. [Bottom] Translation glyphs use deformed extrusions to augment colour, with thicker extrusions linked to greater fitness.*

3.1 Translation View

The translation view focusses on the translations of the solution items. The user gains an understanding of the placement of the solution and its fitness. The visualization also enables users to recognize structures like clustering of solutions, and mirroring of solutions around an arbitrary point.

In this view glyphs are placed at co-ordinate positions that match the x, y, z translation values of the solution. Deformation of a glyph is used to indicate the rotation of the solution. This is done by extruding the surface of the spherical glyph in the direction of the solution's rotation using Simple Constrained Deformation (Scodef) [Borrel and Rappoport 1994]. This deformation scheme allows a set of constrained points to be displaced with the surrounding surface conforming smoothly (according to a B-spline basis). For our purposes, a single displacement constraint is applied in the direction of rotation, while the extent of the surrounding surface undergoing displacement (the radius of effect) is controlled by the fitness of the solution, with fitter solutions implying a larger effect. The actual fitness of the solution is used as the radius of affect. Since the fitness scores lie between zero and one, the resulting deformation reflects the absolute fitness of the solution item, as can be seen in figure 2.

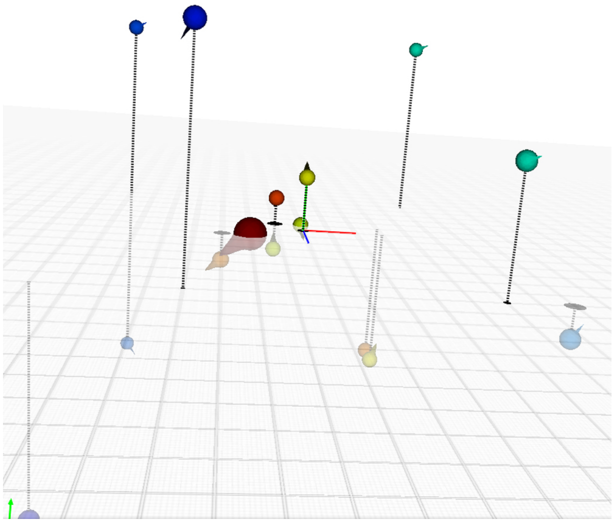


Figure 3: *The translation visualization, with solutions placed according to their x,y,z translations. Solutions are coloured according to fitness and have an extrusion which indicates the rotation direction. The thickness of this extrusion is also used as an indicator of fittness. A semi-transparent ground plane is placed at the level of* $y = 0$ *translation, and height above this ground plane is indicated with dashed drop lines.*

Figure 3 shows the translation view presented to the user. It contains a semi-transparent grid, positioned along the x-z plane, providing users with a sense of space. The user is also able to determine the exact x and z values of each solution item through the unit markers on the grid. The y translation of the solution is indicated by the height the solution is placed from the plane. The x-y and y-z grids are not displayed, thus avoiding cluttering of the view. However, the user can view these planes as well.

3.2 Rotation View

The rotation view, as seen in figure 4, uses the rotation of each solution to determine its position. A vector is constructed, for each solution, by the rotation of that solution. The fitness of the solution dictates the magnitude of the vector. Each vector stretches out from a center-point with the Glyph placed at its end, thus, solutions with higher fitness values will be placed further from the centerpoint than solutions with lower fitness values. A triangular Glyph is added to each vector, indicating the final orientation of the solution item.

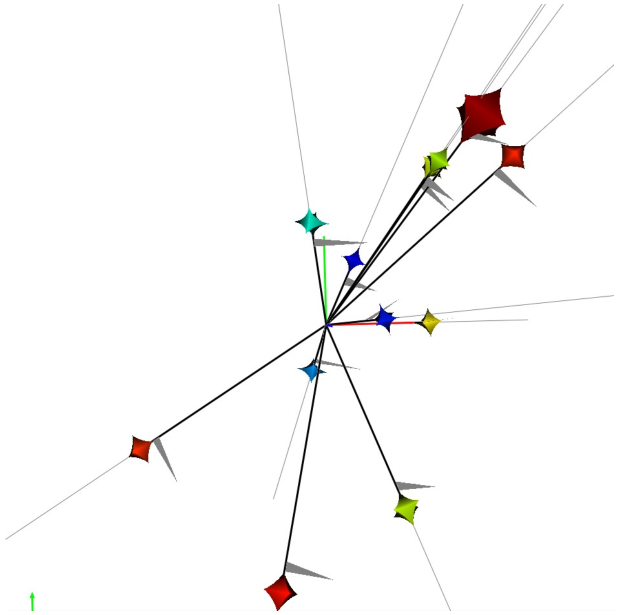


Figure 4: *The rotation window, with rotations extending from a centerpoint. Fitness is indicated by distance from the center, a cool to hot colour scheme and the inflation or deflation of the superquadric glyphs. The flags indicate roll around the direction of the glyph lines and are necessary for a complete specification of orientation.*

The Glyphs representing the solutions, are deformed to indicate the absolute fitness of the solutions, thus reinforcing the link between the two views. Each item is deformed, using superquadrics [Shaw et al. 1999]. This deformation scheme controls the convexity and concavity of the sphere. Figure 2 shows the resulting shapes, ranging from a perfect sphere (fitness of one) to a *diamond* shaped model (fitness of zero).

3.3 View Integration

The final visualization presented to the users is displayed in Figure 5. The side-by-side views are located above the numerical list of solutions, with the translation view on the left and the rotation view on the right. The visualization allows users to select an item in any of the views. The item is highlighted in all the views; items in the translation and rotation views are enlarged, while the solution is selected in the numerical list. The data associated with the item is also presented in a box located between the visualization views and the numerical list.

Various visual cues link the solutions found in the various views. The direction of the translation deformations indicates the associated rotations, but, unfortunately no translation information is presented in the rotation view. The same colour key is applied to the solution items in both windows, therefore, each item is encoded with the same colour across the views.

The visualization allows users to filter the solutions according to translation, rotation, and fitness. The filtering is applied to both views, displaying only solutions that fall within the bounds of the filter. The current translation and rotation filters apply a distance constraint on the selected item, and every other item in the solution space, while the fitness filter includes all solutions that lie within a user-specified range.

	Translation			Rotation			
Solution	x	y	z	x	y	z	Fitness
1	-2	0	1	-140	140	278.1	0.61043
2	-1	1	-1	50	30	179.95	0.57665
3	3	-5	-1	-30	20	196.31	0.5456
4	-3	-1	1	-140	140	278.1	0.54531
5	3	-5	0	60	70	292.42	0.48593
6	0	2	0	-40	140	0	0.48197
7	0	0	-1	10	20	196.31	0.46687
8	-1	-1	-1	150	40	179.95	0.46511
9	5	7	-6	-140	30	224.94	0.26415
10	5	4	8	60	140	16.36	0.26299
11	7	-1	8	20	140	16.36	0.18531
12	-7	-7	-6	-60	30	224.94	0.16823
13	-7	7	-6	-140	20	229.03	0.15737
14	-4	7	8	-60	120	263.93	0.12864
15	-8	-7	8	120	140	16.36	0.11689

Table 1: *Example of a numerical list typically used to evaluate solutions to automated docking.*

Figure 4 also contains two of the high-level properties found in registration spaces. The rotation view depicts the top nine solutions in the space. The solutions form a symmetrical pattern around the x axis, indicating that the molecule has a six fold symmetry. Most of the solutions have also been translated along a straight path, displayed in the translation view.

4 Experimental setup and evaluation

A typical small-scale solution set generated by an automated docking system is shown in Table 1. Currently, the sole means for users to evaluate such numerical results is by visual inspection of the solution set list, with the possible assistance of spreadsheet applications. Thus, we chose to compare our visualization against such a list-based analysis.

A sample of twelve experts in the field of chemistry and structural biology were asked to take part in a longitudinal user experiment. The aim of the study was to evaluate whether users are able to gain a higher level understanding of the solution space, and, whether users find the visualization more useful and intuitive than a list-based method of evaluation.

Based on the work done by Pillat et al. [2005], we determined a set of tasks for our experimental subjects to complete, using either a list-based evaluation or our visualization. Pillat et al. define seven goals that a user should be able to achieve, using a visualization of multidimensional space, namely: identify, determine, compare, infer, locate, visualize and configure. Each of these goals might require the user to complete several low level sub-goals before it can be achieved. In keeping with Pillat et al., subjects were tasked with identifying solutions according to one or more of the quantifying variables (translation, rotation and fitness). Each subject was also asked to compare different solutions and provide an answer to the question relating to that comparison, such as how close together they were in terms of rotation and translation. Subjects needed to spot different items in the list or visualization and extract exact data for that solution (e.g., "Identify the solution that is translated in z only"). Finally, each subject was presented with a hypothesis based on the solution set (such as "Comment on the hypothesis that the best solution relies more heavily on rotation than translation") and was asked to provide reasons that would either support or discredit the validity of the provided hypothesis. The tasks also allowed subjects to explore the solution space and comment on any high level features, such as clustering, mirroring or symmetry, they were able to infer.

The study was longitudinal in that it consisted of two phases: an initial test with relatively small solution spaces (less than twenty solutions) and a later test with larger solution spaces (with 1000 solutions). Each subject took part in both phases and in each phase completed tasks with both a list-based interface and our visualization. In order to compensate for learning effects 4 sets of solutions and associate tasks were created so that subjects would be exposed to different tasks for each interface. Furthermore, both the order of interface presentation (list-based or visualization) and, within each phase, the order of solution sets, was permuted between subjects.

After completing each phase, the subjects were presented with a 5-point Likert-scale questionnaire ($1 =$ to $5 =$) that allowed them to rate both the list-based and visualization interface. A categorization of questions and their mean scores for both list and visualization appear in Table 2. The questions test several of the high level goals outlined by Pillat et al. [2005]. Questions 1 and 2 test the ease with which subjects could identify items and determine attributes associated with the data items. Question 3 aims at measuring the subject's ability to compare different items using the visualization as an aid, while Question 4 tests the subject's ability to infer information such as rules and hypotheses and Question 5 determines the high level understanding gained from the visualization by measuring the subject's ability to locate various structures such as mirroring, clustering and symmetry. We used a two-tailed student t-test to evaluate whether the ratings for each of the questions were significantly different (refer again to Table 2).

5 Results and discussion

Both phases of the test indicates that users found no difference in identifying items according to fitness, between the list and the visualization. This is to be expected since spreadsheets allow users to sort the data items by fitness. In phase one, users indicated that it was easier to identify items by other attributes and to compare these items. However, statistical significance only became apparent in phase two where the solution set was much larger, which can be explained by the fact that users had to sift through a thousand solutions.

An interesting development can be noted in the results of Question 4. In phase one, users found it significantly easier to verify hypotheses with the visualization, while there is no significant difference for this in the second phase. We believe this is due to our program not allowing the user enough freedom to customize the filtering techniques and being overwhelmed by the data. Thus advanced filtering methods would be a fruitful area for future work.

As can be seen from Question 5 of the results table, the most significant result of the user testing is that users found it much easier to obtain a high level understanding of the solution space, which includes noticing structures such as symmetry and clustering. This strongly supports the principle goal of our design.

The results of phase one indicate that the users were less overwhelmed by the information displayed in the visualization (Question 6). However, phase two indicates that the users were as overwhelmed by the data in the visualization as in the list. This can probably be countered through the use of initial clustering, where nearby solutions (in terms of their difference in translation and rotation) are merged into a single representative solution, thus hiding the complexity from the user until such time as it is requested.

The user testing could be extended to larger sample sizes with perhaps more significant results. Since the sample size is less than 20 and since the scores do not necessarily follow a normal distribution, there is the possibility of type-II errors. By cloning the data and performing a t-test on the larger data set, we also found significance for Question 2 and 4, which might serve as an indication that our sample size was too small and thus produced a type-II error.

The results indicate that the visualization enables users to compare items more effectively and gain a high level understanding of the solution space, thus meeting the aim we set out to achieve.

6 Case Study

Here we explore the application of our visualization to a real-world problem, namely the molecular docking of a GroEL subunit into its corresponding electron density map. The GroEL (Bacteriophage growth mutant, restored by mutants in λ head protein gene E, long form) molecule is shown in Figure 1. It plays a critical role in the growth of the *Eschericia Coli* bacteria. The GroEL chaperonin has been studied in detail because of its various properties. It is one of the members of the heat shock regulon and is required for bacterial growth at high temperatures [Fayet et al. 1989]. It has been shown that the GroE (GroEL and GroES) genes are required for bacterial growth across a wide temperature range [Fayet et al. 1989]. The molecule consists of two seven-fold symmetric rings that are docked back to back to form a cylindrical structure [Ranson et al. 2001]. We illustrate the value of our visualization tool using the docking of subunits into GroEL as an example.

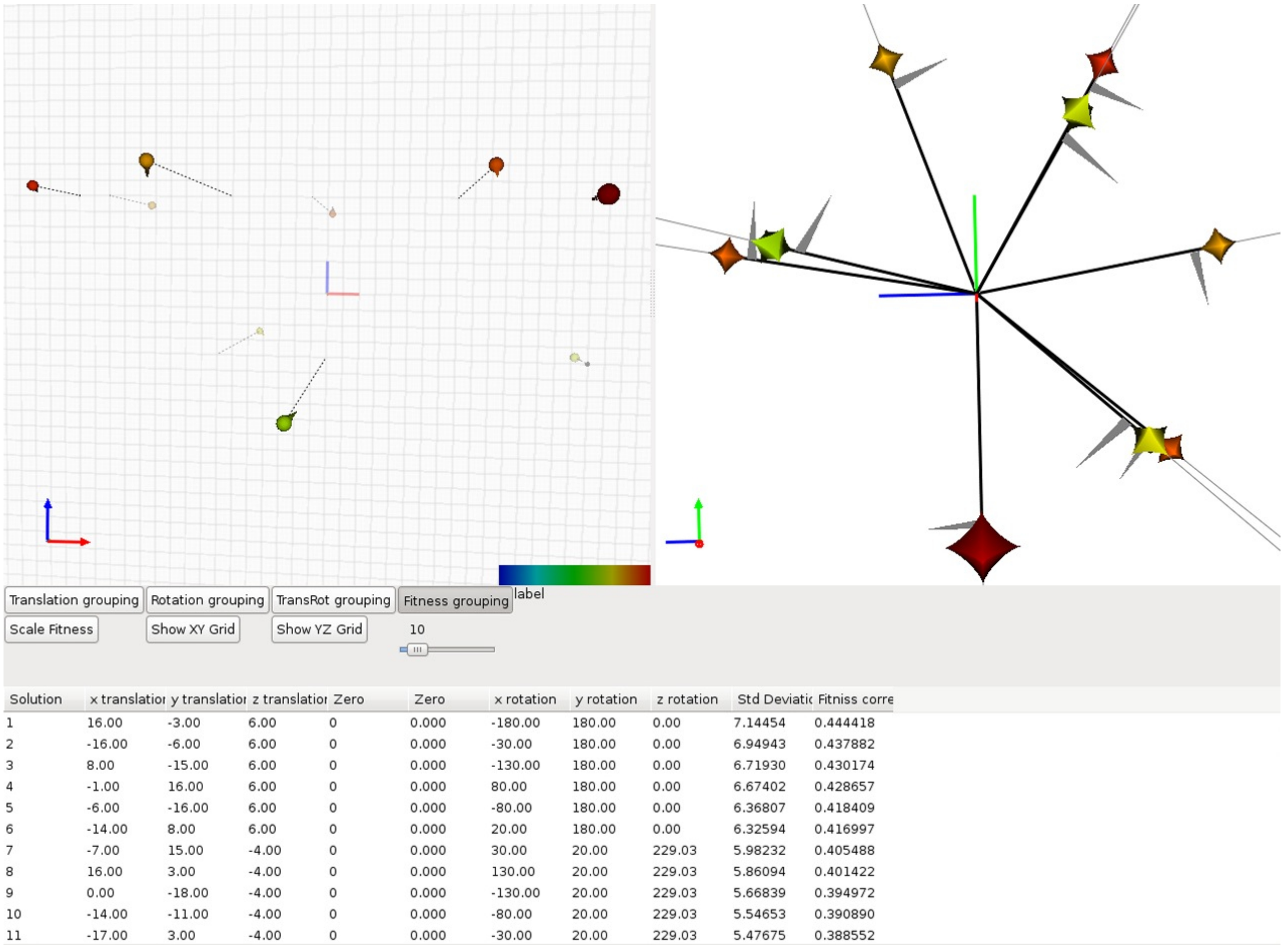
Docking of a single subunit into the GrOEL EM micrograph with DockEM [Roseman 2000] produces a solution set of over 1000 elements. Using our visualization we filter based on fitness and evaluate the top fourteen solutions. Figure 1 provides a reference image of the sub-structure that was docked to produce the solutions that are used for our case study. An electron density map of the GroEL molecule that contains the top ten solutions, is also provided. From the side-on view it is clear that the sub-structure can dock in two separate sections of the map. Furthermore, the top-down view highlights the seven-fold rotational docking symmetry of the structure.

By consulting the numerical listing of solutions found in Table 3 it can be seen that the first six solutions share the same z-translation and that their y- and z-rotations are the same. The same can be said of solutions seven through thirteen, while solution fourteen, once again, shares the same rotation found in the first six solutions. Since we know exactly what molecule is being used in the docking, we can determine that the sub-structure is being docked into two separate parts of the GroEL structure. It is also clear that the first six solutions have rotations of 180 degrees around the y-axis and zero degrees around the z-axis. However, it is less clear what the implication of these rotation values are.

These same conclusions can be arrived at more readily using the dual translation-rotation visualization windows. In particular, the implications of specific rotations are immediately apparent. Based on feedback from the user study, a variety of filtering techniques were implemented. We used these to remove all but the top fourteen solutions, corresponding to Table 3.

Figure 6 depicts the translation view after filtering. From this view, it is easy to spot that seven of the best solutions lie along the same z translation. However, another noticeable result is that another set of seven solutions have been slightly translated along the z-axis but once again follow the same pattern of translation.

Turning to the rotation view of the visualization: Figure 7 shows the initial filtered rotation window. From this filtered view it can be seen that the top solutions are rotated along an arbitrary axis, pointing in seven different directions. Examining the fins on each rotation, it can be seen that the rotations are all oriented in approxi-



Solution	x translatior	y translatior	z translatior	Zero	Zero	x rotation	y rotation	z rotation	Std Deviatic	Fitniss corre
1	16.00	-3.00	6.00	0	0.000	-180.00	180.00	0.00	7.14454	0.444418
2	-16.00	-6.00	6.00	0	0.000	-30.00	180.00	0.00	6.94943	0.437882
3	8.00	-15.00	6.00	0	0.000	-130.00	180.00	0.00	6.71930	0.430174
4	-1.00	16.00	6.00	0	0.000	80.00	180.00	0.00	6.67402	0.428657
5	-6.00	-16.00	6.00	0	0.000	-80.00	180.00	0.00	6.36807	0.418409
6	-14.00	8.00	6.00	0	0.000	20.00	180.00	0.00	6.32594	0.416997
7	-7.00	15.00	-4.00	0	0.000	30.00	20.00	229.03	5.98232	0.405488
8	16.00	3.00	-4.00	0	0.000	130.00	20.00	229.03	5.86094	0.401422
9	0.00	-18.00	-4.00	0	0.000	-130.00	20.00	229.03	5.66839	0.394972
10	-14.00	-11.00	-4.00	0	0.000	-80.00	20.00	229.03	5.54653	0.390890
11	-17.00	3.00	-4.00	0	0.000	-30.00	20.00	229.03	5.47675	0.388552

Figure 5: *Snapshot of the final solution set visualization program evaluated.*

mately the same direction. By scaling the solution set (as illustrated in Figure 8) the solutions can be spaced further apart because the full solution space is stretched to a solution space of zero-to-one. This filtering and scaling indicates that the substructure has probably been docked into a molecule with seven-fold symmetry. This accords with structural knowledge of the GroEL molecule.

In summary, we find that our visualization, particularly when combined with filtering and scaling, allows effective analysis of docking solutions, even in complex cases.

7 Conclusions

We have implemented a novel approach to visualization of seven-dimensional solution sets generated by automated structure docking, using three-dimensional glyphs in two viewing windows. These windows focus attention on the relative translation and orientation of the solutions, respectively, and are linked through various visual cues. A user study shows that we achieved a visualization that enables users to explore the complete set of solutions, compare items more readily and infer structural detail such as symmetry, leading to a high-level appreciation of the solution space important in determining whether the docking procedure has been correct.

Future development of the visualization includes implementing initial clustering of the solutions. This will prevent users from being overwhelmed by the visualization, and allow them to expand regions of interest. The visualization does not currently allow users to view the actual docked structures, and this needs to be addressed as well.

References

AMAR, R., AND STASKO, J. 2004. Best paper: A knowledge task-based framework for design and evaluation of information visualizations. In *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization*, IEEE Computer Society, Washington, DC, USA, 143–150.

ASSA, J., COHEN-OR, D., AND MILO, T. 1997. Displaying data in multidemsional relevance space with 2d visualization maps. *Visualization Conference, IEEE 0*, 127.

BAKER, T. S., AND JOHNSON, J. E. 1996. Low resolution meets high: towards a resolution continuum from cells to atoms. *Current Opinion in Structural Biology 6*, 5, 585–594.

BELLAMY, R. K. E., ERICKSON, T., FULLER, B., KELLOGG, W. A., ROSENBAUM, R., THOMAS, J. C., AND WOLF, T. V. 2007. Seeing is believing: designing visualizations for managing risk and compliance. *IBM Syst. J. 46*, 2, 205–218.

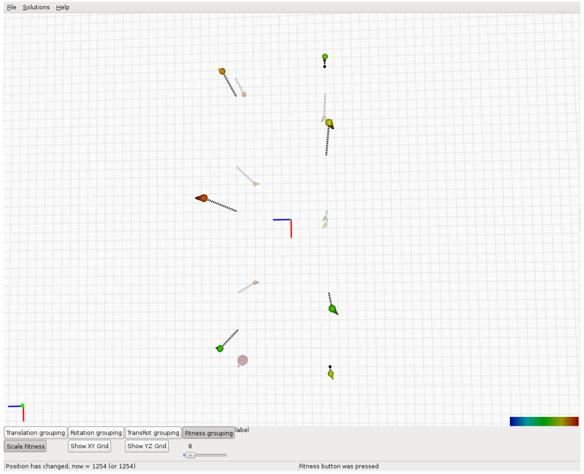
BORREL, P., AND RAPPOPORT, A. 1994. Simple constrained deformations for geometric modeling and interactive design. *ACM Trans. Graph. 13*, 2, 137–155.

Question	Phase1 List	 Visualization	 T-Test	Phase2 Spreadsheet	 Visualization	 T-Test
1. Identification by fitness	4.27	4.18	0.86	4.10	4.00	0.88
2. Identification by translation or orientation.	2.90	3.63	0.18	2.50	4.10	**0.004**
3. Difficulty of comparing items	3.45	2.72	0.26	3.70	2.40	**0.01**
4. Ease of supporting or challenging a hypothesis	2.54	3.36	0.09	3.00	3.60	0.26
5. High level understanding	1.45	4.45	**1.15e-7**	1.60	4.10	**3.66e-7**
6. Overwhelmed with information.	3.45	2.36	**0.03**	3.10	2.80	0.63

Table 2: *Tabulated results from the two phases of the user experiment. Means from a* 5*-point Likert scale are reported for both a list-based and visualization interface. The rightmost column for each phase shows a student t-test for significant difference with significant entries at* $p < 0.05$ *marked in bold.*

Solution	Translation X	 Y	 Z	Rotation X	 Y	 Z	Fitness
1	16.00	-3.00	6.00	-180.00	180.00	0.00	0.444418
2	-16.00	-6.00	6.00	-30.00	180.00	0.00	0.437882
3	8.00	-15.00	6.00	-130.00	180.00	0.00	0.430174
4	-1.00	16.00	6.00	80.00	180.00	0.00	0.428657
5	-6.00	-16.00	6.00	-80.00	180.00	0.00	0.418409
6	-14.00	8.00	6.00	20.00	180.00	0.00	0.416997
7	-7.00	15.00	-4.00	30.00	20.00	229.03	0.405488
8	16.00	3.00	-4.00	130.00	20.00	229.03	0.401422
9	0.00	-18.00	-4.00	-130.00	20.00	229.03	0.394972
10	-14.00	-11.00	-4.00	-80.00	20.00	229.03	0.390890
11	-17.00	3.00	-4.00	-30.00	20.00	229.03	0.388552
12	-1.00	-18.00	-4.00	-130.00	20.00	229.03	0.382164
13	8.00	14.00	-4.00	80.00	20.00	229.03	0.378775
14	12.00	10.00	6.00	130.00	180.00	0.00	0.376549

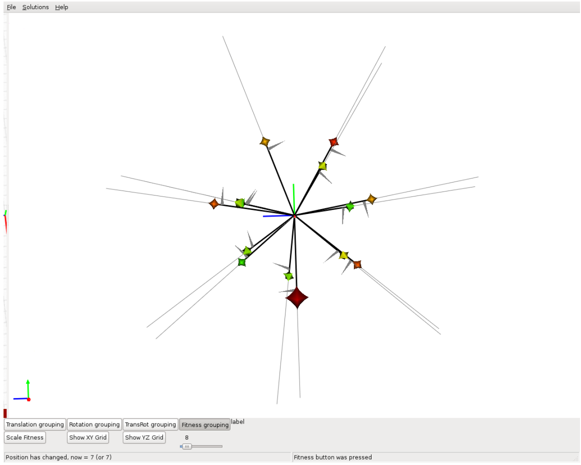
Table 3: *Fourteen fittest numerical solutions of the GroEL docking procedure.*



Position has changed, now = 1254 (or 1254)

Fitness button was pressed

Figure 6: *GroEL Translation View. A filter was applied to remove all but the top fourteen solutions. The parallel clustering of solutions along two lines is immediately apparent.*



Position has changed, now = 7 (or 7)

Fitness button was pressed

Figure 7: *GroEL Rotation View. Seven-fold symmetry is apparent from the grouping of solution into seven clusters.*

BUJA, A., COOK, D., AND SWAYNE, D. F. 1996. Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics 5*, 78–99.

COCKBURN, A. 2004. Revisiting 2d vs 3d implications on spatial memory. In *AUIC '04: Proceedings of the fifth conference on Australasian user interface*, Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 25–31.

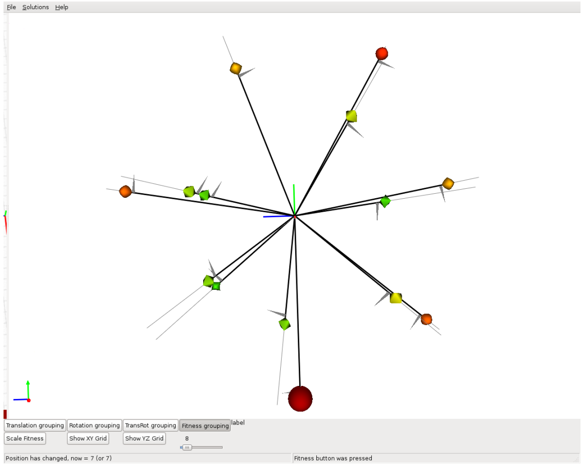
DOS SANTOS, S. R., AND BRODLIE, K. W. 2002. Visualizing and investigating multidimensional functions. *Proceedings of the symposium on data visualization*, 173–ff.

ELLIS, G., AND DIX, A. 2007. A taxonomy of clutter reduction for information visualisation. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (Nov.-Dec.), 1216–1223.

FAYET, O., ZIEGELHOFFER, T., AND GEORGOPOULOS, C. 1989. The groes and groel heat shock gene products of escherichia coli are essential for bacterial growth at all temperatures. *Journal of Bacteriology 107*, 3, 1379–1385.

FEINER, S., AND BESHERS, C. 1999. Worlds within worlds: Metaphors for exploreing n-dimensional virtual worlds. *Readings in Information Visualization: Using vision to think*, 96–103.

HIBBARD, B. 2003. Visbio: a biological tool for visualization and



Position has changed, now = 7 (or 7)

Fitness button was pressed

Figure 8: *GroEL Rotation View, with normalised scaling. Normalising the rotation fitness to the range [0, 1] emphasises disparities among solutions.*

analysis. *SIGGRAPH Comput. Graph. 37*, 2, 5–7.

INSELBERG, A. 1999. Multidimensional detective. *Readings in Information Visualization: Using vision to think*, 107–114.

INSELBURG, A. 1985. The plane with parallel coordinates. *The Visual Computer 1*, 2, 69–91.

KEIM, D. A. 2002. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics 8*, 1, 1–8.

LASKER, K., DROR, O., SHATSKY, M., NUSSINOV, R., AND WOLFSON, H. J. 2007. Ematch: Discovery of high resolution structural homologues of protein domains in intermediate resolution cryo-em maps. *IEEE/ACM Trans. Comput. Biol. Bioinformatics 4*, 1, 28–39.

MOLL, A., HILDEBRANDT, A., LENHOF, H., AND KOHLBACHER, O. 2006. Ballview: a tool for research and education in molecular modeling. *Bioinformatics 22*, 3, 365–366.

PETTERSEN, E., GOADDARD, T., HUANG, C., COUCH, G., GREENBLATT, D., MENG, E., AND FERRIN, T. 2004. Ucsf chimera-a visualization system for exploratory research and analysis. *J. Comput. Chem. 25*, 13, 1605–1612.

PILLAT, R. M., VALIATI, E. R. A., AND FREITAS, C. M. D. S. 2005. Experimental study on evaluation of multidimensional information visualization techniques. In *CLIHC '05: Proceedings of the 2005 Latin American conference on Human-computer interaction*, ACM, New York, NY, USA, 20–30.

PINTILIE, G. D., TUEKAM, B., AND HOGUE, C. W. V. 2005. Generation of glyphs for conveying complex information with application to protein representations. *Lecture notes in computer science*, 90–102.

PLUIM, J., MAINTZ, J., AND VIERGEVER, M. 2000. Image registration by maximization of combined mutual information and gradient information. *Medical Imaging, IEEE Transactions on 19*, 8 (Aug.), 809–814.

POST, F. H., POST, F. J., WALSUM, T. V., AND SILVER, D. 1995. Iconic techniques for feature visualization. In *VIS '95: Proceedings of the 6th conference on Visualization '95*, IEEE Computer Society, Washington, DC, USA, 288.

RANSON, N., FARR, G., ROSEMAN, A., GOWEN, B., FENTON, W., HORWICH, A., AND SAIBIL, H. 2001. Atp-bound states of groel captured by cryo-electron microscopy. *Cell 107*, 869–879.

ROSEMAN, A. M. 2000. Docking structures of domains into maps from cryo-electron microscopy using local correlation. *ACTA Crystallographica 56*, 1332–1340.

ROTH, S. F., AND MATTIS, J. 1990. Data characterization for intelligent graphics presentation. In *CHI '90: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, USA, 193–200.

SCHÖLKOPF, B., SMOLA, A. J., AND MÜLLER, K.-R. 1999. Kernel principal component analysis. *Advances in kernel methods: support vector learning*, 327–352.

SHAW, C. D., HALL, J. A., BLAHUT, C., EBERT, D. S., AND ROBERTS, D. A. 1999. Using shape to visualize multivariate data. In *NPIVM '99: Proceedings of the 1999 workshop on new paradigms in information visualization and manipulation in conjunction with the eighth ACM internation conference on Information and knowledge management*, ACM, New York, NY, USA, 17–20.

SHNEIDERMAN, B. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *VL '96: Proceedings of the 1996 IEEE Symposium on Visual Languages*, IEEE Computer Society, Washington, DC, USA, 336.

SIGFRIDSSON, A., EBBERS, T., HEIBERG, E., AND WIGSTRÖM, L. 2002. Tensor field visualisation using adaptive filtering of noise fields combined with glyph rendering. In *Vis '02: Proceedings of the conference on Visualization '02*, IEEE Computer Society, Washington, DC, USA, 371–378.

SPEARS, W. M. 1999. An overview of multidimensional visualization techniques. In *Evolutionary Computation Visualiztion*, Morgan Kaufmann, 104–105.

STUMP, G. M., YUKISH, M., SIMPSON, T. W., AND HARRIS, E. N. 2003. Design space visualization and its applications to a design by shopping paradigm. *ASME 2003 design engineering technical conferences and computer and information in engineering conference.*.

VAN WIJK, J. J. 2006. Views on visualization. *IEEE Transactions on Visualization and Computer Graphics 12*, 4, 1000–433.

WALTER, J. A., AND RITTER, H. 2002. On interactive visualization of high-dimensional data using the hyperbolic plane. ACM, New York, NY, USA.

WARD, M. O. 2002. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization 1*, 3/4, 194–210.

WEHREND, S., AND LEWIS, C. 1990. A problem-oriented classification of visualization techniques. In *VIS '90: Proceedings of the 1st conference on Visualization '90*, IEEE Computer Society Press, Los Alamitos, CA, USA, 139–143.

WONG, P. C., AND BERGERON, R. D. 1997. Multivariate visualization using metric scaling. In *VIS '97: Proceedings of the 8th conference on Visualization '97*, IEEE Computer Society Press, Los Alamitos, CA, USA, 111–ff.

WRIGGERS, W., MILLIGAN, R., AND MCCAMMON, A., 1999. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy.

YEUNG, K., AND RUZZO, W. 2001. Principal component analysis for clustering gene expression data. *Bioinformatics 17*, 9, 763–774.

ZITOV, B., AND FLUSSER, J. 2003. Image registration methods: a survey. *Image and Vision Computing 21*, 11, 977 – 1000.