A STRUCTURAL AND ENERGETIC DESCRIPTION OF PROTEIN-PROTEIN INTERACTIONS IN ATOMIC DETAIL

A Thesis

by

TIFFANY BRINK FISCHER

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2006

Major Subject: Biochemistry

A STRUCTURAL AND ENERGETIC DESCRIPTION OF PROTEIN-PROTEIN INTERACTIONS IN ATOMIC DETAIL

A Thesis

by

TIFFANY BRINK FISCHER

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved by:

Chair of Committee.	Jerry W. Tsai
Committee Members,	James C. Hu
	Thomas R. Ioerger
	Patricia J. LiWang
Head of Department,	Gregory D. Reinhart

December 2006

Major Subject: Biochemistry

ABSTRACT

A Structural and Energetic Description of Protein-Protein Interactions in Atomic Detail.

(December 2006)

Tiffany Brink Fischer, B.S., Texas A&M University

Chair of Advisory Committee: Dr. Jerry Tsai

Here, we present the program QContacts, which implements Voronoi polyhedra to determine atomic and residue contacts across the interface of a protein-protein interaction. While QContacts also describes hydrogen bonds, ionic pair and salt bridge interactions, we focus on QContacts' identification of atomic contacts in a protein interface compared against the current methods. Initially, we investigated in detail the differences between QContacts, radial cutoff and Change in Solvent Accessible Surface Area (ΔSASA) methods in identifying pair-wise contacts across the binding interface. The results were assessed based on a set of 71 double cycle mutants. QContacts excelled at identifying knob-in-hole contacts. QContacts, closest atom radial cutoff and the ΔSASA methods performed well at picking out direct contacts; however, QContacts was the most accurate in excluding false positives. The significance of the differences identified between QContacts and previous methods was assessed using pair-wise contact frequencies in a broader set of 592 protein interfaces. The inaccuracies introduced by commonly used radial cutoff methods were found to produce misleading bias in the residue frequencies. This bias could compromise pair-wise potentials that are based on such frequencies. Here we show that QContacts provides a more accurate

description of protein interfaces at atomic resolution than other currently available methods. QContacts is available in a web-based form at http://tsailab.tamu.edu/qcons (Fischer et al., 2006).

DEDICATION

To my husband Tim and my son Dylan

ACKNOWLEDGEMENTS

I would like to thank Dr. Jerry Tsai for giving me the opportunity to work on this project. I would also like to thank the Tsai lab, in particular J. Bradley Holmes, David Schell, Xiaotao Qu and Rosemarie Swanson for their helpful discussions. I also want to acknowledge Gwen Knapp and Dr. James Hu for their discussions and guidance. I would also like to extend gratitude to my committee members, Dr. Patricia LiWang and Dr. Thomas Ioerger for their helpful comments, their support and for taking the time to listen.

TABLE OF CONTENTS

Page

ABSTRACT iii

DEDICATION v

ACKNOWLEDGEMENTS vi

TABLE OF CONTENTS vii

LIST OF FIGURES viii

LIST OF TABLES ix

CHAPTER

I INTRODUCTION 1

Current Methods in Identifying Binding Interfaces 3
Voronoi Polyhedra Approach 6

II ASSESSING METHODS FOR IDENTIFYING PAIR-WISE ATOMIC CONTACTS ACROSS BINDING INTERFACES 8

Summary 8
Introduction 9
Methods and Materials 13
Results and Discussion 20
Conclusion 44

III CONCLUSIONS 46

Future Work 47

REFERENCES 50

VITA 56

LIST OF FIGURES

Page

Figure 1. Calculating radial cutoffs. 5

Figure 2. Calculating Vorinoi polyhedra. 12

Figure 3. Comparing the QContacts method to ΔSASA in identifying a protein interface. 23

Figure 4. A closer look at the knob-in-hole interaction. 24

Figure 5. Indirect water mediated contacts. 28

Figure 6. Indirect residue mediated contacts. 29

Figure 7. Comparison of radial cutoff methods to QContacts. 34

Figure 8. Differences in pair-wise contact frequencies: the QContacts method against radial cutoff methods. 40

Figure 9. Direct comparison of QContacts to the closest atom radial cutoff method. ... 42

LIST OF TABLES

Page

Table 1. Comparison of QContacts to changes in solvent accessible surface area (ΔSASA): atoms/residues in the protein interface 22

Table 2. Predictive ability for identifying experimentally defined residue interactions 27

Table 3. Detailed predictive ability for identifying experimentally defined interactions 30

Table 4. Comparison of QContacts to radial cutoff methods: pair-wise residue contacts 36

Table 5. Amino acid propensities 38

CHAPTER I

INTRODUCTION

Interactions between proteins mediate most of the communication and regulation in the cell. Determining the rules governing such protein-protein interactions would provide a better understanding of protein interaction networks as well as approaches to control/intervene in signal transduction pathways and protein complexes. Once these rules of interactions are established, then we can use them to predict what proteins interact and how they interact. For this reasons, increasing effort is being directed towards understanding the determinants of protein interface recognition and affinity of protein-protein interactions. A measure of a true understanding of the characteristics that governs protein interactions is to use those characteristics to predict protein interactions. The Critical Assessment of Protein Interactions (CAPRI) was designed to measure the performance of methods for predicting protein interactions. One of the known problems that CAPRI has identified in various protein docking methods is the poor performance of scoring functions (Wodak and Mendez, 2004). Pair-wise contact potentials are incorporated into many of the scoring functions used in protein docking (Mendez et al., 2003). Even small inaccuracies in these potentials can add up to a substantial amount when they are summed over an entire protein interface. Thus,

---

This thesis follows the style of the Journal of Structural Biology.

programs that accurately identify contacts across the protein-protein interface are very important in protein docking methods.

An additional problem is the need for an accurate method for identification of atomic contacts for analyzing the features that make up protein binding interfaces. There are predicted to be 10,000 interface types of which 2,000 have been experimentally determined and the number of co-crystallized proteins is increasing by 300-400 every year (Aloy and Russell, 2004). The Alanine Scanning Energetics Database (AseDB) (Thorn and Bogan, 2001) and the Binding Interface Database (BID) (Fischer et al., 2003) together contain over 500 protein interactions and thousands of mutants. This wealth of protein-protein interaction data can be used to examine the amino acid frequencies (Janin and Chothia, 1990; Jones and Thornton, 1996), surface area (Janin and Chothia, 1990; Jones and Thornton, 1996), hot spot propensities (Bogan and Thorn, 1998), structural conservation (Ma et al., 2003), contact conservation (Hu et al., 2000) and the correlation between such attributes (Hu et al., 2000). Currently sequence, structural and contact conservation in combination with experimental information such as Hot spots are commonly used in protein docking as well as predicting protein interactions (Aytuna et al., 2005; Russell et al., 2004; Wodak and Mendez, 2004). In using this volume of information an accurate method for analysis of protein interactions is needed. The current methods used to identify contacts across the protein interface include radial cutoff methods and measuring a change in the solvent accessible surface area.

Current Methods in Identifying Binding Interfaces

More accurate computational tools are in need to combat the surge of data from increasing biological databases. Specifically these tools include ones that identify contacting residues and atoms across the protein interface. A number of computational methods have been developed to identify residues interacting across a protein binding interface. These methods include the most basic which uses radial cutoffs and the more complex which uses an interaction-dependent change in the solvent accessible surface area (ΔSASA) (Bahadur et al., 2003; Bahadur et al., 2004)**.** The fact that these methods are easy to implement and in most cases rapidly calculated demonstrate their utility. However there is also an indirect relationship between speed and accuracy such that the fastest of these methods do not give as good of atomic resolution. Radial cutoff methods have been implemented at both the residue and atom level in identifying protein interfaces. For the less precise, residue version, one centers the radial cutoff on the center of each residue's Cβ atom (Cα for Gly) (Glaser et al., 2001) or the residue's center of mass (centroid) (Caffrey et al., 2004). In Figure 1 you can see that each of the dots represents a Cβ or center of mass. Any dots that are within the radial cutoff (dotted line in Figure 1) of the Cβ or center of mass are considered to be in contact. For the atom version, a 6 Å radial cutoff is measured for each heavy atom (Ofran and Rost, 2003). These methods implement unified atom models, where hydrogens are incoorporated into an atomic group because the hydrogens are not defined in the crystal structure. The unified atom model does not accurately represent the variations in atom sizes and irregular packing (Tsai and Gerstein, 2002; Tsai et al., 1999). This

discrepancy further degrades the accuracy of the radial cutoff methods. The radial cutoff method is also limited in that it can not be used to calculate interaction area. If there is no way of calculating interaction area or identifying contacts accurately, then the method is not suited for calculating interaction energies.

ΔSASA is the most computationally expensive of the currently used methods, which can be attributed to the fact that it has to calculate the solvent accessible surface area for individual subunits and when the subunits are bound. The benefit to using this method over the radial cutoff methods is that it can give you an estimated surface area of the interface. The ΔSASA method identifies interacting residues by comparing the solvent accessible surface area (SASA) of a protein complex to the individual subunits (Bahadur et al., 2003; Bahadur et al., 2004). A change in the SASA when comparing the bound and unbound states is considered due to an interaction. An adaptation of this method is more similar to the radial cutoff methods in which any atoms that exhibit a ΔSASA are examined to see what neighboring atoms are within the 1.4 Å SASA probe plus the VDW radii distance (Mancini et al., 2004; Sobolev et al., 1999). Two problems when using the ΔSASA method are as follows: 1) the surface area of an interface is over-estimated (McConkey et al., 2002; Sobolev et al., 1999) and 2) the implementation does not always adequately identify pair-wise contacts. Pair-wise contact potentials and solvation are both used in predicting specific protein-protein interactions (Bahadur et al., 2004) and in many protein-protein docking scoring functions (Mendez et al., 2003). Overestimating the interface area and inaccurately identifying atoms in contact is problematic for programs that use these measures in protein interaction prediction.

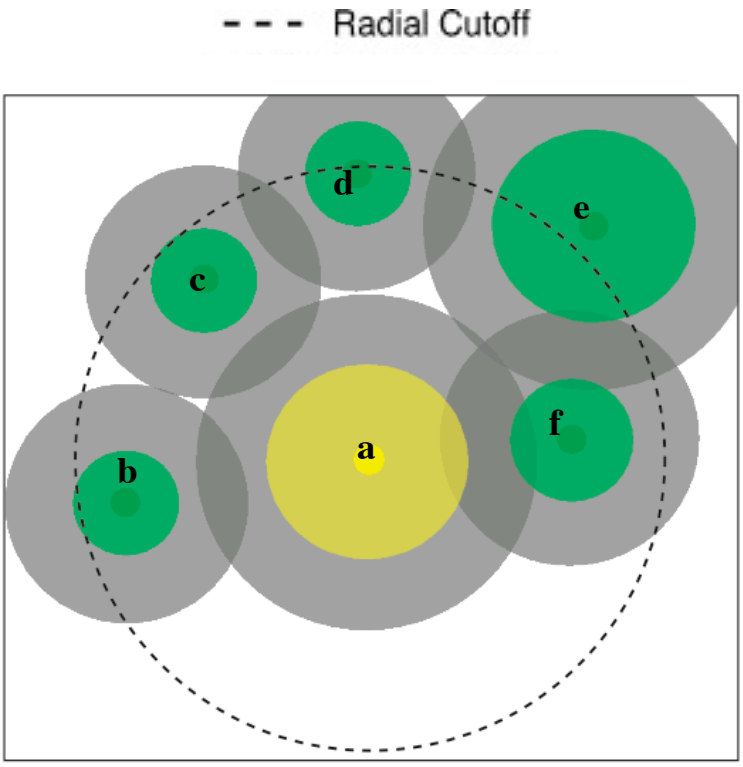

Figure 1. Calculating radial cutoffs. The center of the above atoms are shown as green or yellow points inside the unified atom radii representing their van der Waals (vdW) area which are also shaded in light green or yellow. A 1.4 Å radius is shaded light gray representing the water radii, and overlapping atoms are indicated by dark gray. Here the radial cutoff method is used to identify atoms contacting the atom labeled "a". The radial cutoff used is centered on the atom a and is indicated by the dashed circle. To be considered a contact by the radial cutoff method, another atom center must be enclosed by the dashed line. No overlapping area between atoms a and d indicates that there is no direct contact between the two atoms. Since the radial cutoff method does include the a-d contact, it is considered a false positive. Also, the radial cutoff does not include atom e as being in contact with atom a and is therefore considered a false negative.

Voronoi Polyhedra Approach

The Voronoi polyhedra method (Voronoi, 1908) is proposed here for identifying contacting atoms, residues and their interaction area in a protein-binding interface. Using Voronoi polyhedra volumes to identify nearest neighbors has been implemented previously in proteins (Richards, 1974) and in measuring the extent of packing in protein interfaces (Lo Conte et al., 1999). A caveat of this method is that as long as two atoms have no other atom between them, then they are considered neighbors regardless of how far apart they are. Therefore, the QContacts method uses an additional constraint to identify neighbors. The constraint requires that atoms must be close enough that a water molecule could not intervene. This limits contacts to physically real contacts by adding an atom dependent radial cutoff as described by the Laguerre polyhedral decomposition (Gellatly and Finney, 1982) or by adding a water radius to each atom radii termed the radical plane method (McConkey et al., 2002). Contacts within the added water radii are considered packing or water excluding interactions, such as hydrophobic interactions. This method identifies contacts while trying to minimize false contacts. This method is also unique in that it calculates the interface area as polyhedron edges bisecting contacting atoms. The ΔSASA method estimates the interface area as the buried surface area but it is not an atomic pair-wise interface area.

The accuracy of QContacts, radial cutoff and ΔSASA methods is compared in the present study. Contacts identified using each method are compared to mutagenesis studies that have addressed the validity of those contacts by measuring the interaction energy. A much larger dataset is then used to further address the significance of any

differences in contacts identified. QContacts is found to identify contacts in the interface accurately by minimizing false positives. We find that the ΔSASA method misses knob-in-hole contacts. In addition we find that the inaccuracies found using the radial cutoff methods create a bias in calculating residue frequencies.

CHAPTER II

ASSESSING METHODS FOR IDENTIFYING PAIR-WISE ATOMIC CONTACTS ACROSS BINDING INTERFACES[*]

Summary

An essential step in understanding the molecular basis of protein–protein interactions is the accurate identification of inter-protein contacts. We evaluate a number of common methods used in analyzing protein–protein interfaces: a Voronoi polyhedra-based approach, changes in solvent accessible surface area (ΔSASA) and various radial cutoffs (closest atom, Cβ, and centroid). First, we compared the Voronoi polyhedra-based analysis to the ΔSASA and show that using Voronoi polyhedra finds knob-in-hole contacts. To assess the accuracy between the Voronoi polyhedra-based approach and the various radial cutoff methods, two sets of data were used: a small set of 75 experimental mutants and a larger one of 592 structures of protein–protein interfaces. In an assessment using the small set, the Voronoi polyhedra-based methods, a solvent accessible surface area method, and the closest atom radial method identified 100% of the direct contacts defined by mutagenesis data, but only the Voronoi polyhedra-based method found no false positives. The other radial methods were not able to find all of

[*] Portions of this chapter have been reprinted with permission from "Assessing methods for identifying pair-wise atomic contacts across binding interfaces" by Fischer, T.B., Holmes, J.B., Miller, I.R., Parsons, J.R., Tung, L., Hu, J.C., Tsai, J., 2006. Journal of Structural Biology, 153, 103-112. With permission from Elsevier Inc.

the direct contacts even using a cutoff of 9 Å. With the larger set of structures, we compared the overall number of contacts found using the Voronoi polyhedra-based method as a standard. All the radial methods using a 6 Å cutoff identified more interactions, but these putative contacts included many false positives as well as missed many contacts. While radial cutoffs are quicker to calculate as well as to implement, this result highlights why radial cutoff methods do not have the proper resolution to describe in detail the non-homogeneous packing within protein interfaces, and suggests an inappropriate bias in pair-wise contact potentials. Of the radial cutoff methods, using the closest atom approach exhibits the best approximation to the more intensive Voronoi calculation. Our version of the Voronoi polyhedra-based method, QContacts is available at http://tsailab.tamu.edu/qcons (Fischer et al., 2006).

Introduction

A number of approaches have been developed to determine the residues participating in a protein interface. A common method for studying protein interfaces looks for an interaction-dependent change in the solvent accessible surface area (ΔSASA) (Bahadur et al., 2003; Bahadur et al., 2004). Adding a 1.4 Å probe to the van der Waals (vdW) radius of an atom, the solvent accessible surface area (SASA) is the sum of the non-overlapping atomic surfaces and effectively, defines the distance of closest approach for a water molecule to the protein (Lee and Richards, 1971). To analyze protein interfaces, the ΔSASA method simply identifies residues contributing to the molecular interaction by comparing the SASA of a protein complex compared to the

individual subunits. Those residues that exhibit a change in SASA between the two states are considered to be involved in the interaction. Any SASA changes in the complex compared to the monomers are examined to see what neighboring atoms are within the SASA probe plus vdW radii distance (Mancini et al., 2004; Sobolev et al., 1999). There are two main drawbacks in the ΔSASA method: the surface area of an interface is overestimated (McConkey et al., 2002; Sobolev et al., 1999) and in many cases the implementation does not adequately identify pair-wise contacts.

More simply, a basic approach to analyzing protein interfaces uses a radial cutoff to find pair-wise interactions. Radial cutoffs have been implemented at both the residue and atom level. The premise of this approach is that a radius can approximate an atom/residue's sphere of interaction. For the residue version, the radial cutoff is centered on a point representative of the residue such as the Cβ atom of a residue (Glaser et al., 2001) or the residue's center of mass, commonly referred to as its centroid (Caffrey et al., 2004). In the case of glycine, both of these approaches place the cutoff on the Cα. For the atom version, the radial cutoff is applied at the level of individual heavy atoms (Ofran and Rost, 2003), since hydrogen atoms are many times not resolved in structure files. A radial cutoff of 6 Å is typically chosen to maximize true positives and minimize false negatives. The unified atom models, where hydrogens are subsumed into an atomic group, do not account for variations in atom sizes and irregular packing (Tsai and Gerstein, 2002; Tsai et al., 1999). As a result, a uniform 6 Å radial cutoff is problematic for defining interactions between non-uniform atoms/residues. Also, the radial cutoff is a binary description of interactions, in which the extent to which atoms or

residues are in contact is not clearly distinguished. An interaction-free energy potential could be based on distances, but would suffer from inaccuracies intrinsic to the radial calculation of interactions. For example, two atoms with equal distances to the same atom may not have the same amount of contact area or interaction-free energy. This lack of precision in using radial cutoffs limits correct identification of interactions and prohibits accurate calculation of contact energetics.

One alternative method for identifying protein interaction area and atoms/residues across a protein interface is to use Voronoi polyhedra (Voronoi, 1908), which was first applied to proteins by Richards (1974). For simplicity, we will refer to this Voronoi polyhedra-based approach as the QContacts method. Calculating polyhedra is more computationally intensive, but the benefits are that the method can exactly locate nearest neighbors. Polyhedra volumes have been implemented previously in measuring the extent of packing in protein interfaces (Lo Conte et al., 1999). One drawback of polyhedra method is that two atoms can be in contact regardless of distance so long as no other atoms intervene between them. Therefore, variations have been developed that limit identification of physically real contacts by adding an atom-dependent radial cutoff as described by the Laguerre polyhedral decomposition (Gellatly and Finney, 1982) or by adding a water radius to each atom radii termed the radical plane method (McConkey et al., 2002). The radical plane method further reduces false positives by excluding neighbors that are separated by a large distance (Figure 2). A vdW radial overlap is considered a direct interaction, whereas those within the added water radii are considered water excluding interactions.

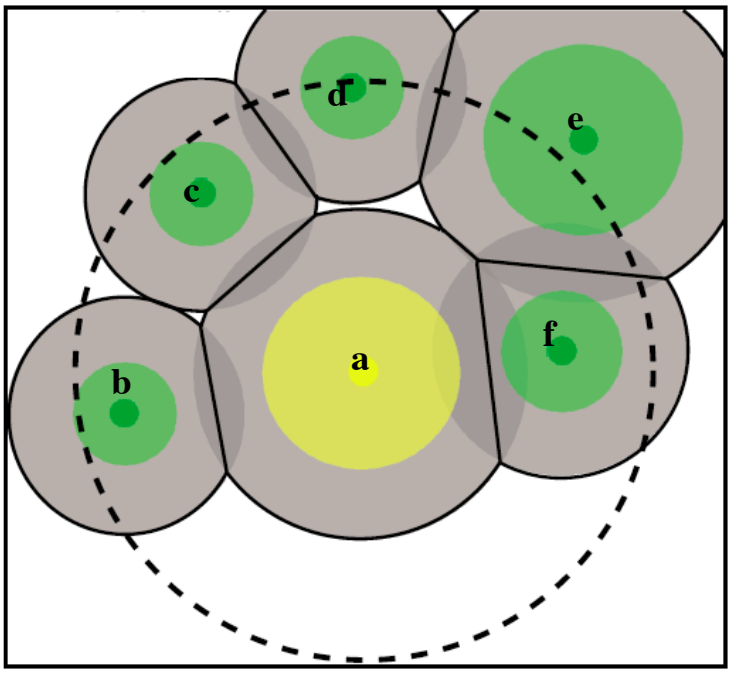

Figure 2. Calculating Vorinoi polyhedra. The center of atoms and the unified atom radii representing their van der Waals (vdW) area are shaded in light green or yellow. The extended 1.4 Å water radius is shaded light gray, and atom overlaps are shown in dark gray. Boundaries to atoms calculated by the radical plane method are shown in solid black lines. A radial cutoff is shown by the dashed circle. The QContacts method finds direct contacts with central atom a, which are interacting with atoms b, c, e, and f, and properly excludes the potential indirect interaction with atom d. The a to d contact is not a direct contact because the distance between the two atoms is large enough to allow a water molecule between them, where the center of the water molecule could occupy any point in the white space formed by the edges of atoms a, c, d, and e. If atoms a and d were polar and a water were to bridge between the two atoms in the white space, then these atoms could be considered an indirect water-mediated interaction. In addition, another water could potentially occupy the white space created by atoms a, b, and c, but in this case, all of these atoms contact each other. In comparison, the radial cutoff identified atom d, which is not in contact as described previously, identifying a false positive. Also, the radial cutoff does not include atom e, producing a false negative. While the radial cutoff passes through the vdW area of atom e, the center of atom e needs to be within the radial cutoff to be considered interacting with atom a. The QContacts method is also more accurate than the ΔSASA method, in that ΔSASA does not identify atoms in crevices as being able to interact. For instance the atoms b and c in the figure are not far enough for a water molecule to fit in between, but are far enough apart for a small atom to squeeze between these atoms and contact atom a.

Here, we compare all of the above methods against each other in identifying interactions across protein interfaces. Initially, we investigated in detail the differences between ΔSASA and a radical plane implementation of the QContacts method in associating individual atoms/residues with a binding interface. Next, the accuracies of the QContacts method and radial cutoff methods in identifying pair-wise contacts were assessed for their predictive ability on a small set of experimental data. Using a broader set of protein interfaces, we then expand this analysis on the significance of the differences in pair-wise contact frequencies. In particular, we find that the inaccuracies introduced by commonly used radial cutoff methods can produce misleading analyses of the residues involved in protein interfaces. We find that the QContacts approach provides a more accurate description of protein interfaces at atomic resolution than the ΔSASA or radial cutoff methods.

Methods and Materials

Radical plane contacts

The Laguerre polyhedral decomposition (Gellatly and Finney, 1982) of the radical plane method was implemented with the Chothia radii set (Chothia, 1974), in which atom specific and hydrogen volumes are unified into one value. The radii are "extended" by the equivalent of a water radius (1.4 Å) in order to account for water exclusion. This extended radical plane Voronoi polyhedra-based method (McConkey et al., 2002) calculates the amount of contact as the area of the plane between the two atoms. For simplicity, we refer to our implementation of this radical plane Voronoi

polyhedra-based calculation in the text as the QContacts method. The total contact area between two residues is defined as the sum of the inter-residue atomic contacts.

Identifying protein binding interfaces using the ΔSASA method

The SASA is calculated using a previously described method (Lee and Richards, 1971) except that the same Chothia (1974) radii set used for the radical plane method was implemented. In this method, the protein is defined by calculating the vdW radius for each atom. To define the surface of the protein, each atom's radius is extended by 1.4 Å. The calculation steps through the atoms at 0.25 Å increments to classify whether any overlapping surface area is buried. The remaining unidentified surface is considered accessible to the water solvent. As done previously (Bahadur et al., 2003; Bahadur et al., 2004), the SASA is calculated twice to identify those atoms/residues involved in a protein surface: once for the monomer (s) and once for the complex. If there is a change in the SASA (ΔSASA) of an atom/residue when going from the monomer to the dimer form, then it is considered involved in the protein interface. For pair-wise interactions using a SASA approach, we used a cutoff between the two atoms. If the two atoms were within the distance defined by summing their respective radii from Chothia (1974) radii set plus 2.8 Å (the diameter of a water atom), they were considered in contact.

The water probe is calculated as the vdW radii plus a 1.4 Å probe. The water probe method is unique from the other radial cutoff methods in that it does not use a fixed radial cutoff. Instead, this method uses Chothia (1974) radii set to calculate the vdW radii. If two atoms were found to be within the sum of their water probe, then they were considered to be in contact (Lo Conte et al., 1999; Mancini et al., 2004).

Identifying protein binding interfaces using radial cutoff methods

Similar to previous work (Ofran and Rost, 2003), atoms were used as centers (closest atom method). In another method (Glaser et al., 2001), the Cα for Gly and Cβ for the remaining residues were used as centers (Cβ method). We also used a third center, which is a residue's centroid or center of mass of a residue (Kazmierkiewicz et al., 2003). A residue centroid is defined as the average of all side-chain atoms (centroid method). For Gly, the Cα atom is considered as residue's centroid. For each of the aforementioned centers, atoms or residues were considered in contact if they fell within the defined cutoff radius. Three radial cutoffs (water probe, 3 Å, 6 Å, and 9 Å) were used to calculate the interactions between three types of residue/atom centers. These cutoffs were used as they have been used in previous studies (Bordner and Abagyan, 2005; Glaser et al., 2001; Kazmierkiewicz et al., 2003; Ofran and Rost, 2003). While many studies use a range of values (Dall'Acqua et al., 1998; Li et al., 2003; Lu et al., 2003; Papageorgiou et al., 1997), radial cutoffs generally fall within this range. We use the 3 Å, 6 Å, and 9 Å values for simplicity and consistency.

Datasets, statistics, and analysis

Two sets of structures were used in this study. To qualitatively compare the ability of the various methods to identify experimentally found contacts, a set of 71 double cycle mutants with solved co-crystals were chosen from the Binding Interface Database (Fischer et al., 2003). A total of five structure files and eight interfaces were used and are listed as follows with the Protein Data Bank (Berman et al., 2000) or PDB codes along with the number of mutations found across an interface: 1a4y (Papageorgiou

et al., 1997) with 5 mutations across the AB interface (Chen and Shapiro, 1999); 1brs (Buckle et al., 1994) with 33 mutations across the CF interface (Schreiber and Fersht, 1995); 1dqj (Li et al., 2000) with 5 and 8 mutations across the AC and BC interfaces, respectively (Li et al., 2003); 1vfb (Bhat et al., 1994) with 7 and 7 mutations across the AC and BC interfaces, respectively (Dall'Acqua et al., 1998); 3hfm (Padlan et al., 1989) with 4 and 6 mutations across the HY and LY interfaces, respectively (Pons et al., 1999). Atoms are categorized as being direct contact, indirect water-mediated, and indirect atom/residue-mediated. Direct contacts are defined as two types, vdW contacts (the distance between two atoms is less than the sum of their vdW radii) or water exclusion contacts (the distance between the two atoms being less than the total distance of a vdW radii plus a water radii). Either must exhibit a significant interaction energy: experimentally measured $\Delta\Delta G$ of greater or equal to 0.5 kcal/mol (attractive contacts) or less than or equal to –0.5 kcal/mol (repulsive contacts). An indirect water-mediated interaction occurs between two atoms, where at least one water molecule separates their radii and they exhibit a significant interaction energy. Indirect atom/residue-mediated contacts occur when atoms are separated by at least one other atom and exhibit a significant, experimentally determined interaction energy. Those atoms not falling into one of the above categories are considered to be not in contact or no contact class. This fourth class exhibits an experimentally determined insignificant energetic interaction between 0.49 and –0.49 kcal/mol and structurally do not overlap vdW radii or exclude waters.

To analyze the difference in calculated contact frequencies, a set of 592 co-crystallized protein–protein interactions was used in this work. The chains are separated from the four letter PDB code by a dash, for example the PDB id 6gsv chains A and B are shown as 6gsv-AB. These PDB codes were chosen to create a non-redundant set by limiting sequence identity to less than 30% between chains of different PDB entries as described previously (Glaser et al., 2001). The list of Protein Data Bank or PDB (Berman et al., 2000) codes are as follows: 1anw-AB, 1aor-AB, 1aoz-AB, 1apx-AB, 1apy-AB, 1apy-AC, 1apy-BD, 1asy-AB, 1atn-AD, 1avd-AB, 1bar-AB, 1bbb-AB, 1bbp-BD, 1bbr-EK, 1bbt-14, 1bbt-23, 1bbt-24, 1bcf-AB, 1bcm-AB, 1bdm-AB, 1bgl-AD, 1bgs-FG, 1bhm-AB, 1bin-AB, 1blb-AB, 1bmf-AG, 1bmf-CD, 1bmf-DF, 1bmf-DG, 1bmt-AB, 1bmv-12, 1bnc-AB, 1bnd-AB, 1bov-BC, 1bpl-AB, 1bql-LY, 1brs-CF, 1bsr-AB, 1bun-AB, 1bvp-23, 1c2r-AB, 1cax-AC, 1cax-CF, 1cbi-AB, 1cdk-AB, 1cdl-AC, 1cea-AB, 1cgj-EI, 1chk-AB, 1chm-AB, 1cho-EI, 1chr-AB, 1cki-AB, 1cle-AB, 1clx-AB, 1cmc-AB, 1cns-AB, 1col-AB, 1cpc-AB, 1cpc-AK, 1csk-AD, 1csm-AB, 1cud-AB, 1cwe-AC, 1cwp-AB, 1cyd-CD, 1d66-AB, 1daa-AB, 1dbq-AB, 1dcp-GH, 1dea-AB, 1dek-AB, 1dfn-AB, 1dif-AB, 1dir-AB, 1dkt-AB, 1dky-AB, 1dlh-BE, 1dmx-AB, 1dnp-AB, 1dok-AB, 1dpg-AB, 1dpp-AC, 1dpr-AB, 1dth-AB, 1dut-AB, 1dvf-BD, 1dvr-AB, 1dyn-AB, 1ebd-BC, 1ece-AB, 1ecf-AB, 1ecm-AB, 1ecp-BD, 1ecz-AB, 1edh-AB, 1efn-AB, 1efn-BD, 1efu-AB, 1efu-AC, 1efu-BD, 1epa-AB, 1ept-AB, 1ept-AC, 1ept-BC, 1esf-AB, 1etp-AB, 1ext-AB, 1fat-AB, 1fba-BC, 1fbi-QY, 1fc1-AB, 1fc2-CD, 1fcb-AB, 1fcc-AC, 1fcd-AB, 1fcd-BD, 1fia-AB, 1fie-AB, 1fin-AB, 1fjm-AB, 1fki-AB, 1fle-EI, 1fod-12, 1fod-13, 1fos-GH, 1frp-AB, 1frt-AC, 1frt-BC, 1frv-AC, 1frv-CD, 1fss-AB,

1fuj-AB, 1fuq-AB, 1fvp-AB, 1fxi-AD, 1fxr-AB, 1g6n-AB, 1gad-OP, 1gam-AB, 1gar-AB, 1gdh-AB, 1gdt-AB, 1ges-AB, 1gfl-AB, 1ggg-AB, 1ghs-AB, 1gif-BC, 1gla-FG, 1glq-AB, 1glu-AB, 1got-AB, 1got-BG, 1gp1-AB, 1gpm-BD, 1gri-AB, 1gto-BC, 1gtp-BI, 1gtq-AB, 1gua-AB, 1gyl-AB, 1hav-AB, 1hbh-CD, 1hcg-AB, 1hcn-AB, 1hde-AB, 1hge-AC, 1hge-CD, 1hiw-AR, 1hjr-BD, 1hle-AB, 1hlp-AB, 1hmp-AB, 1hng-AB, 1hpc-AB, 1hpl-AB, 1hrd-BC, 1hrh-AB, 1hro-AB, 1hsa-AD, 1hsl-AB, 1hst-AB, 1htm-DF, 1htt-AB, 1huc-AB, 1hul-AB, 1hur-AB, 1hxp-AB, 1hyh-AB, 1hyl-AB, 1ice-AB, 1ids-AC, 1ihf-AB, 1ilr-12, 1inh-AB, 1isu-AB, 1ith-AB, 1jst-AC, 1kba-AB, 1kif-BF, 1kir-BC, 1kny-AB, 1kob-AB, 1kp8-FG, 1kpt-AB, 1lcp-AB, 1leh-AB, 1lgb-AC, 1lmb-34, 1lmk-EG, 1lmw-BD, 1lpb-AB, 1lts-AC, 1lts-DE, 1luc-AB, 1lwi-AB, 1lya-BD, 1lyl-AC, 1lyn-AB, 1mac-AB, 1mas-AB, 1mda-HJ, 1mda-HL, 1mdp-12, 1mdt-AB, 1mdy-AB, 1mec-14, 1mee-AI, 1mhl-AC, 1mhl-CD, 1mka-AB, 1mld-AB, 1mmo-BC, 1mmo-CE, 1mmo-CH, 1mmo-DE, 1mmo-EH, 1mol-AB, 1mpm-BC, 1msa-AD, 1mtn-BF, 1mtn-FH, 1mtn-GH, 1myk-AB, 1nal-23, 1nba-AB, 1nci-AB, 1nco-AB, 1nfk-AB, 1nip-AB, 1noy-AB, 1npo-AC, 1nsn-HS, 1nsn-LS, 1oac-AB, 1obp-AB, 1occ-FS, 1occ-GN, 1occ-HU, 1occ-NO, 1occ-NP, 1occ-NQ, 1occ-NS, 1occ-NU, 1occ-NW, 1occ-NX, 1occ-NY, 1occ-NZ, 1occ-OQ, 1occ-OR, 1occ-OU, 1occ-OV, 1occ-OX, 1occ-PS, 1occ-PT, 1occ-PU, 1occ-QR, 1occ-QS, 1occ-QV, 1occ-QX, 1occ-QZ, 1occ-RS, 1occ-RV, 1occ-ST, 1occ-SW, 1occ-TU, 1occ-WY, 1occ-YZ, 1onr-AB, 1ord-AB, 1oro-AB, 1ort-BF, 1osj-AB, 1otf-BE, 1otg-BC, 1ova-AB, 1ova-CD, 1ovo-AB, 1pag-AB, 1pam-AB, 1pdg-AB, 1pfx-CL, 1pge-AB, 1pio-AB, 1pky-AC, 1pma-12, 1pma-CD, 1pma-CP, 1pml-BC, 1pnk-AB, 1pov-03, 1pox-AB, 1ppf-EI, 1prc-CH, 1prc-CL, 1prc-CM, 1prc-LM, 1pre-AB, 1prt-AB,

1prt-AE, 1prt-AF, 1prt-EF, 1prt-HJ, 1prt-HL, 1psa-AB, 1psd-AB, 1pvc-12, 1pvc-13, 1pvc-24, 1pvc-34, 1pvd-AB, 1pvu-AB, 1pxt-AB, 1pya-CD, 1pya-CE, 1pya-DF, 1pyt-AB, 1pyt-AC, 1pyt-BD, 1qap-AB, 1qas-AB, 1qbe-BC, 1qor-AB, 1qpa-AB, 1qrd-AB, 1rah-BD, 1rba-AB, 1rcm-AB, 1rcp-AB, 1rdl-12, 1reg-XY, 1rfb-AB, 1rgf-AB, 1rhg-AC, 1rlb-AF, 1rn1-AC, 1rth-AB, 1rtm-12, 1rtp-23, 1rva-AB, 1sac-CD, 1sce-BD, 1sch-AB, 1scm-AB, 1scm-BC, 1scu-BE, 1scu-DE, 1seb-AB, 1seb-EH, 1sei-AB, 1sem-AB, 1set-AB, 1sgp-EI, 1slt-AB, 1slu-AB, 1smn-AB, 1smp-AI, 1spb-PS, 1sph-AB, 1sri-AB, 1stf-EI, 1stm-BC, 1sva-23, 1tab-EI, 1tah-AC, 1tbr-KS, 1tcb-AB, 1tco-AB, 1tco-AC, 1tco-BC, 1tcr-AB, 1tgx-AB, 1the-AB, 1thj-BC, 1tht-AB, 1tii-AC, 1tlf-AB, 1tmc-AB, 1tme-12, 1tme-13, 1tme-23, 1tmf-13, 1tmf-14, 1tmf-23, 1tmf-24, 1tmf-34, 1tnd-AC, 1tnf-AB, 1tph-12, 1trk-AB, 1tro-AC, 1tsd-AB, 1tsr-AB, 1tta-AB, 1tvx-BD, 1ubs-AB, 1ucy-HK, 1udi-EI, 1umu-AB, 1una-AB, 1vcp-BC, 1vfb-AB, 1vfb-AC, 1vhi-AB, 1vmo-AB, 1vok-AB, 1vol-AB, 1vrt-AB, 1vsc-AB, 1vsg-AB, 1wap-BC, 1wdc-AC, 1wgt-AB, 1wht-AB, 1wtl-AB, 1xik-AB, 1xim-AC, 1xso-AB, 1xva-AB, 1xxa-DF, 1xyp-AB, 1ycb-AB, 1ygp-AB, 1yha-AB, 1ypt-AB, 1yrn-AB, 1ytf-AD, 1ytf-BD, 1ytt-AB, 1zop-AB, 256b-AB, 2abx-AB, 2ach-AB, 2adm-AB, 2afn-BC, 2bbk-HJ, 2bbv-BC, 2bpa-13, 2btf-AP, 2ccy-AB, 2cht-DE, 2cst-AB, 2dhf-AB, 2dld-AB, 2drp-AD, 2eip-AB, 2gls-BH, 2hhm-AB, 2hmq-CD, 2hnt-CF, 2hpp-HP, 2kai-AI, 2kai-BI, 2kau-AB, 2kau-AC, 2kau-BC, 2lig-AB, 2ltn-AC, 2mev-12, 2mev-23, 2mev-34, 2mta-AC, 2nac-AB, 2pcc-AB, 2pcd-BC, 2pcd-BN, 2pcd-MP, 2pel-BC, 2phl-BC, 2pka-AB, 2pka-BY, 2plv-14, 2psp-AB, 2ptc-EI, 2rbi-AB, 2rmc-EG, 2rsl-AB, 2rsp-AB, 2scp-AB, 2spc-AB, 2tmd-AB, 2trx-AB, 2utg-AB, 2zta-AB, 3cro-LR, 3hhr-AB, 3hhr-BC, 3ins-BD, 3lad-AB, 3mde-AB, 3mon-BD, 3mon-

CD, 3mon-CE, 3pga-24, 3pmg-AB, 3sdh-AB, 4aah-AC, 4aah-CD, 4ake-AB, 4cha-AB, 4cts-AB, 4dfr-AB, 4kbp-BC, 4sbv-AB, 4sgb-EI, 5cna-AB, 6chy-AB, 6gsv-AB.

Contact frequencies were calculated as described previously (Ofran and Rost, 2003), which is shown below:

$$f_{ij} = \sum_{i=1}^{20}\sum_{j=1}^{20} \text{aa}_{ij}/\text{total contacts} \qquad (1)$$

To calculate these pair-wise contact frequencies *fij*, the total number of amino acids (aa) *i* and *j* in contact is divided by the total number of residues involved in the interface. Interaction energies were scaled by 0.045 kcal/mol/A^2 (Raschke et al., 2001), which is a fit of hydrophobic surface area calculated using Voronoi to experimentally determine solvation energies.

Program runtime comparison

The QContacts, ΔSASA, and the radial cutoff method's speed were compared using the structure 2tmd (Barber et al., 1992) chains A and B. The run time for each program was measured using the A and B chains of 2tmd.

Results and Discussion

QContacts versus ΔSASA in identifying protein interface atoms/residues Table 1 compares the number of atoms and residues identified in binding interfaces using the QContacts and ΔSASA over a set of 592 co-crystals (see Methods and Materials for selection criteria and a full listing of structures). As a detailed description of types of interactions has been described previously (Lo Conte et al., 1999), here we discuss the

differences between the two methods. Since both of these methods are similar in that they add a water radius, we expected the differences to be small, as illustrated in Figure 3. Computationally, both of these methods require the same amount of time to calculate, where the ΔSASA has a slight 10% advantage over the QContacts method. Overall, the QContacts method found about 10% more atoms and 1% more residues than ΔSASA in protein–protein interfaces. That the QContacts method identified far more atoms than residues compared to ΔSASA is not surprising. Although ΔSASA missed many atoms involved in the interface, the method needs to only identify one atom of a residue to consider it involved in the protein–protein interaction. Therefore, decreasing the resolution from atoms to residues improves the agreement between these two approaches. In addition, less than 1% of the total atoms and residues were identified by ΔSASA only. A closer inspection of the atom and residue differences for both of these classes distinguished the following types. Out of our set, only 463 atoms and 63 residues were found by ΔSASA but not by the QContacts method. We have found two explanations for these. First, since the ΔSASA uses a smaller cutoff of 0.001 $Å^2$ to the QContacts's 0.01 $Å^2$, more atoms/residues with a small contribution are considered as part of the interface in comparison to the QContacts method. Lastly, similar to above, crystallographic errors were found in two cases. For example, the Cγ atom of Pro57, chain W from the structure 1ooc (Tsukihara et al., 1996) is within 0.59 Å of the terminal O atom of the same residue, in which case they are effectively overlapping. Typically, the distance between these atoms should be approximately 4.0 Å, as they are covalently linked by three bonds.

Table 1. Comparison of QContacts to changes in solvent accessible surface area (ΔSASA): atoms/residues in the protein interface

Type	QContacts total	ΔSASA total	Both[a]	QContacts only	ΔSASA only
Atoms	165,790	150,385	149,922	15,868	463
Residues	38,565	38,252	38,189	376	63

[a] Those atoms/residues identified by both QContacts and ΔSASA as in the interface.
Those atoms and residues that are indicated as the "total" number means that all atoms/residues found in the interface using that method are counted. Atoms and Residues that are indicated as "only" one method, means that the atoms/residues identified were not found using the other method.

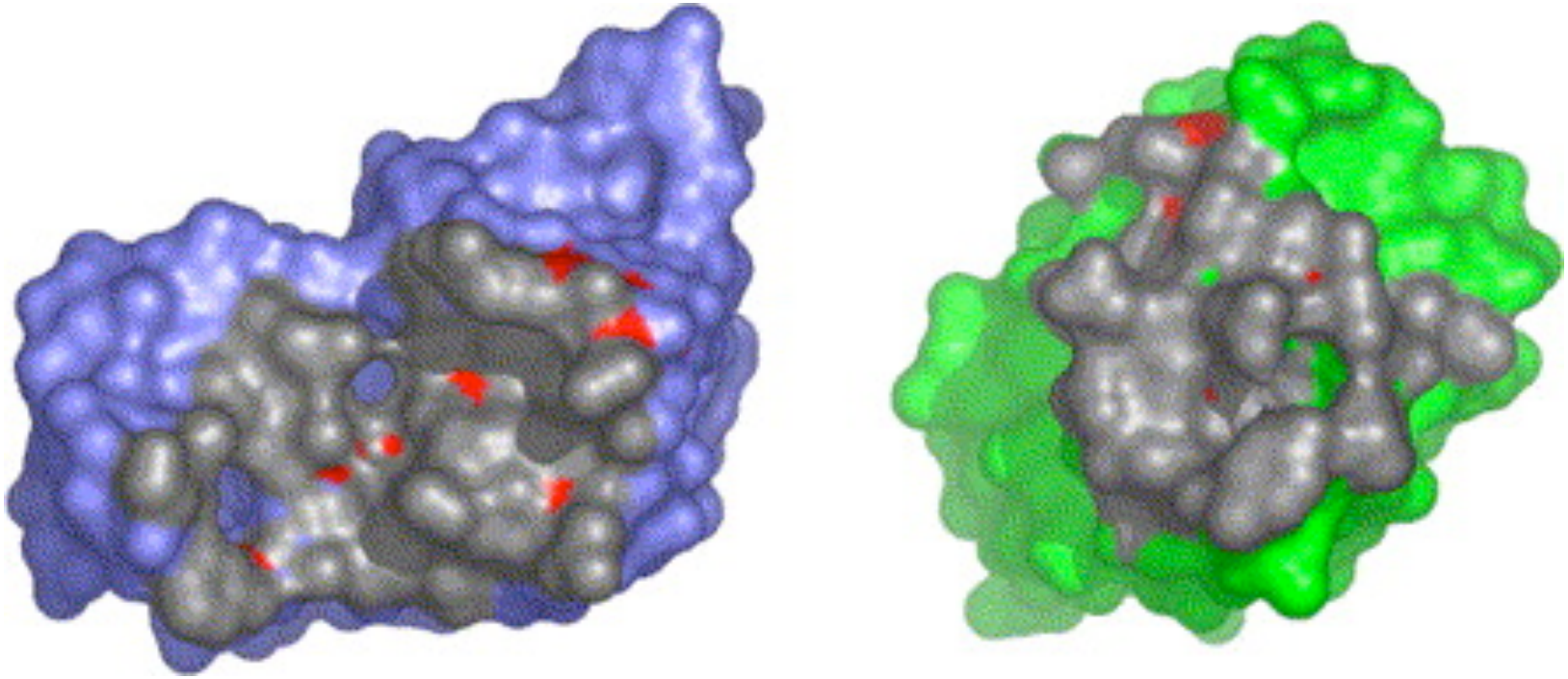

Figure 3. Comparing the QContacts method to ΔSASA in identifying a protein interface. The barnase–barstar interface from structure 1brs (Buckle et al., 1994) drawn using PyMol (DeLano, 2002). In each of the parts, barnase is represented in violet and barstar is represented in green. The interface is shown as an open book, where the respective interacting surfaces are displayed. The gray and red surfaces are calculated by the QContacts method, whereas the ΔSASA only identifies the gray surface. Therefore, ΔSASA underestimates the interaction since the red surface is not found by the ΔSASA method due to knob-in-hole contacts.

More significantly, contacts found by the QContacts method but not ΔSASA consisted of 15,868 atoms and 376 residues. We discovered that this occurs because of three reasons. First, knob-in-hole interactions (red surface in Figure 3) occur when the atoms/residues pack in the interface, but they are not accessible to solvent in the monomers based on the SASA calculation. These make up the majority of the differences between the QContacts calculation and not ΔSASA. Basically, due to the 1.4 Å addition to a neighbor atom's radii, the ΔSASA considers the atom/residue buried in both the monomer and the interaction complex. So, although this atom/residue forms part of the interaction interface, it exhibits no change in SASA and so it is not considered as part of the interface (Figure 4).

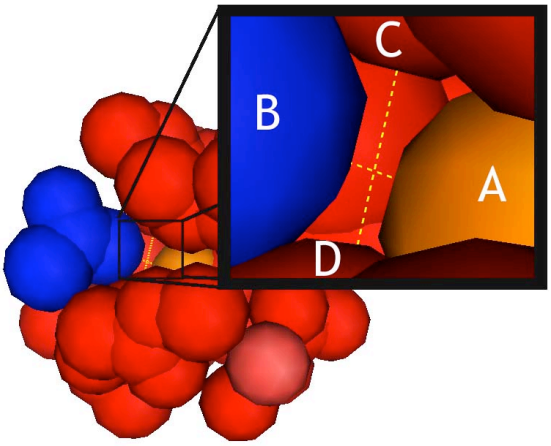

Figure 4. A closer look at the knob-in-hole interaction. In the background is a 6 Å expansion around atom ND2 of residue 43 chain A PDB 1aap. A cross section of the resulting area is shown. The orange atom (Atom ND2, residue 43 chain A (labeled A) is found to be in contact with the blue atom from chain B (labeled B) using QContacts but not ΔSASA. This is because the added water radii from the surrounding atoms (C and D) occlude atom A from the solvent according to ΔSASA (as indicated by a distance between atoms C and D of 1.85 Å). However, even though a water molecule cannot fit between C and D, an atom can. The distance between atom B and A is 0.47 Å and has a surface area of 9.8 $Å^2$.

Therefore, knob-in-hole interactions are found in deep crevices within the protein as a general underestimation of contacts in concave surfaces. To understand the significance of these missed contacts, a formula developed previously was implemented to quantify the energetic contribution of the knob-in-hole interactions. Twenty percent of the PDB structures had knob-in-hole interactions that amounted to a significant (0.5 kcal/mol) contribution to the binding energy. As a result, overlooking knob-in-hole contacts has been found to affect protein-docking prediction (Connolly, 1986; Norel et al., 1994; Norel et al., 1995; Norel et al., 1999; Schneidman-Duhovny et al., 2003; Wang and Wade, 2003). Second, approximations in the SASA step size can cause certain differences. If this step size is larger than some very small contact interfaces, those will not be counted. In such a case, the QContacts method finds a very small surface area. This error only accounts for a very small portion of the data. This was confirmed by calculating the difference between a step size of 0.25 Å and 0.05 Å, which changes the total of these type of contacts by 0.12%. Lastly, crystallographic errors also account for a very small portion of the contacts found. For example, the chains A and C from the structure 1frt (Burmeister et al., 1994) have atoms from opposing chains that overlap so much that they are closer than if they were covalently bonded: atom O of residue 114 from chain A and atom $C\gamma_1$ of residue 153 from chain C are separated by only 1.22 Å. In this case, atoms in the crystal structure are seen to overlap otherwise buried atoms on the opposing chain.

Validating pair-wise contact identification with experimental double cycle mutants

As explained above, a primary analysis of a protein interface needs to accurately identify pair-wise contacts. A set of data from double cycle mutant experiments of protein interfaces was gathered that contained a data on a total of 75 pair-wise mutants. These mutant pairs were grouped into the following four classes based on their experimentally determined energy of interaction and structural attributes as detailed in Methods and Materials. As shown in Table 2, 49 residue pairs were found to be direct contact; 3 indirect water-mediated interactions; 4 indirect residue-mediated; and 19 were considered not to be in contact or no contact. Although they did not cite where these cutoff values were derived from, most used the following values to determine contacting atoms (Dall'Acqua et al., 1998; Li et al., 2003; Papageorgiou et al., 1997; Sheriff et al., 1987): 4.1 Å between two carbons; 3.8 Å between a carbon and a nitrogen; 3.7 Å between a carbon and an oxygen; 3.3 Å between two oxygens; 3.4 Å between an oxygen and a nitrogen; and 3.4 Å between two nitrogens. Unlike these, the barnase–barstar analysis used a 7 Å cutoff to define contacts (Schreiber and Fersht, 1995) (as seen in Figures 5A and Figure 6A and B). It is somewhat interesting to discover that at least seven interactions (sum of the indirect classes) fell well outside of these cutoff values, but the authors did not report how they deduced that these residues should be interacting (see Figure 5B). Also, the large radial cutoff used in the barnase–barstar analysis produced the largest number of interactions (17) in the no contact class, as seen in Table 3.

Table 2. Predictive ability for identifying experimentally defined residue interactions

Class	Muta-genesis	QContacts	SASA	Closest atom			Centroid			Cβ		
				3 Å	6 Å	9 Å	3 Å	6 Å	9 Å	3 Å	6 Å	9 Å
Total [a]	75	388	483	107	1179	4273	0	333	1423	0	175	1065
Direct	49	49	49	12	49	49	0	28	43	0	8	25
Indirect water	3	0	0	0	0	1	0	0	3	0	0	0
Indirect residue	4	0	1	0	1	1	0	0	1	0	0	0
No contact	19	0	0	0	0	10	0	0	1	0	0	0

A more detailed version of this table listing each mutation and energies can be found in Table 3.

[a] The interaction in the 1brs structure between Y29 of the F chain and the H102 of the C chain was mutated twice, but this interaction was only considered once for this table. In Table 3, this interaction is listed twice with both energies for comparison

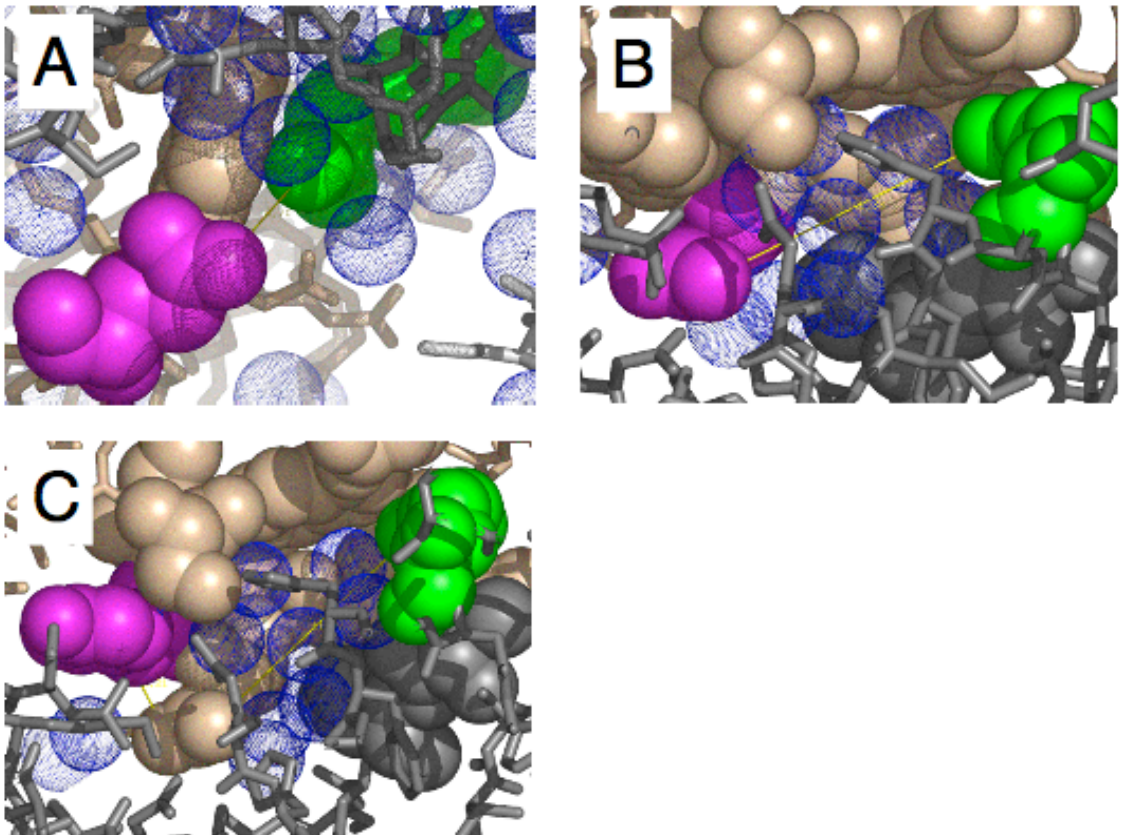

Figure 5. Indirect water mediated contacts. Mutated residues from opposing proteins are shown in green and pink space fill respectively. Other residues are colored dark grey and brown, respectively, and are either space fill or stick for clarity. Waters are displayed as purple dot surfaced spheres. Yellow lines depict direct distances between atoms, and distances are measure from atom centers. All images made using PyMol (Copyright © 2004 DeLano Scientific). A) chain C Arg59 to chain F Glu80 of the barnase-bastar complex 1brs are 6.7 Å apart. B) chain B Arg5 to chain A Asp435 of the angiogenin-RNaseI complex 1a4y are over 11 Å apart, but are bridged by the crystallographically resolved waters in the binding interface. C) chain B Gln5 to chain A Tyr434 of 1a4y are separated by 9 Å. If the atom radii are subtracted, the remaining distance would allow for at least one water molecule to fit between them.

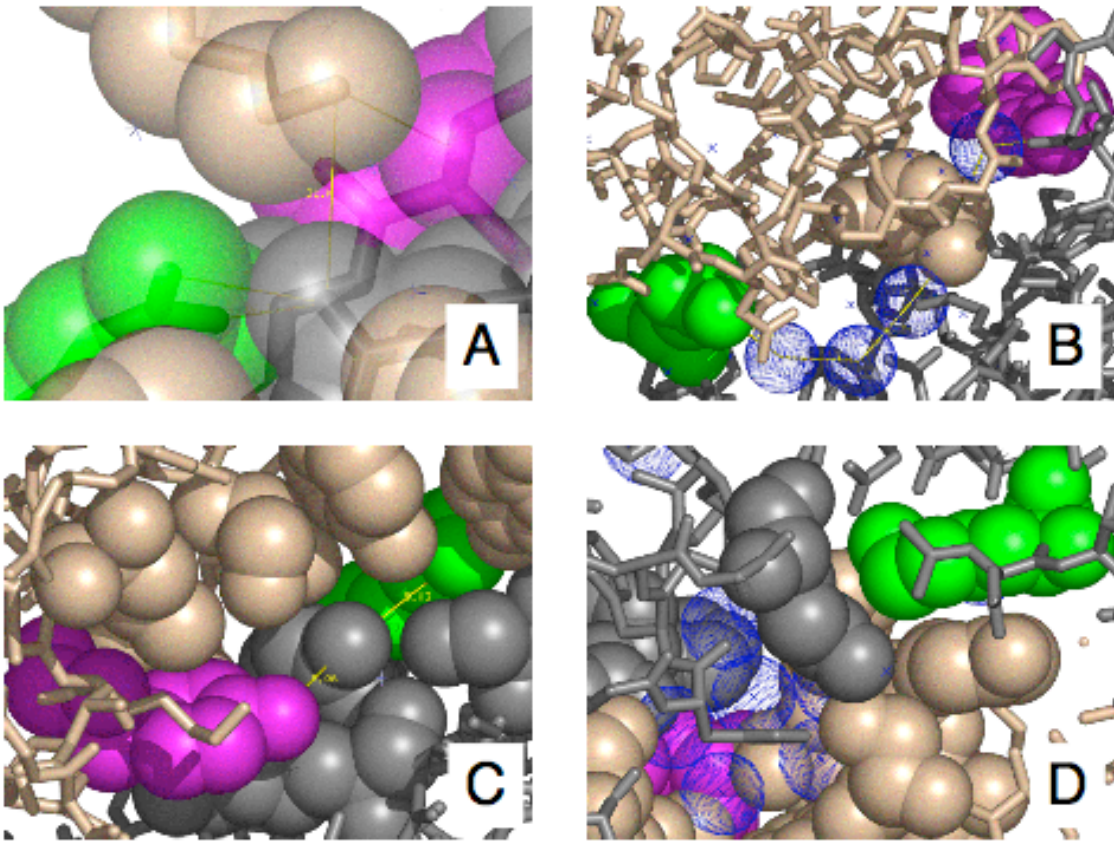

Figure 6. Indirect residue mediated contacts. Mutated residues from opposing are shown in green and pink space fill respectively. Other residues are colored dark grey and brown, respectively, and are either space fill or stick for clarity. Waters are displayed as purple dot surfaced spheres. Yellow lines depict direct distances between atoms, and distances are measure from atom centers. All images made using PyMol (Copyright © 2004 DeLano Scientific). A) chain C Arg83 to chain F Thr42 of 1brs are separated by 5 Å. B) chain C Arg59 to chain F Tyr29 of 1brs are separated by 15 Å. C) chain Y Arg21 to chain L Tyr50 of 3hfm are separated by 9.5 Å. D) chain B Lys40 to chain A Tyr437 of 1a4y are separated by 11 Å.

Table 3. Detailed predictive ability for identifying experimentally defined interactions

PDB ID	Chain 1	Residue		Chain 2	Residue		Mutant	Class	Mut.	QC	Cβ 6Å	Cent 9Å	Cl at 3Å	Cl at 9Å	Δ SASA
1brs	C	27	K	F	42	T	A/A	E	1.5	+	+	+	-	+	+
1brs	C	27	K	F	39	D	A/A	E	4.8	+	-	+	-	+	+
1brs	C	27	K	F	38	W	A/F	E	0.6	+	-	+	-	+	+
1brs	C	27	K	F	35	D	A/A	W	0.4	-	-	-	-	+	-
1brs	C	27	K	F	76	E	A/A	W	0.1	-	-	-	-	-	-
1brs	C	27	K	F	29	Y	A/A	W	0.2	-	-	-	-	-	-
1brs	C	27	K	F	80	E	A/A	W	0.4	-	-	-	-	+	-
1brs	C	59	R	F	42	T	A/A	W	0.2	-	-	+	-	+	-
1brs	C	59	R	F	38	W	A/F	V	0.6	+	-	+	-	+	+
1brs	C	59	R	F	35	D	A/A	V	3.4	+	+	+	+	+	+
1brs	C	59	R	F	76	E	A/A	V	1.7	+	-	+	+	+	+
1brs	C	59	R	F	29	Y	A/A	W	0.6	-	-	-	-	-	-
1brs	C	59	R	F	80	E	A/A	W	0.6	-	-	-	-	+	-
1brs	C	83	R	F	39	D	Q/A	V	6.7	+	-	+	+	+	+
1brs	C	83	R	F	38	W	Q/F	O	0.3	-	-	-	-	+	-
1brs	C	83	R	F	35	D	Q/A	O	0.3	-	-	-	-	+	-
1brs	C	83	R	F	29	Y	Q/A	V	0.6	+	-	+	+	+	+
1brs	C	83	R	F	42	T	Q/A	O	1.2	-	-	+	-	+	+
1brs	C	83	R	F	76	E	A/A	O	0.1	-	-	-	-	-	-
1brs	C	83	R	F	80	E	Q/A	O	0.2	-	-	-	-	-	-
1brs	C	87	R	F	38	W	A/A	O	0.2	-	-	-	-	+	-
1brs	C	87	R	F	39	D	A/A	E	6.1	+	-	+	-	+	+
1brs	C	87	R	F	29	Y	A/A	E	1	+	-	+	-	+	+
1brs	C	87	R	F	42	T	A/A	O	0.4	-	-	-	-	+	-
1brs	C	87	R	F	76	E	A/A	O	0.1	-	-	-	-	-	-
1brs	C	87	R	F	80	E	A/A	O	0	-	-	-	-	-	-
1brs	C	102	H	F	39	D	A/A	V	4.9	+	-	-	+	+	+
1brs	C	102	H	F	42	T	A/A	O	-0.1	-	-	-	-	+	-
1brs	C	102	H	F	38	W	A/F	O	0.2	-	-	-	-	+	-
1brs	C	102	H	F	76	E	A/A	O	0	-	-	-	-	-	-
1brs	C	102	H	F	80	E	A/A	O	0.1	-	-	-	-	-	-
1brs	C	102	H	F	29	Y	A/A	E	3.3	+	+	+	-	+	+
1brs	C	102	H	F	29	Y	A/F	E	0.5	+	+	+	-	+	+
1brs	C	73	E	F	39	D	A/A	E	2.9	+	-	-	-	+	+
1a4y	B	5	R	A	434	Y	A/A	O	1.4	-	-	+	-	-	-
1a4y	B	5	R	A	435	D	A/A	W	0.9	-	-	+	-	-	-
1a4y	B	40	K	A	434	Y	G/A	E	0.2	+	+	+	-	+	+
1a4y	B	40	K	A	435	D	G/A	E	-2.6	+	-	+	+	+	+
1a4y	B	40	K	A	437	Y	G/A	O	2.4	-	-	+	-	-	-

Change in binding free energy when mutated, Contact, no contact, QContacts, centroid and closest atom are abbreviated as Mut., "+", "-", QC, Cent and Cl respectively.

Table 3. Continued…

PDB ID	Chain 1	Residue		Chain 2	Residue		Mutant	Class	Mut.	QC	Cβ 6Å	Cent 9Å	Cl at 3Å	Cl at 9Å	SASA
1vfb	C	121	Q	A	32	Y	A/A	E	2	+	-	+	-	+	+
1vfb	C	18	D	A	50	Y	A/A	V	-0.4	+	-	+	+	+	+
1vfb	C	119	D	A	50	Y	A/A	W	0.3	-	-	-	-	-	-
1vfb	C	121	Q	A	92	W	A/A	E	2.7	+	-	+	-	+	+
1vfb	C	124	I	A	32	Y	A/A	E	0	+	-	+	-	+	+
1vfb	C	124	I	A	92	W	A/A	E	0.7	+	-	+	-	+	+
1vfb	C	125	R	A	92	W	A/A	E	1.7	+	-	+	-	+	+
1vfb	C	129	L	A	92	W	A/A	W	0.2	-	-	-	-	+	-
1vfb	C	116	K	B	32	Y	A/A	E	0.2	+	-	+	-	+	+
1vfb	C	119	D	B	52	W	A/A	V	-0.3	+	-	+	-	+	+
1vfb	C	118	T	B	54	D	A/A	E	0.6	+	+	+	-	+	+
1vfb	C	24	S	B	100	D	A/A	V	0.3	+	+	+	+	+	+
1vfb	C	119	D	B	101	Y	A/F	V	-0.1	+	-	+	-	+	+
1vfb	C	120	V	B	101	Y	A/F	E	0	+	-	+	-	+	+
3hfm	Y	96	K	L	31	N	A/A	V	4.7	+	-	+	+	+	+
3hfm	Y	96	K	L	50	Y	A/A	V	3.8	+	-	+	-	+	+
3hfm	Y	21	R	L	96	Y	A/A	V	-1.9	+	-	+	+	+	+
3hfm	Y	20	Y	L	50	Y	A/F	E	1	+	-	+	-	+	+
3hfm	Y	21	R	L	50	Y	A/A	O	-0.7	-	-	-	-	-	-
3hfm	Y	97	K	L	50	Y	A/A	E	3.5	+	-	-	-	+	+
3hfm	Y	21	R	H	50	Y	A/A	V	0.5	+	-	-	+	+	+
3hfm	Y	96	K	H	98	W	A/A	E	4.8	+	-	+	-	+	+
3hfm	Y	97	K	H	32	D	A/A	E	3.5	+	-	-	-	+	+
3hfm	Y	97	K	H	33	Y	A/A	V	5	+	-	-	-	+	+
1dqj	A	32	N	C	96	K	A/A	V	4.4	+	-	+	+	+	+
1dqj	A	91	S	C	21	R	A/A	E	-0.5	+	-	+	-	+	+
1dqj	A	91	S	C	20	Y	A/A	E	1.1	+	-	+	-	+	+
1dqj	A	96	Y	C	21	R	A/A	E	-1.1	+	-	+	-	+	+
1dqj	A	96	Y	C	100	S	A/A	E	1	+	-	+	-	+	+
1dqj	B	32	D	C	97	K	A/A	V	3	+	-	+	+	+	+
1dqj	B	53	Y	C	62	W	A/A	E	0.7	+	-	+	-	+	+
1dqj	B	53	Y	C	63	W	A/A	E	0.3	+	-	+	-	+	+
1dqj	B	53	Y	C	75	L	A/A	V	1.5	+	+	+	-	+	+
1dqj	B	53	Y	C	101	D	A/A	V	-0.2	+	+	+	-	+	+
1dqj	B	98	W	C	100	S	A/A	E	0.4	+	-	+	-	+	+
1dqj	B	98	W	C	97	K	A/A	E	1.8	+	-	+	-	+	+
1dqj	B	98	W	C	20	Y	A/A	E	3.1	+	-	+	-	+	+

Change in binding free energy when mutated, Contact, no contact, QContacts, centroid and closest atom are abbreviated as Mut., "+", "-", QC, Cent and Cl respectively.

For automatic methods, contact identification is focused on the direct class. Therefore, these methods try to minimize categorizing no contact interactions as false positives and thereby also minimizing the total number of interactions. As the cause of indirect interactions is more complex, finding them based solely on structural data presents a considerable challenge to any computational analysis, and is outside of the scope of this assessment provided in this work. Table 2 summarizes how well the QContacts method compares against a SASA method and the three radial methods at three distance cutoffs: closest atom, centroid, and Cβ over the distances of 3 Å, 6 Å, and 9 Å (see Methods and Materials for a detailed description of the three radial methods) in identifying the above experimentally determined protein interface contacts. Out of a total 388 interactions, the QContacts method finds all 49 direct contacts and none of the indirect or no contacts. Importantly, the QContacts method is able to find the direct contacts without any false positives from the no contact class and with the minimum number of total contacts in comparison to the other methods. The SASA method performs almost as well by finding all 49 direct contacts, no false positives (no contacts class), but about 25% more total interactions. This method does find one of the indirect residue-mediated interactions. In general, radial cutoff methods do not perform as well at finding direct contacts, where the closest atom seems to do the best. However, one advantage of the radial cutoff approach is that they calculate one order of magnitude faster than the QContacts method. Because the Cβ method puts the interaction center so close to the peptide backbone, it requires a large cutoff at 6 Å to begin to find interactions and even at a 9 Å cutoff finds just over 50% of the direct contacts with over

twice as many total contacts than the QContacts method, which can potentially introduce many more false positives. While the centroid method places the interaction center slightly farther out, the results follow the same trend as the Cβ method, but are somewhat better. Again, even at a 9 Å cutoff, the centroid radial method does not include all of the direct contacts and even finds a false positive (contact differences are shown in Figure 7A). In contrast, the closest atom radial method using a 6 Å cutoff mimics the results from the SASA approach, but produces three times as many total contacts as the QContacts method. These differences can also be seen in the comparison of interface contacts found using QContacts and the closest atom 6 Å method shown in Figure 7B. The similarity of the 6 Å closest atom and SASA methods is expected since the SASA probes with atom specific radial cutoffs between 5.6 Å and 6.54 Å is approximately the same cutoff as a constant 6 Å value. These results serve to illustrate the problem in using a regular sphere of a radial cutoff to approximate the shape of quite irregular residues. Overall, this comparison with experimental double cycle mutants shows that the QContacts method and the SASA methods prove to be the best at finding true positives (direct class) without including false positives (no contact class; Figures 4 and 5 show examples of false positives). Of the radial methods, only the closest atom at 6 Å came close to performing as well. For a more in-depth comparison of this data, individual mutations with their energies are provided in Table 3. The radial cutoff methods are seen to overestimate the binding interface in all cases when compared to QContacts towards the protein interior, as seen by the open blue mesh in Figure 7.

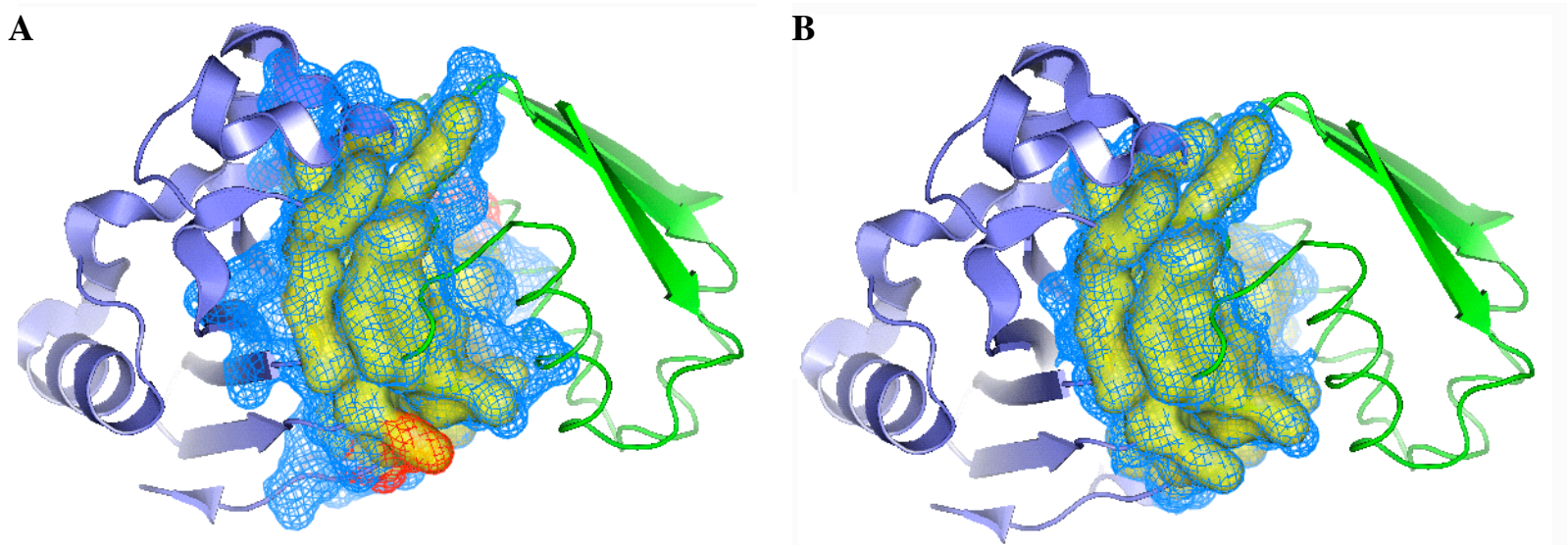

Figure 7. Comparison of radial cutoff methods to QContacts.
Barnase is shown in purple cartoon and Barstar in green cartoon. Contacts identified using QContacts are shown in a yellow surface. Contacts identified using the radial cutoff method are indicated by a mesh surface. The contacts found using the radial cutoff method but not QContacts are seen as a blue mesh. Contacts that were found using QContacts only is shown in red. A) This is a comparison of Centroid 9 Å and QContacts. B) Here the comparison of Closest atom 6 Å is shown.

Identifying pair-wise contacts

As the experimental mutagenesis set is small, a more thorough exploration of the differences between the QContacts method and the radial cutoff methods was performed using the expanded dataset of 592 protein pairs (the structures are listed in Methods and Materials). Table 4 shows the number of residue pairs identified in the binding interfaces identified by the different methods. In our discussion, we will use the 57,052 pair-wise interactions found by QContacts method as a reference since the algorithm used precisely identifies vdW overlaps and water exclusion contacts. While we do not suggest that all of these interactions have a significant energetic contribution to binding, the mutagenesis data from the previous section suggests that many of these contacts are important.

Once again, the trends between the radial methods are similar to the results from the identification of experimentally found protein–protein contacts. The approximation of a residue by a singular point provides such a decrease in the resolution of the Cβ and centroid methods that they require larger radial cutoffs of 6 Å and 9 Å to find the vdW and water exclusion contacts. Even at a large cutoff of 9 Å, there still are vdW and water exclusion contacts that these methods did not find (see column ”QContacts only” in Table 4). This suggests that many interacting residues in the interface do not interdigitate, but rather interact head on. Also, as an unwanted result of using such large cutoffs, many extra potentially non-specific contacts are identified (see column “Radial only” in Table 4), which could produce many misleading false positives. For both the Cβ and centroid methods, increasing the radial cutoff to get all of the vdW overlaps and

water exclusion contacts would produce even more false positives, which would make finding significant interactions difficult in all the noise added by the extra, spurious interactions. As expected, of the three radial methods, the closest atom method finds the most residue contacts at each cutoff of all the radial methods and is much better at finding the vdW overlaps and water exclusion interactions.

Table 4. Comparison of QContacts to radial cutoff methods: pair-wise residue contacts

	Method	Total	Both [a]	QContacts only	Radial only
	QContacts	57,052	—	—	—
Closest atom	3 Å	4,667	4,664	52,388	3
	6 Å	70,113	55,670	1,382	14,443
	9 Å	253,377	57,052	0	196,325
Centroid	3 Å	25	25	57,027	0
	6 Å	23,296	23,248	33,804	48
	9 Å	85,636	52,490	4,526	33,146
Cβ	3 Å	3	3	57,049	0
	6 Å	14,736	14,676	42,376	60
	9 Å	73,016	46,078	10,974	26,938

The number of residue interactions across a protein interface are listed.
[a] Residue pair interactions identified by both the QContacts method and the respective radial method as contacting across a protein interface.

In the previous section, the closest atom method using a 6 Å cutoff performed the best of the radial methods by finding all the direct contacts, but produced three times more interactions than the QContacts method (see Table 4). In this larger set, closest atom method using a 6 Å cutoff produced only about 25% more contacts, but actually missed 1382 water exclusion contacts or about 2 per interface (Table 4). We know these

must be water exclusion contacts because standard vdW radii sets for protein atoms do not go over 2 Å (Tsai et al., 1999). To not to be within a 6 Å radius would suggest a radius of 3 Å or more for the atoms. The QContacts method finds these water exclusion interactions that are over 6 Å apart are reasonable. The largest atom in the radii set used is 1.87 Å for sp^3 carbon atoms and two of these with a 2.8 Å water diameter are over the 6 Å limit. At a 9 Å cutoff, the closest atom method finds all the interactions QContacts does as seen in Table 4, but in a total of 253,377 residue contacts. This is over three times more interactions or about 332 extra contacts per binding interface, which increases the likelihood of including false positives in an analysis.

Effect on pair-wise contact frequencies

Because these radial methods are used so often to calculate pair-wise frequencies of individual amino acid pairs for use in statistical scoring functions (Fischer et al., 1995; Krippahl et al., 2003; Moont et al., 1999; Murphy et al., 2003; Palma et al., 2000), the propensity of interactions between amino acid pairs in binding interfaces was calculated using the QContacts method (for values see Table 5) and compared to the frequencies of all three radial cutoff methods. Figure 8 shows difference plots for each of the radial cutoff methods compared to the QContacts approach. The value of the radial cutoff was chosen based on the closest performing results from Table 4.

Table 5. Amino acid propensities

	R	K	N	D	Q	E	H	P	Y	W	S	T	G	A	M	C	F	L	V	I
R	.0117	.0087	.0098	.0175	.0102	.0160	.0094	.0100	.0093	.0094	.0094	.0093	.0116	.0085	.0075	.0082	.0075	.0078	.0086	.0077
K	.0087	.0144	.0105	.0169	.0108	.0191	.0085	.0088	.0103	.0083	.0106	.0095	.0105	.0083	.0079	.0071	.0067	.0065	.0080	.0071
N	.0098	.0105	.0195	.0123	.0112	.0086	.0097	.0114	.0108	.0105	.0111	.0107	.0096	.0100	.0077	.0073	.0086	.0069	.0075	.0077
D	.0175	.0169	.0123	.0133	.0112	.0076	.0143	.0077	.0103	.0090	.0106	.0095	.0102	.0094	.0058	.0056	.0062	.0061	.0067	.0074
Q	.0102	.0108	.0112	.0112	.0167	.0099	.0090	.0108	.0100	.0089	.0103	.0096	.0118	.0088	.0075	.0078	.0090	.0090	.0085	.0075
E	.0160	.0191	.0086	.0076	.0099	.0137	.0118	.0093	.0093	.0064	.0111	.0090	.0085	.0089	.0082	.0083	.0077	.0070	.0082	.0072
H	.0094	.0085	.0097	.0143	.0090	.0118	.0188	.0089	.0100	.0097	.0095	.0097	.0109	.0114	.0080	.0104	.0087	.0091	.0084	.0080
P	.0100	.0088	.0114	.0077	.0108	.0093	.0089	.0148	.0122	.0124	.0102	.0098	.0106	.0088	.0100	.0119	.0093	.0087	.0091	.0091
Y	.0093	.0103	.0108	.0103	.0100	.0093	.0100	.0122	.0127	.0092	.0110	.0088	.0102	.0092	.0095	.0111	.0099	.0089	.0095	.0095
W	.0094	.0083	.0105	.0090	.0089	.0064	.0097	.0124	.0092	.0207	.0083	.0112	.0110	.0095	.0117	.0125	.0130	.0091	.0078	.0111
S	.0094	.0106	.0111	.0106	.0103	.0111	.0095	.0102	.0110	.0083	.0159	.0102	.0093	.0097	.0082	.0117	.0074	.0088	.0098	.0079
T	.0093	.0095	.0107	.0095	.0096	.0090	.0097	.0098	.0088	.0112	.0102	.0162	.0116	.0094	.0088	.0080	.0100	.0088	.0097	.0096
G	.0116	.0105	.0096	.0102	.0118	.0085	.0109	.0106	.0102	.0110	.0093	.0116	.0153	.0094	.0092	.0109	.0077	.0074	.0084	.0091
A	.0085	.0083	.0100	.0094	.0088	.0089	.0114	.0088	.0092	.0095	.0097	.0094	.0094	.0168	.0112	.0108	.0111	.0098	.0104	.0102
M	.0075	.0079	.0077	.0058	.0075	.0082	.0080	.0100	.0095	.0117	.0082	.0088	.0092	.0112	.0251	.0104	.0118	.0125	.0122	.0115
C	.0082	.0071	.0073	.0056	.0078	.0083	.0104	.0119	.0111	.0125	.0117	.0080	.0109	.0108	.0104	.0606	.0114	.0102	.0083	.0081
F	.0075	.0067	.0086	.0062	.0090	.0077	.0087	.0093	.0099	.0130	.0074	.0100	.0077	.0111	.0118	.0114	.0180	.0120	.0115	.0139
L	.0078	.0065	.0069	.0061	.0090	.0070	.0091	.0087	.0089	.0091	.0088	.0088	.0074	.0098	.0125	.0102	.0120	.0199	.0126	.0130
V	.0086	.0080	.0075	.0067	.0085	.0082	.0084	.0091	.0095	.0078	.0098	.0097	.0084	.0104	.0122	.0083	.0115	.0126	.0177	.0122
I	.0077	.0071	.0077	.0074	.0075	.0072	.0080	.0091	.0095	.0111	.0079	.0096	.0091	.0102	.0115	.0081	.0139	.0130	.0122	.0191
Ave	.0099	.0099	.0101	.0099	.0099	.0098	.0102	.0102	.0101	.0105	.0101	.0100	.0102	.0101	.0102	.0120	.0101	.0097	.0098	.0098
SD	.0026	.0033	.0027	.0035	.0020	.0032	.0025	.0017	.0010	.0030	.0018	.0017	.0018	.0018	.0040	.0116	.0029	.0031	.0025	.0030

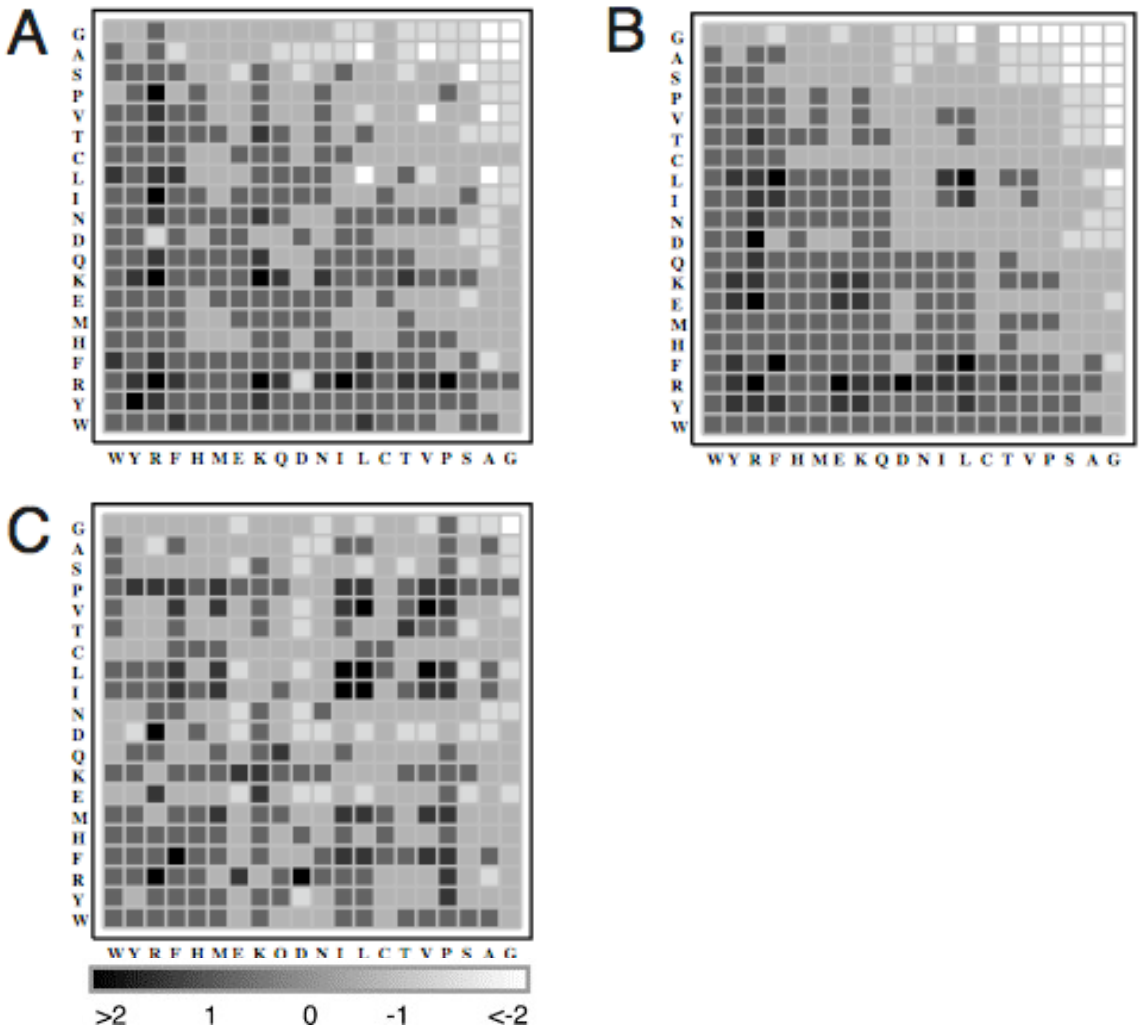

Figure 8. Differences in pair-wise contact frequencies: the QContacts method against radial cutoff methods. The difference in propensities of residue-to-residue contacts as determined by QContacts and three radial cutoff methods were calculated and plotted in a 20-by-20 amino acid matrix. To simplify the numbers graphically, plots were colored according to the number of standard deviations that the difference between the QContacts method and the radial cutoff method value fell or a *Z* score. The standard deviation was calculated individually for each plot from the calculated 400 difference values. The gray-scale color scheme goes from black indicating overestimation of over 2 standard deviations to white indicating underestimation of over 2 standard deviations. Fifty percentage gray indicates that the propensities were approximately the same. In all plots, amino acids are roughly ordered according to size and are abbreviated using the one letter code. (A) The difference in pair-wise contact frequencies between the QContacts method and the 6 Å centroid radial cutoff method. (B) QContacts method and the 9 Å Cβ radial cutoff method. (C) QContacts method and the 6 Å closest atom radial cutoff method.

Generally, the 6 Å centroid and 9 Å Cβ methods (Figure 8A and B, respectively) overestimate pair-wise contact frequencies for the larger amino acids like Trp, Tyr, and Arg (darker squares in the bottom left hand corners of the plots) and underestimate them for the smaller amino acids like Ala and Gly (lighter to white squares in the upper right corner). More symmetric amino acids like Leu and Val exhibit similar frequencies in comparison to the QContacts method. Again, these differences are attributable to the placement of the center of the probe radius for the centroid and Cβ methods. In addition, placing the center at the centroid potentially underestimates main-chain-to-main-chain contacts because the distance between the centroid and the main-chain is proportional to the size of the side-chain. While this situation could occur for amino acids with longer side-chains, it seems to happen frequently with His-to-His, Met, and Glu as well as Met-to-Met contacts. The effect is that these pair-wise frequencies are underestimated in the centroid method such that they are similar to the frequencies found by QContacts. The Cβ method measures distances between the first atoms of two residues' side-chains. This approximation avoids the centroid method's main-chain-to-main-chain underestimation, but leads to a problem where the side-chains are close despite the fact that the Cβ atoms are not. To compensate, the radius must be so large that contacts between larger amino acids are more consistently overestimated (darker squares in the lower left corner of Figure 8B) based purely on the size of the side-chain. Due to the asymmetry of side-chains, using a single point to represent an amino acid produces pair-wise contact frequencies that favor contact of larger amino acids and disfavor smaller ones. This inherent bias by the centroid and Cβ methods using such a

large radial cutoff would be detrimental to statistical potential functions. When comparing the frequencies calculated using the 6 Å closest atom method and QContacts, there are far fewer differences (50% gray squares in Figure 8C) because distances are measured between the residues' atoms, which provide better resolution. Even so, the trend for larger amino acids to overestimate interactions still holds especially for Trp and Phe. Surprisingly, the smaller and more symmetric Leu seems to be overestimated as well as the Ile and Pro. The underestimation holds true for Gly but is also evident for interactions with Asp. Much of this can be attributed to packing of side-chains in protein interfaces (Li and Nussinov, 1998; Lo Conte et al., 1999). While not as drastic as the other radial cutoff methods, the 6 Å closest atom method produces pair-wise frequencies that would produce inconsistencies in scoring functions.

For a more direct comparison of QContacts and the closest atom method, contact maps are shown for the coiled coil domain of the GCN4 leucine zipper 2zta (O'Shea et al., 1991) from the output of QContacts and the closest atom method at each of three cutoff values (Figure 9). These contact maps provide a tractable representation of the various approaches' ability to identify specific contacts, because they provide a quick, visual determination of asymmetrical regions along an interface, in this case a homodimer. The diagonal provides a line of symmetry across the plots. When examined more closely, asymmetric atomic contacts can be found in all the plots, but the points of asymmetry are less accurate using radial cutoff methods than QContacts. These slight differences in symmetry have been seen as dynamic regions in the binding interface of leucine zippers and are important in determining affinity (Junius et al., 1995).

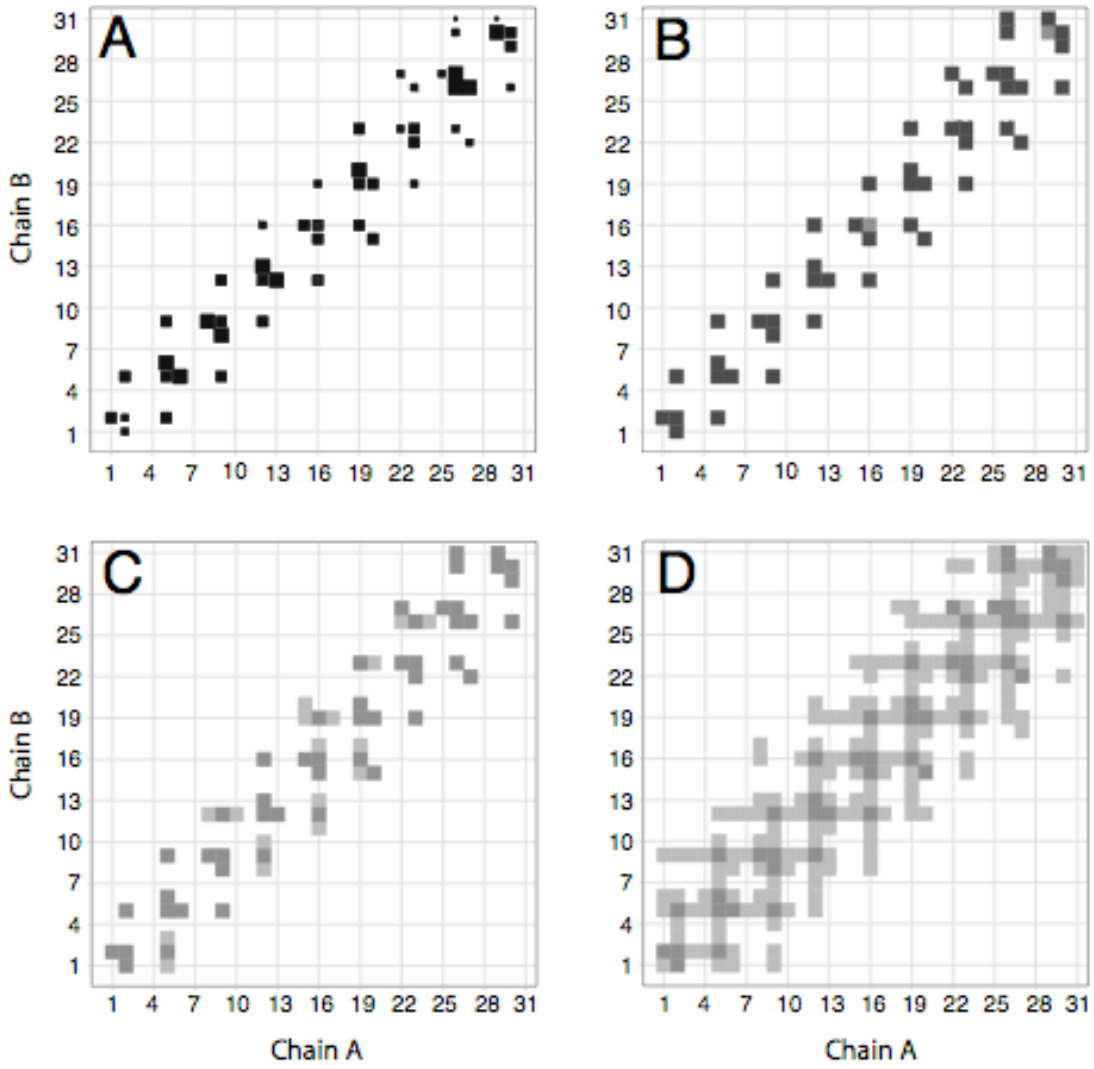

Figure 9. Direct comparison of QContacts to the closest atom radial cutoff method. Contact maps for the leucine zipper structure 2zta (O'Shea et al., 1991)are shown, where the x-axis is the residues on the A chain and the y-axis is the residues on the B chain. A) QContacts analysis showing only the output from the Voronoi polyhedra based analysis. Contact between two residues is indicated by a square, and the size of the square directly relates to the magnitude of the residue-residue interaction. Parts B), C) and D) are comparisons of the QContacts to the closest atom radial method at 3, 6 and 9 Å cutoffs, respectively. The sizes of the squares are the same as all the radial cutoff methods are a binary indicator of an interaction. In these three parts for each cutoff, if the contact is found only by QContacts, the square is colored dark grey. If both QContacts and the closest atom method find the contact, then the square is colored medium grey. If the contact is found only by the closest atom method, the square is colored light grey.

Figure 9A displays the QContacts map, which differs from the remaining plots in that the size of the square directly relates to the magnitude of the residue-residue interaction. In the remaining plots, the size of the square does not vary since the radial methods can only produce a binary identification of interactions. Figures 8B-D compare the results of the QContacts analysis to the various radial cutoffs used by the closest atom method, where dark grey indicates that only QContacts finds that interaction, medium grey indicates both, and light grey are contacts found only by the closest atom method. As expected, the 3 Å cutoff finds few contacts (2 medium grey squares in Figure 9B), all of which are in the set of interactions found by QContacts. At 3 Å, the closest atom method clearly underestimates the number of vdW and water exclusion contacts. Using the 6 Å cutoff, the closest atom method (Figure 9B) is the most similar to the QContacts' map than the any of the other cutoffs used. This 6 Å closest atom method finds all the QContacts interactions (medium grey squares in Figure 9C), but includes 17 or about 28% more interactions than QContacts. These 17 non-vdW or water exclusion interactions could contribute to the binding interface through conformational changes, but they also could be false positives. Figure 9D shows the clear overestimation of pair-wise interactions by the closest atom method with a 9 Å cutoff when compared to QContacts. The 9 Å closest atom method identifies over 4 times as many pair-wise interactions as QContacts, and most likely the majority of these are false positives. In addition, the excess interactions would hinder any useful analysis of the binding interface by obscuring real interactions.

Conclusion

In this work, we have shown that the QContacts method provides an accurate analysis of a protein interface for vdW overlaps and water excluded interactions. Compared to the ΔSASA method, the QContacts method exactly calculates a surface area of contact more similar to a molecular surface, whereas the ΔSASA method inflates the interaction surface area by a value related to a water radius. In addition, the ΔSASA method misses knob-in-hole contacts. While only 1% of residues are incorrectly identified as not participating in the binding interface (Table 1), ΔSASA will miss on average 10% of the atom contacts because of these knob-in-hole contacts. Potentially, this would underestimate the real surface area of interaction. In comparison to the radial methods, our results suggest that neither the centroid nor the Cβ methods possess the resolution necessary to accurately describe the pair-wise interactions in the protein interface. The best of these methods uses the closest atoms. Even so, because of the asymmetry of a united atom model as well as amino acid side-chains, such radial cutoff methods cannot find an appropriate cutoff value that finds direct vdW and water exclusion contacts without either underestimating (producing false negatives) with smaller radii or overestimating (producing false positives) with larger radii (Table 2 and Table 3). The inaccuracies of the radial cutoff based methods are even more evident in the pair-wise frequencies shown in Figure 8, which would incorrectly bias a statistical potential function. In addition, their poor resolution prevents these radial approaches from correctly finding unique asymmetries in homodimeric interfaces (data not shown). These slight differences in symmetry have been seen as dynamic regions in the binding

interface of leucine zippers and are important in determining affinity (Junius et al., 1995). The method most similar to the QContacts method for finding pair-wise interactions is the SASA. As shown in Table 2, the SASA method also finds all the experimentally significant, direct contacts as well as one of the indirect contacts. This SASA does find slightly more total contacts. For all of these methods, the overestimation of contacts would add bias to a statistical scoring function currently used in protein-docking programs. Our implementation of Voronoi polyhedra for finding contacts has been incorporated into our analysis program of protein interfaces called QContacts, which also finds hydrogen bonds, ionic pairs, and salt bridges. A web-based version of QContacts is available for use at http://tsailab.tamu.edu.ezproxy.tamu.edu:2048/Qcons.

CHAPTER III

CONCLUSIONS

QContacts is a robust program for identifying protein-protein interactions in atomic detail. The most rigorous analysis of protein interactions is provided by site-directed mutagenesis. We have used this technique as our standard to measure the performance of computational methods for identifying protein interface contacts. QContacts, closest atom radial cutoff and the ΔSASA methods performed well at picking out direct contacts, however QContacts was the most accurate in excluding false positives as exhibited in Table 2. Table 2 also demonstrated that the Centroid method severely overestimated contacts while the Cβ method greatly underestimated contacts. These inaccuracies can be more easily seen in the bias of amino acid frequencies as seen in Figure 8. ΔSASA lacks in identifying pair-wise and knob-in-hole atomic contacts (Figure 3) which prohibits an accurate picture of atomic packing in the interface. Accuracy is especially important in identifying different features in closely related proteins because it allows for the identification of contacts specific to its interface family. Accuracy is also critical in the analysis of protein binding interfaces. Such analysis include identify features such as hydrophobicity, size and amino acid content that are used to differentiate between non-specific crystal packing and specific dimerization interactions from crystal structures (Bahadur et al., 2004; Thornton, 1996). Other binding interface features such as residue-residue frequencies are used in pair-wise contact potentials that predict how two proteins interact (Fischer et al., 1995; Krippahl et al., 2003; Moont et al., 1999; Murphy et al., 2003; Palma et al., 2000). Accurately

identifying protein interactions is needed to create accurate pair-wise contact potentials. QContacts was designed as a tool to achieve a highly accurate description of protein interactions at the atomic level. QContact's can be used to improve analysis of protein binding interfaces by reducing the bias associated with inaccuracies in identifying contacts.

Future Work

"Very conserved residues are very important" (Oliveira et al., 2003) has become a central theme to most predictive measures of protein function. To identify structurally conserved residues in a protein interface an accurate interface alignment method is needed. Current methods for identifying contacting atoms in the binding interface are rough estimations and in turn this inaccuracy limits the capabilities of structural alignment. There are two problems with aligning protein interfaces when using the current methods: 1) macromolecular complexes are generally too large to perform numerous alignments for large scale analysis or multiple alignments and 2) interfaces need to be precisely identified in order to create an accurate alignment. These methods that attempt to align protein interfaces are not true interface alignments due to the inability to accurately define the interface. Instead these programs were developed to align the overall structure assuming that this would also align the interfaces (Aloy et al., 2003; Hu et al., 2000; Aytuna et al., 2005). The most basic method is a pseudo interface alignment method in which the alignment is independent of the interface residues. Instead, domain centers are aligned and the RMDS of these centers after transformation

are measured (iRMSD) (Aloy et al., 2003). However, two proteins may not have similar interfaces regardless of similar sequence or fold. Thus the assumption that an aligned structure or sequence is representative of the aligned interface becomes inaccurate. Another method performs a structure alignment independent of the interface (Aytuna et al., 2005). Because this method aligns the entire structure of all four proteins in any given alignment of two protein complexes it is very computationally expensive and would not be suited for large scale or multiple alignments. The third method for aligning interfaces uses select residues in the interface to align (Hu et al., 2000). Residues are identified as in-contact using an arbitrary radial cutoff and only those residues that participate in hydrogen bonds or hydrophobic contacts are considered. This negates many interacting residues that would be described as energetically repulsive as well as a handful of attractive interactions such as pi-cation and the alphatic chain of Arg and Lys interacting with hydrophobic residues. One or both of the interface alignment caveats described above limit these alignment methods.

A prospect for QContacts is its implementation in protein interface alignment. QContacts's potential to improve interface alignment programs lies in its accuracy. Essentially some of the less accurate methods rely on the overall structure to perform the alignment since they can not identify the interfaces well enough to properly align them. QContacts accurately identifies the interface so that a true interface alignment can be achieved. Restricting the alignment to only interface residues also leads to less computational power needed to align proteins based on their interface. This

advancement in interface alignments would allow for more and larger molecules in multiple interface alignments.

REFERENCES

Aloy, P., Ceulemans, H., Stark, A., Russell, R.B., 2003. The relationship between sequence and interaction divergence in proteins. J Mol Biol 332, 989-998.

Aloy, P., Russell, R.B., 2004. Ten thousand interactions for the molecular biologist. Nat Biotechnol 22, 1317-1321.

Aytuna, A.S., Gursoy, A., Keskin, O., 2005. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. Bioinformatics 21, 2850-2855.

Bahadur, R.P., Chakrabarti, P., Rodier, F., Janin, J., 2003. Dissecting subunit interfaces in homodimeric proteins. Proteins 53, 708-719.

Bahadur, R.P., Chakrabarti, P., Rodier, F., Janin, J., 2004. A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 336, 943-955.

Barber, M.J., Neame, P.J., Lim, L.W., White, S., Matthews, F.S., 1992. Correlation of x-ray deduced and experimental amino acid sequences of trimethylamine dehydrogenase. J Biol Chem 267, 6611-6619.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res 28, 235-242.

Bhat, T.N., Bentley, G.A., Boulot, G., Greene, M.I., Tello, D., Dall'Acqua, W., Souchon, H., Schwarz, F.P., Mariuzza, R.A., Poljak, R.J., 1994. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. Proc Natl Acad Sci USA 91, 1089-1093.

Bogan, A.A., Thorn, K.S., 1998. Anatomy of hot spots in protein interfaces. J Mol Biol 280, 1-9.

Bordner, A.J., Abagyan, R., 2005. Statistical analysis and prediction of protein-protein interfaces. Proteins 60, 353-366.

Buckle, A.M., Schreiber, G., Fersht, A.R., 1994. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution. Biochemistry 33, 8878-8889.

Burmeister, W.P., Huber, A.H., Bjorkman, P.J., 1994. Crystal structure of the complex of rat neonatal Fc receptor with Fc. Nature 372, 379-383.

Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., Huang, E.S., 2004. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 13, 190-202.

Chen, C.Z., Shapiro, R., 1999. Superadditive and subadditive effects of "hot spot" mutations within the interfaces of placental ribonuclease inhibitor with angiogenin and ribonuclease A. Biochemistry 38, 9273-9285.

Chothia, C., 1974. Hydrophobic bonding and accessible surface area in proteins. Nature 248, 338-339.

Connolly, M.L., 1986. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. Biopolymers 25, 1229-1247.

Dall'Acqua, W., Goldman, E.R., Lin, W., Teng, C., Tsuchiya, D., Li, H., Ysern, X., Braden, B.C., Li, Y., Smith-Gill, S.J., Mariuzza, R.A., 1998. A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. Biochemistry 37, 7981-7991.

DeLano, W.L., 2002. Unraveling hot spots in binding interfaces: progress and challenges. Curr Opin Struct Biol 12, 14-20.

Fischer, D., Lin, S.L., Wolfson, H.L., Nussinov, R., 1995. A geometry-based suite of molecular docking processes. J Mol Biol 248, 459-477.

Fischer, T.B., Arunachalam, K.V., Bailey, D., Mangual, V., Bakhru, S., Russo, R., Huang, D., Paczkowski, M., Lalchandani, V., Ramachandra, C., Ellison, B., Galer, S., Shapley, J., Fuentes, E., Tsai, J., 2003. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics 19, 1453-1454.

Fischer, T.B., Holmes, J.B., Miller, I.R., Parsons, J.R., Tung, L., Hu, J.C., Tsai, J., 2006. Assessing methods for identifying pair-wise atomic contacts across binding interfaces. J Struct Biol 153, 103-112.

Gellatly, B.J., Finney, J.L., 1982. Calculation of protein volumes: an alternative to the Voronoi procedure. J Mol Biol 161, 305-322.

Glaser, F., Steinberg, D.M., Vakser, I.A., Ben-Tal, N., 2001. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 43, 89-102.

Hu, Z., Ma, B., Wolfson, H., Nussinov, R., 2000. Conservation of polar residues as hot spots at protein interfaces. Proteins 39, 331-342.

Janin, J., Chothia, C., 1990. The structure of protein-protein recognition sites. J Biol Chem 265, 16027-16030.

Jones, S., Thornton, J.M., 1996. Principles of protein-protein interactions. Proc Natl Acad Sci USA 93, 13-20.

Junius, F.K., Mackay, J.P., Bubb, W.A., Jensen, S.A., Weiss, A.S., King, G.F., 1995. Nuclear magnetic resonance characterization of the Jun leucine zipper domain: unusual properties of coiled-coil interfacial polar residues. Biochemistry 34, 6164-6174.

Kazmierkiewicz, R., Liwo, A., Scheraga, H.A., 2003. Addition of side chains to a known backbone with defined side-chain centroids. Biophys Chem 100, 261-280.

Krippahl, L., Moura, J.J., Palma, P.N., 2003. Modeling protein complexes with BiGGER. Proteins 52, 19-23.

Lee, B., Richards, F.M., 1971. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55, 379-400.

Li, A.J., Nussinov, R., 1998. A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. Proteins 32, 111-127.

Li, Y., Li, H., Smith-Gill, S.J., Mariuzza, R.A., 2000. Three-dimensional structures of the free and antigen-bound Fab from monoclonal antilysozyme antibody HyHEL-63 (,). Biochemistry 39, 6296-6309.

Li, Y., Urrutia, M., Smith-Gill, S.J., Mariuzza, R.A., 2003. Dissection of binding interactions in the complex between the anti-lysozyme antibody HyHEL-63 and its antigen. Biochemistry 42, 11-22.

Lo Conte, L., Chothia, C., Janin, J., 1999. The atomic structure of protein-protein recognition sites. J Mol Biol 285, 2177-2198.

Lu, H., Lu, L., Skolnick, J., 2003. Development of unified statistical potentials describing protein-protein interactions. Biophysical Journal 84, 1895-1901.

Ma, B., Elkayam, T., Wolfson, H., Nussinov, R., 2003. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 100, 5772-5777.

Mancini, A.L., Higa, R.H., Oliveira, A., Dominiquini, F., Kuser, P.R., Yamagishi, M.E., Togawa, R.C., Neshich, G., 2004. STING Contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. Bioinformatics 20, 2145-2147.

McConkey, B.J., Sobolev, V., Edelman, M., 2002. Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. Bioinformatics 18, 1365-1373.

Mendez, R., Leplae, R., De Maria, L., Wodak, S.J., 2003. Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins 52, 51-67.

Moont, G., Gabb, H.A., Sternberg, M.J., 1999. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 35, 364-373.

Murphy, J., Gatchell, D.W., Prasad, J.C., Vajda, S., 2003. Combination of scoring functions improves discrimination in protein-protein docking. Proteins 53, 840-854.

Norel, R., Lin, S.L., Wolfson, H.J., Nussinov, R., 1994. Shape complementarity at protein-protein interfaces. Biopolymers 34, 933-940.

Norel, R., Lin, S.L., Wolfson, H.J., Nussinov, R., 1995. Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. J Mol Biol 252, 263-273.

Norel, R., Petrey, D., Wolfson, H.J., Nussinov, R., 1999. Examination of shape complementarity in docking of unbound proteins. Proteins 36, 307-317.

O'Shea, E.K., Klemm, J.D., Kim, P.S., Alber, T., 1991. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. Science 254, 539-544.

Ofran, Y., Rost, B., 2003. Analysing six types of protein-protein interfaces. J Mol Biol 325, 377-387.

Oliveira, L., Paiva, P.B., Paiva, A.C., Vriend, G., 2003. Identification of functionally conserved residues with the use of entropy-variability plots. Proteins 52, 544-552.

Padlan, E.A., Silverton, E.W., Sheriff, S., Cohen, G.H., Smith-Gill, S.J., Davies, D.R., 1989. Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex. Proc Natl Acad Sci USA 86, 5938-5942.

Palma, P.N., Krippahl, L., Wampler, J.E., Moura, J.J., 2000. BiGGER: a new (soft) docking algorithm for predicting protein interactions. Proteins 39, 372-384.

Papageorgiou, A.C., Shapiro, R., Acharya, K.R., 1997. Molecular recognition of human angiogenin by placental ribonuclease inhibitor--an X-ray crystallographic study at 2.0 A resolution. Embo J 16, 5162-5177.

Pons, J., Rajpal, A., Kirsch, J.F., 1999. Energetic analysis of an antigen/antibody interface: alanine scanning mutagenesis and double mutant cycles on the HyHEL-10/lysozyme interaction. Protein Sci 8, 958-968.

Raschke, T.M., Tsai, J., Levitt, M., 2001. Quantification of the hydrophobic interaction by simulations of the aggregation of small hydrophobic solutes in water. Proc Natl Acad Sci USA 98, 5965-5969.

Richards, F.M., 1974. The Interpretation of Protein Structures: Total Volume, Group Volume Distributions and Packing Density. J Mol Biol 82, 1-14.

Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pichaud, M., Topf, M., Sali, A., 2004. A structural perspective on protein-protein interactions. Curr Opin Struct Biol 14, 313-324.

Schneidman-Duhovny, D., Inbar, Y., Polak, V., Shatsky, M., Halperin, I., Benyamini, H., Barzilai, A., Dror, O., Haspel, N., Nussinov, R., Wolfson, H.J., 2003. Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. Proteins 52, 107-112.

Schreiber, G., Fersht, A.R., 1995. Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. J Mol Biol 248, 478-486.

Sheriff, S., Hendrickson, W.A., Smith, J.L., 1987. Structure of myohemerythrin in the azidomet state at 1.7/1.3A resolution. Journal of Molecular Biology 197, 273-296.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., Edelman, M., 1999. Automated analysis of interatomic contacts in proteins. Bioinformatics 15, 327-332.

Thorn, K.S., Bogan, A.A., 2001. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 17, 284-285.

Thornton, S.J.a.J., 1996. Principles of Protein-Protein Interactions. Proc. Natl. Acad. Sci. USA 93, 13-20.

Tsai, J., Taylor, R., Chothia, C., Gerstein, M., 1999. The packing density in proteins: standard radii and volumes. J Mol Biol 290, 253-266.

Tsai, J., Gerstein, M., 2002. Calculations of protein volumes: sensitivity analysis and parameter database. Bioinformatics 18, 985-995.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., Yoshikawa, S., 1996. The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. Science 272, 1136-1144.

Voronoi, G.F., 1908. Nouvelles applications des paramétres continus á la théorie de formes quadratiques. J. Reine Angew. Math. 134, 198-287.

Wang, T., Wade, R.C., 2003. Implicit solvent models for flexible protein-protein docking by molecular dynamics simulation. Proteins 50, 158-169.

Wodak, S.J., Mendez, R., 2004. Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. Curr Opin Struct Biol 14, 242-249.

VITA

Name: **Tiffany Brink Fischer**
Address: 818 Water Oak Dr
Allen, TX 75002
Email: tifischer@gmail.com

Education:

Bachelor of Science, Genetics, Texas A&M University 2002
Master of Science, Biochemistry, Texas A&M University 2006

Experience:

08/03 - 12/06 Graduate Student
- 01/04 - 05/06 Teaching Assistant, TAMU Dept. of BICH
- 01/05 - 02/06 Web Administrator for http//:Tsailab.org

09/01 - 08/03 Research Assistant, TAMU Dept. of Biochemistry/Biophysics
- Curator for the Binding Interface Database (http//:Tsailab.org/BID)

09/01 - 01/02 Teaching Assistant, TAMU Dept. of Chemistry

05/00 - 01/01 Lab Technician, Sigma Genosys, The Woodlands, Texas
- DNA sequencing and purification

Conferences:

Brink, T.M., Bliss, R.B., and Tsai, J. W., 2002. A Detailed Protein-Protein Interaction Database. Sixteenth Symposium of the Protein Society.

Publications:

Fischer, TB, Arunachalam, KV, Bailey, D, Mangual, V, Bakhru, S, Russo, R, Huang, D, Paczkowski, M, Lalchandani, V, Ramachandra, C, Ellison, B, Galer, S, Shapley, J, Fuentes, E, and Tsai, J., 2003. The Binding Interface Database (BID): A Compilation of Amino Acid Hot Spots in Protein Interfaces. Bioinformatics 19, 1453-4.

Walker, JM. The Proteomics Protocols Handbook: Fischer, TB, Paczkowski, M., Zettel, M, and Tsai, J. A Guide to Protein Interaction Databases. March 2005. Humana Press.

Fischer TB, Holmes JB, Miller IR, Parsons JR, Tung L, Hu JC, Tsai J., 2006. Assessing methods for identifying pair-wise atomic contacts across binding interfaces. J Struct Biol 153, 103-12.

Rohl, C, Price, Y, Fischer, TB, Paczkowski, M, Zettel, MF, Tsai, J., 2006. Cataloging the Relationship Between Proteins. Mol Biotechnology 34, 69-93.