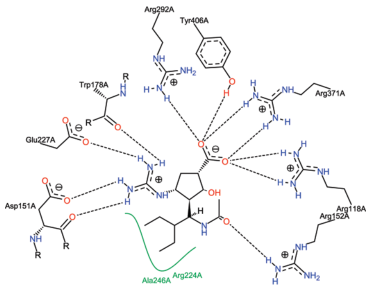
ACS Medicinal Chemistry Letters
TECHNOLOGY NOTE
pubs.acs.org/acsmedchemlett

Drawing the PDB: Protein−Ligand Complexes in Two Dimensions

Katrin Stierand and Matthias Rarey*

Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

ABSTRACT The two-dimensional representation of molecules is a popular communication medium in chemistry and the associated scientific fields. Computational methods for drawing small molecules with and without manual investigation are well-established and widely spread in terms of numerous software tools. Concerning the planar depiction of molecular complexes, there is considerably less choice. We developed the software PoseView, which automatically generates two-dimensional diagrams of macromolecular complexes, showing the ligand, the interactions, and the interacting residues. All depicted molecules are drawn on an atomic level as structure diagrams; thus, the output plots are clearly structured and easily readable for the scientist. We tested the performance of PoseView in a large-scale application on nearly all druglike complexes of the PDB (approximately 200000 complexes); for more than 92% of the complexes considered for drawing, a layout could be computed. In the following, we will present the results of this application study.



KEYWORDS molecular visualization, PDB, protein−ligand complexes, structure diagram, two dimensions

Because of the growing portfolio of computational drug design methods, medicinal chemists are confronted with large data amounts in a short time. During a drug design process, the molecular candidate selection after each working step is crucial for the overall result. Pipelining tools like Pipeline Pilot[1] or KNIME[2] offer a resource and time optimization by automating parts of the workflow. However, the visual investigation of the results by scientific experts cannot be completely replaced. Consequently, there is need for clearly arranged and informative visualization of output data. The visualized data can be presented in either a three-dimensional (3D) or a two-dimensional (2D) layout. While three dimensions offer a large richness of detail, the visual exploration is often more time-consuming, and the user needs some practice to handle the tools. In comparison, the information content of 2D layouts is more limited, which allows a focus on only a subset of attributes but offers the possibility to scan quickly over a lot of data. Two-dimensional visualization includes the color-coded comparison of special numerical values (e.g., heat maps[3]) as well as the visualization of multimolecular relationships like metabolic networks[4] or molecule depiction on atomic level, for example, by means of structure diagrams.[5]

Three-dimensional visualization of ligands in complex with macromolecules and the resulting interaction pattern are widely used and available in many different applications, whereas the corresponding 2D field offers only three different software tools.[6] One of them is Ligplot,[7] which was the first published approach to visualize hydrophilic interactions, metal interactions, and hydrophobic contacts between ligands and proteins. The residues bound to the ligand by directed interactions are, like the ligand itself, drawn in atomic detail, but the protonation states of all compounds are omitted. The algorithm transfers the bond angles and bond lengths of the 3D complex to the 2D plot as far as possible. A more recent development is embedded in MOE.[8] Beyond the interactions shown in Ligplot, π interactions are included in the complex visualization. While the ligand is represented as a structure diagram, the interaction partners are drawn at the residue level as labeled and colored disks. During the past few years, we developed and implemented PoseView,[9,10] which automatically generates 2D diagrams of molecular complexes with a focus on the interaction network between the complex partners. The aim is to compute collision-free layouts by representing the interacting molecules on the atomic level following the IUPAC recommendations[11] for the depiction of structure diagrams. By drawing the individual molecules as structure diagrams, the learning effort for a medicinal chemist to use PoseView is negligible.

PoseView is able to draw diagrams of complexes consisting of a small molecule and a receptor molecule that can be either a protein or a DNA/RNA. If not user-specified, interactions between the complex partners are estimated using simple geometric criteria, such as distances and angles. PoseView considers five different interaction types. Four of

Received Date: July 8, 2010
Accepted Date: August 24, 2010
Published on Web Date: August 31, 2010

© 2010 American Chemical Society

them are directed interactions: hydrogen bonds, metal interactions, π−cation interactions, and π−π stacking. The fifth interaction type is the undirected hydrophobic contact.

We tested the output quality and performance of our tool on a large set of complexes provided by the RCSB Protein Data Bank[12] (PDB). PoseView was able to generate about 80% collision-free layouts with an average computing time of 0.09 s per complex. The other 20% could not be drawn completely collision-free, either because of the reduction of dimensions from the original 3D coordinates to the 2D visualization or because of the accumulation of many interaction atoms at a small part of the ligand. For these cases, the computing time averages out at about 10 s per complex. In the following, we will present the formerly mentioned interaction model and the results of the PDB visualization in detail.

Besides the simple visualization of given data, PoseView is able to roughly estimate interactions between a ligand and its receptor based on the 3D coordinates given by the input files of the two molecules. Therefore, it is essential that the ligand coordinates are assigned relative to the active site of the receptor as it can be found in proteins with cocrystallized ligands; that is, PoseView does not perform a docking calculation before generating the sketches. Depending on type, valence, and hybridization of a ligand atom, the surrounding receptor atoms are scanned to find potential interaction partners. For the different available interaction types—hydrogen bonds, metal interactions, hydrophobic contacts, π−π stacking, and π−cation interactions—different geometric criteria have to be fulfilled. Figure 1 gives a graphical representation of the five interaction types. While ligand file formats often contain a specification of the protonation state and the coordinates of hydrogen atoms, commonly used formats for macromolecular structures like the PDB format provide no such information. In these cases, ProToss,[13] a program to place polar hydrogen atoms in protein−ligand complexes, is called preliminary to the interaction estimation to protonate and adjust the active site residues automatically.

Hydrogen bonds are implemented mainly following the measures published by Desiraju and Steiner in 2001.[14] The optimal distance between two atoms connected by a hydrogen bond is set to 1.9 Å with a tolerance of 0.5 Å. Additional to this measure, the acceptor−hydrogen−donor angle must not fall below 120°. Hydrogen atoms that are bound to a noncarbon atom are treated as hydrogen donor candidates. Potential acceptor atoms are either nitrogen, oxygen, or sulfur atoms provided that they are uncharged or negatively charged and that their surface is accessible.

Metal interactions are calculated between metal atoms embedded in the receptor and metal acceptor atoms, which are identical to the hydrogen bond acceptor atoms. Their geometry is based on the calculated coordination geometry of the metal.[15] Each coordination point that is not occupied by a receptor atom is checked for close ligand atoms, and the maximal distance deviation is set to 0.8 Å. If no geometry can be calculated, a sphere with a radius of 2 Å is placed around the metal. In this case, all atoms lying on the sphere, again with a tolerance of 0.8 Å, are regarded as potential interaction partners.

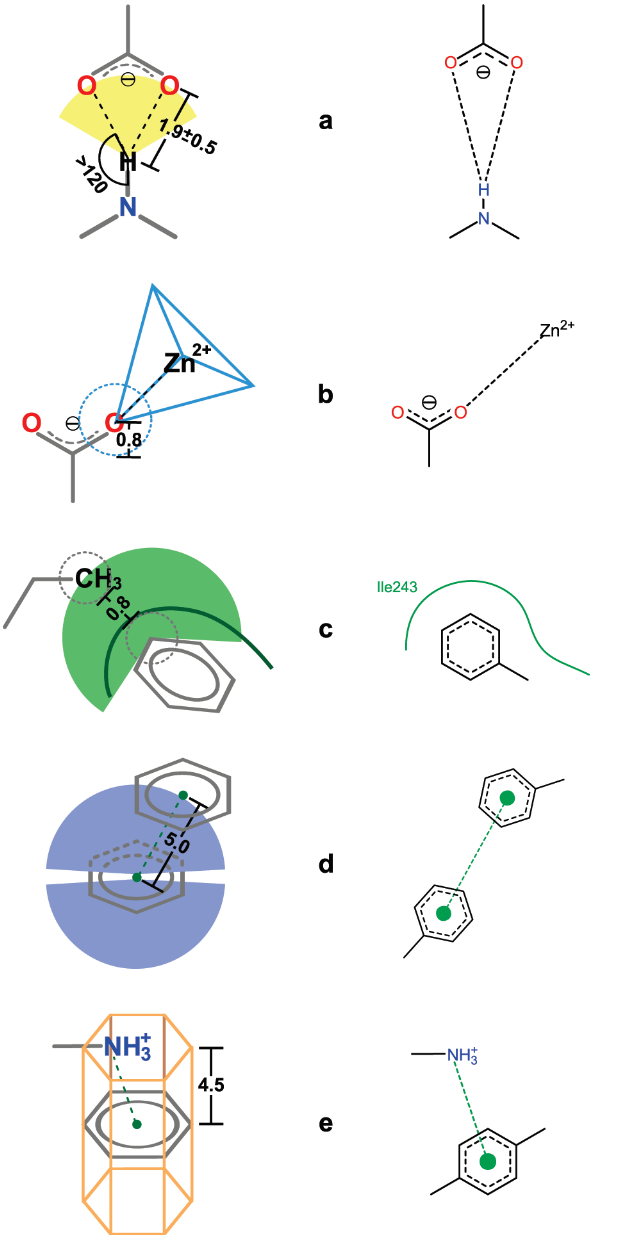


Figure 1. Five available interaction types in PoseView. The sketches on the left-hand side are labeled with the geometric criteria of the interaction model, and on the right-hand side, the corresponding PoseView layout is depicted. A detailed description of each interaction type is given in the text. Each row shows one type: (a) hydrogen bond, (b) metal interaction, (c) hydrophobic contact, (d) π−π stacking, and (e) π−cation interaction.

In contrast to the formerly mentioned interactions, hydrophobic contacts are estimated based on the distance between two hydrophobic atoms only. They are visualized not by a dashed interaction line but by drawing the label of the contacting residue and a spline segment denoting the hydrophobic part of the ligand. Because many atoms typically form a hydrophobic subpocket, this representation reflects the interaction geometry better. A prerequisite for a hydrophobic contact is that at least three hydrophobic ligand

© 2010 American Chemical Society

atoms lie in the range of the currently examined receptor residue. Hydrophobic atoms are in this context carbon atoms with accessible surface and halogens. The maximum distance is set to the sum of the van der Waals radii of atoms in question and a tolerance of 0.8 Å.

According to their significance in drug design, the computation of π interactions including $\pi-\pi$ stacking and $\pi-$cation interactions was added to the model. The parameters for these types are derived from the publications of McGaughey et al.[16] and Gallivan and Dougherty.[17] In contrast to the other directed interactions, π interactions are formed between molecular substructures like phenyl rings rather than single atoms. Hence, the determination of π interactions starts with the identification of aromatic systems, which are a part of the ligand or the active site. Aromatic ring systems are defined as planar ring systems, the atom set of which follows Hückel's rule.[18] Aromaticity is a common feature in the considered receptor residues: Four of the 20 amino acids have side chains containing aromatic rings, and both pyrimidines and purines are dominated by an aromatic ring system. Moreover, drug molecules frequently contain aromatic systems due to their conformational stability. As reference values for π interaction determination, both the centroid and the surface normal of the plane spanned by the selected set of aromatic atoms are calculated.

Although McGaughey et al.[16] described a preference of very small angles between the two corresponding surface normals, the identification of a $\pi-\pi$ stacking interaction in PoseView depends only on the distance of the centroids because all angles up to 90° show lower but nearly constant numbers of occurrences in the examined ring system pairs. The angle between the centroid connection line and one of the normals shows no clear trend; this leads to a very simple model where the maximum distance between the two centroids is set to 5 Å.

$\pi-$Cation interactions are computed depending on two criteria: First, the distance between an aromatic system centroid and a cation must be less than 4.5 Å. To describe the relative orientation, a right prism is created with a base face shape identical to the polygon defined by the atom coordinates of the aromatic system. Provided that the cation is situated inside the prism, a $\pi-$cation interaction is assigned to the complex.

When drawing the PDB, the performance of PoseView and the layout quality of the diagrams highly depend on the number of directed interactions and the number of receptor residues forming at least one directed interaction to the ligand. Hydrophobic contacts have no noticeable influence on quality or computing time due to their representation as labels and the lack of a connecting line to the ligand.

As of June 8, 2010, 65802 structures were contained in the PDB. For many of these structures, one or more cocrystallized small molecules are provided in the Ligand Expo database,[19] which was developed as part of the RCSB PDB project and contains about 700000 ligands. The macromolecular structures are outnumbered by small molecules due to the existence of multiple binding sites, cocrystallized solvent molecules, metal ions, and the occurrence of multiple coordinate models for one ligand. The primary source of information stored in the Ligand Expo database is the wwPDB Chemical Component Dictionary.[20]

Table 1. Filter Criteria of the PoseView Input

filtering step	number of remaining ligands
all ligands from LigandExpo	691417
ligands with less than five or more than 80 atoms	383412
application of exclusion list (PoseView website)	213907
corresponding PDB file not found	212759
large receptor molecule (> 50000 heavy atoms)	201245

To use the ligands as PoseView input, we downloaded the SD files of all small molecules, which provide atom and bond types and experimental coordinates, from Ligand Expo (Version 1, downloaded on April 29, 2010). In preparation for the diagram generation, we filtered the ligand set. Molecules with less than five atoms including hydrogens were excluded because most of them are metals or solvent molecules. Apart from counting atoms, we applied an exclusion list that consists mainly of metal-containing compounds like iron sulfur clusters or heme and small molecules whose existence in the files originates from the protein crystallization process (e.g., ethanol, ammonium, and sulfate ions); the list can be found on the PoseView Web site at http://poseview.zbh.uni-hamburg.de. In the case of multiple models for one ligand, all models were included in the test set since the difference of coordinates could lead to a difference in interaction patterns. The ligand-filtering process reduced the input size to approximately 210000 complexes. Prior to the diagram computation, an analysis of the receptor files pertaining to the remaining ligands was performed, which led to a further reduction to 201000 input complexes. This analysis considered the existence of the receptor file and the size of the macromolecule, which did not have to exceed 50000 heavy atoms. For detailed figures on the filtering process, see Table 1.

In over 85% of the remaining complexes, PoseView succeeds by generating a plot. The main reason for the absence of an output diagram—in about 32500 of the cases—is the lack of interactions between ligand and receptor. To reduce the overall computing time, complexes with more than 18 directed interactions or more than 14 amino acids were omitted (~900 complexes). For roughly 1000 additional diagrams, the algorithm was aborted at a time of 450 s. As a consequence of different technical reasons, 11000 files remained empty. Possible reasons are difficulties in drawing the ligand structure containing bridged ring systems or macrocycles, uncommonly formatted PDB files in which no protein or RNA/DNA is defined, ligand locations outside the protein preventing the definition of an active site, etc.

The resulting diagrams are subdivided into three quality categories (see Figure 2) focusing on collisions between the different diagram components, which can be either structure diagrams of the ligand or the interacting receptor residues, dashed lines visualizing the directed interactions, spline segments denoting hydrophobic contact areas of the ligand, or the residue labels. An example diagram that contains all possible

© 2010 American Chemical Society

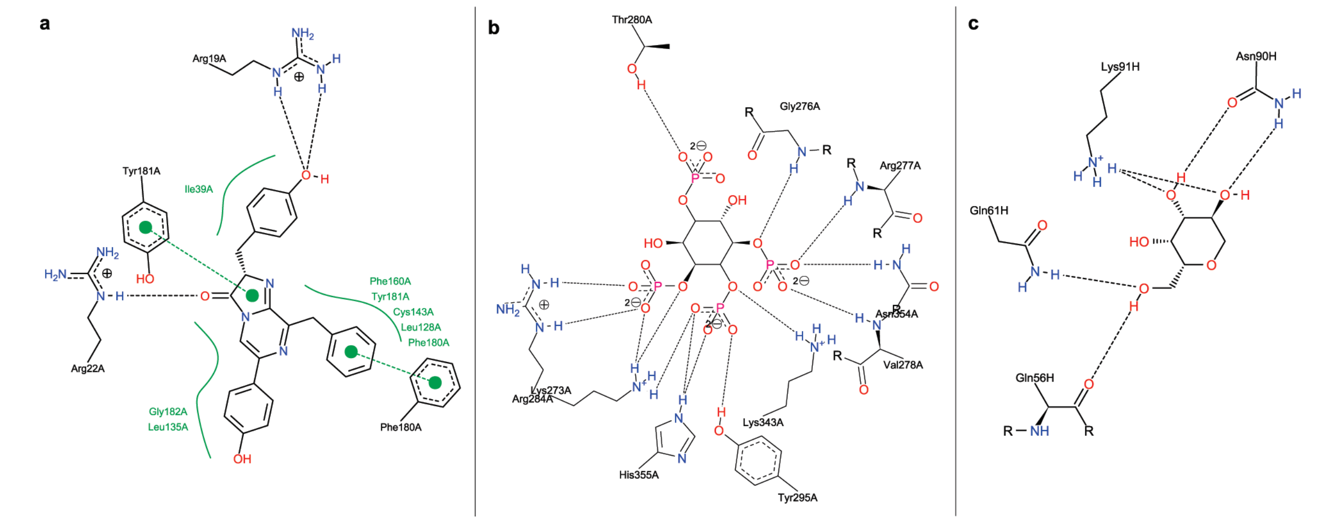


Figure 2. Examples for the different layout qualities. (a) A good layout was computed for the complex of coelenterazine-binding protein and C2-hydroxy-coelenterazine (PDB code: 2HPS[21]), while in panel b, the diagram of a Pleckstrin homology domain in complex with inositol 1,3,4,5-tetrakisphosphate (PDB code: 1FHX[22]) is of improvable quality. The labels of Lys233A and Arg284A collide, and there is an overlap of the structure diagrams of Val278A and Asn354A. (c) In the case of heat labile enterotoxin type I in complex with β-D-galactose (PDB code: 1LTI[23]), no collision free diagram could be computed due to the order of interaction atoms.

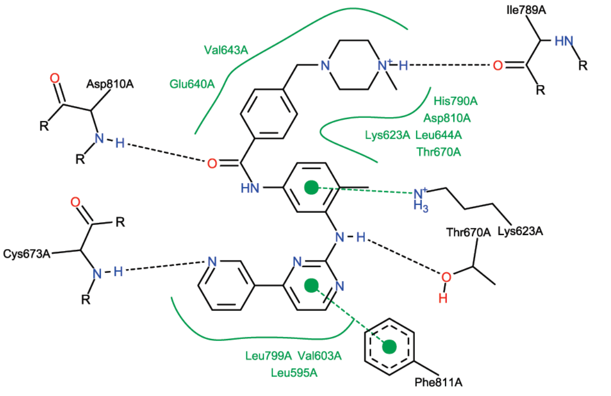


Figure 3. Diagram of *Homo sapiens* v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homologue in complex with imatinib (PDB code: 1T46[24]). The interaction pattern is composed of hydrogen bonds, visualized as black dashed lines; π interactions, shown as green dashed lines with dots denoting the participating π systems; and hydrophobic contacts, which are represented by the residue labels and spline segments along the contacting hydrophobic ligand parts.

components is shown in Figure 3. A good layout quality is characterized by a collision-free arrangement of all components. In some cases, the arrangement of interaction atoms of the ligand provides the possibility of a collision-free arrangement that was not found by the PoseView algorithm. These layouts are referred to as improvable ones. In the third category, the components cannot be arranged in a collision-free manner due to the reduction from three to two dimensions. A description of the underlying algorithm, which modifies the interaction atom arrangement and which is able to find and quantify collisions, can be found in former publications[9,10] concerning PoseView. For almost 80 % of all drawn diagrams, a good layout could be computed, while 17 % are of improvable quality; the remaining 3 % suffer from unsolvable collisions (see Table 2). The number of interactions and the different layout qualities are not evenly distributed over the set of complexes: Two-thirds of all complexes are characterized by less than five directed interactions between ligand and receptor; the number of complexes converges asymptotically to zero with a growing number of interactions (see Figure 4a). Figure 4b shows the layout quality ratio for the different numbers of interactions. While the percentage of good layouts decreases with growing numbers of interactions, the improvable layout fraction has a maximum at 14 interactions, and it is exceeded by the unsolvable layouts at 18 interactions. This substantiates the increasing complexity of the layout optimization problem for highly cross-linked complexes. The computing time (Figure 4a) reflects the dependency on the input size as well.

Table 2. Results of the PoseView Test Run on 201245 from the PDB, Selected by the Preceding Filter Steps (Table 1)[a]

PoseView computing results			count
calculated diagrams			155612
−good layouts	123535	79.4%	
−improvable layouts	26624	17.1%	
−layouts with unsolvable collisions	5453	3.5%	
no interactions between ligand and receptor			32549
more than 18 directed interactions			897
not calculated due to technichal reasons			11149
computing timeout (>450 s)			1038

[a] Additionally, the number of all drawn diagrams is subdivided in the three different quality categories.

In summary, PoseView offers the opportunity to facilitate the evaluation and communication of molecular complex information. We applied PoseView to visualize most of the complexes available in the PDB; over 90 % of the calculated complex diagrams contain less than 11 directed interactions

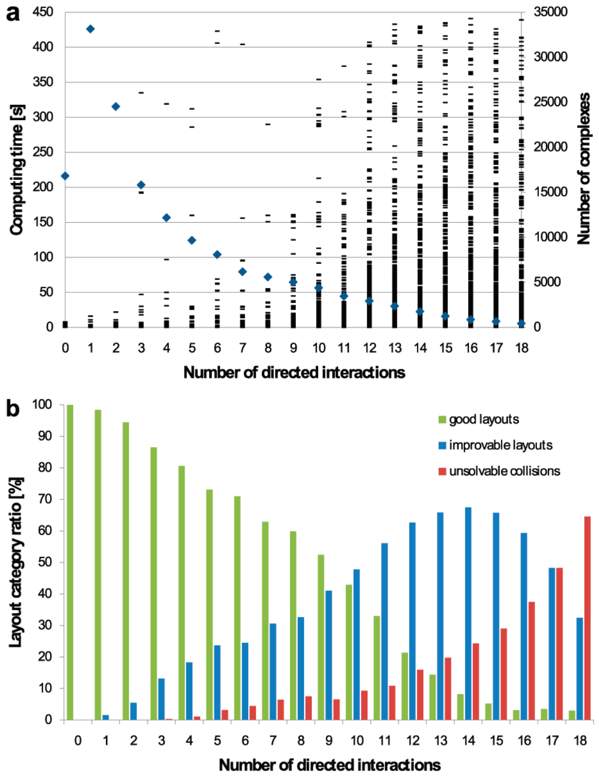


Figure 4. Graphical analysis of the PoseView results. In panel a, the computing times are denoted by black dashes, and the number of complexes are denoted by blue diamonds. Panel b shows the ratio of layout qualities for each occurring number of directed interactions.

and could be calculated in the range of milliseconds to seconds. For 80 % of the drawn diagrams, collision-free layouts could be computed. The remaining set featured collisions, which could be partly resolved by introducing appropriate upgrades to the existing methods.

Because of the illustration of ligand and receptor molecules as structure diagrams, the information of the complex interaction pattern is easily ascertainable by the scientist. In a time where an increasing amount of structural data is available because of internal or Internet resources, we believe that automated 2D complex diagram generation is an important application, making medicinal chemists' daily lives easier.

PoseView can be used free of charge as a webservice at http://poseview.zbh.uni-hamburg.de. It is also available as a licensed standalone version at http://www.biosolveit.de/PoseView. Moreover, the next release of LeadIT (R.1.2.0, http://www.biosolveit.de/LeadIT/) will embed PoseView to complement its 3D visualization of protein—ligand complexes.

AUTHOR INFORMATION

Corresponding Author: *To whom correspondence should be addressed. E-mail: rarey@zbh.uni-hamburg.de.

Funding Sources: The development of PoseView is funded by the Klaus Tschira Stiftung gGmbH.

REFERENCES

(1) Pipeline Pilot; http://accelrys.com/products/pipeline-pilot/.

(2) Berthold, M.; Cebron, N.; Dill, F.; Gabriel, T.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. *KNIME: The Konstanz Information Miner.* Proceedings of the 4th annual industrial simulation conference, Workshop on multi-agent systems and simulations, Palermo, 2006.

(3) Wilkinson, L.; Friendly, M. The history of the cluster heat map. *Am. Statistician* **2009,** *63*, 179–184.

(4) Gehlenborg, N.; O'Donoghue, S.; Baliga, N.; Goesmann, A.; Hibbs, M.; Kitano, H.; Kohlbacher, O.; Neuweger, H.; Schneider, R.; Tenenbaum, D.; Gavin, A.-C. Visualization of omics data for systems biology. *Nat. Methods* **2010,** *7*, 56–68.

(5) Helson, H. Structure diagram generation. *Rev. Comput. Chem.* **1999,** *13*, 313–398.

(6) O'Donoghue, S.; Goodsell, D.; Frangakis, A.; Jossinet, F.; Laskowski, R.; Nilges, M.; Saibil, H.; Schafferhans, A.; Wade, R.; Westhof, E.; Olson, A. Visualization of macromolecular structures. *Nat. Methods* **2010,** *7*, 42–55.

(7) Wallace, A.; Laskowski, R.; Thornton, J. LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. *Protein Eng., Des. Sel.* **1995,** *8*, 127–134.

(8) Clark, A.; Labute, P. 2D depiction of protein-ligand complexes. *J. Chem. Inf. Model.* **2007,** *47*, 1933–1944.

(9) Stierand, K.; Maass, P.; Rarey, M. Molecular complexes at a glance: Automated generation of two-dimensional complex diagrams. *Bioinformatics* **2006,** *22*, 1710–1716.

(10) Stierand, K.; Rarey, M. From modeling to medicinal chemistry: Automatic generation of twodimensional complex diagrams. *ChemMedChem* **2007,** *2*, 853–860.

(11) Brecher, J. Graphical representation standards for chemical structure diagrams (IUPAC Recommendations 2008). *Pure Appl. Chem.* **2008,** *80*, 277–410.

(12) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. The protein data bank. *Nucleic Acids Res.* **2000,** *28*, 235–242.

(13) Lippert, T.; Rarey, M. Fast automated placement of polar hydrogen atoms in protein-ligand complexes. *J. Cheminformatics* **2009,** *1*, 13.

(14) Desiraju, G.; Steiner, T. *The Weak Hydrogen Bond: In Structural Chemistry and Biology*; Oxford University Press: United States, 2001.

(15) Seebeck, B.; Reulecke, I.; Kämper, A.; Rarey, M. Modeling of metal interaction geometries for protein-ligand docking. *Proteins: Struct., Funct., Bioinf.* **2008,** *71*, 1237–1254.

(16) McGaughey, G.; Gagne, M.; Rappe, A. π-Stacking interactions. *J. Biol. Chem.* **1998,** *273*, 15458–15463.

(17) Gallivan, J.; Dougherty, D. Cation-π interactions in structural biology. *Proc. Natl. Acad. Sci.* **1999,** *96*, 9459–9464.

(18) Hückel, E. Quantentheoretische Beiträge zum Benzolproblem. *Z. Phys. A Hadrons Nuclei* **1931,** *70*, 204–286.

(19) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H.; Westbrook, J. Ligand Depot: A data warehouse for ligands bound to macromolecules. *Bioinformatics* **2004,** *20*, 2153–2155.

(20) Berman, H.; Henrick, K.; Nakamura, H. Announcing the worldwide protein data bank. *Nat. Struct. Biol.* **2003,** *10*, 980–980.

(21) Stepanyuk, G.; Liu, Z.; Markova, S.; Frank, L.; Lee, J.; Vysotski, E.; Wang, B. Crystal structure of coelenterazine-binding protein from Renilla muelleri at 1.7Å: Why it is not a calcium-regulated photoprotein. *Photochem. Photobiol. Sci.* **2008,** *7*, 442–447.

(22) Ferguson, K.; Kavran, J.; Sankaran, V.; Fournier, E.; Isakoff, S.; Skolnik, E.; Lemmon, M. Structural basis for discrimination of

© 2010 American Chemical Society

3-phosphoinositides by pleckstrin homology domains. *Mol. Cell* **2000,** *6*, 373–384.

(23) Van den Akker, F.; Steensma, E.; Hol, W. Tumor marker disaccharide D-Gal-beta 1, 3- GalNAc complexed to heat-labile enterotoxin from Escherichia coli. *Protein Sci.* **1996,** *5*, 1184–1188.

(24) Mol, C.; Dougan, D.; Schneider, T.; Skene, R.; Kraus, M.; Scheibe, D.; Snell, G.; Zou, H.; Sang, B.; Wilson, K. Structural basis for the autoinhibition and STI-571 inhibition of c-Kit tyrosine kinase. *J. Biol. Chem.* **2004,** *279*, 31655–31663.

NOTE ADDED AFTER ASAP PUBLICATION This paper was published on the Web on August 31, 2010 with an error in the title. The corrected version was reposted on September 3, 2010.

© 2010 American Chemical Society