Winter 12-15-2009

Shortest Geometric Paths Analysis in Structural Biology

Ryan G. Coleman
University of Pennsylvania, ryan.g.coleman@gmail.com

Shortest Geometric Paths Analysis in Structural Biology

Abstract

The surface of a macromolecule, such as a protein, represents the contact point of any interaction that molecule has with solvent, ions, small molecules or other macromolecules. Analyzing the surface of macromolecules has a rich history but analyzing the distances from this surface to other surfaces or volumes has not been extensively explored. Many important questions can be answered quantitatively through these analyses. These include: what is the depth of a pocket or groove on the surface? what is the overall depth of the protein? how deeply are atoms buried from the surface? where are the tunnels in a protein? where are the pockets and what are their shapes? A single algorithm to solve one graph problem, namely Dijkstra's shortest paths algorithm, forms the basis for algorithms to answer these many questions. Many distances can be measured, for instance the distance from the convex hull to the molecular surface while avoiding the interior of the surface is defined as Travel Depth. Alternatively, the distance from the surface to every atom can be measured, giving a measure of the Burial Depth of given residues. Measuring the minimum distance to the protein surface for all points in solvent, combined with topological guidance, allows tunnels to be located. Analyzing the surface from the deepest Travel Depth upwards allows pockets to be catalogued over the entire protein surface for additional shape analysis. Ligand binding sites in proteins are significantly deep, though this does not affect the binding affinity. Hyperthermostable proteins have a less deep surface but bury atoms more deeply, forming more spherical shapes than their mesophilic counterparts. Tunnels through proteins can be identified, for the first time tunnels that are winding or bifurcated can be analyzed. Pockets can be found all over the protein surface and these pockets can be tracked through time series, mutational series, or over protein families. All of these results are new and for the first time provide quantitative and statistical verification of some previous hypotheses about protein shape.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Genomics & Computational Biology

First Advisor

Kim A. Sharp

Keywords

protein pockets depth holes shape geometry

Subject Categories

Computational Biology | Structural Biology | Theory and Algorithms

SHORTEST GEOMETRIC PATHS ANALYSIS IN STRUCTURAL BIOLOGY

Ryan G. Coleman

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2009

Supervisor of Dissertation: Kim A. Sharp, Ph.D., Associate Professor of Biochemistry and Biophysics

Kim A Sharp
____

Graduate Group Chairperson: Maja Bucan, Ph.D., Professor of Genetics

Maja Bucan
____

Dissertation Committee:

Warren Ewens, Ph.D., Professor of Biology

Jean Gallier, Ph.D., Professor of Computer and Information Science

Ravi Radhakrishnan, Ph.D., Assistant Professor of Bioengineering

Jeffrey G. Saven, Ph.D., Associate Professor of Chemistry

SHORTEST GEOMETRIC PATHS ANALYSIS IN STRUCTURAL BIOLOGY

Acknowledgement

The author wishes to thank friends, family, teachers, professors, mentors, thesis committee members, advisors, fellow students and lab members for the many excellent shared experiences in the past and yet to come.

The author particularly wishes to thank his advisor for offering an academic home for many years, providing scientific and non-scientific guidance, helping to curb my awkward sentence construction, handing out key advice at just the right moment, giving me a great rotation project that became the basis for this thesis, and providing generous financial support.

Financial support from the NSF (MCB02-35400) and NIH (GM48130) is gratefully acknowledged, as well as support from NIH Structural Biology Training Grant (GM008275) and the Computational Genomics Training Grant (T32HG000046) from the National Human Genome Research Institute.

ABSTRACT

SHORTEST GEOMETRIC PATHS ANALYSIS IN STRUCTURAL BIOLOGY

Ryan G. Coleman

Kim A. Sharp

The surface of a macromolecule, such as a protein, represents the contact point of any interaction that molecule has with solvent, ions, small molecules or other macromolecules. Analyzing the surface of macromolecules has a rich history but analyzing the distances from this surface to other surfaces or volumes has not been extensively explored. Many important questions can be answered quantitatively through these analyses. These include: what is the depth of a pocket or groove on the surface? what is the overall depth of the protein? how deeply are atoms buried from the surface? where are the tunnels in a protein? where are the pockets and what are their shapes? A single algorithm to solve one graph problem, namely Dijkstra’s shortest paths algorithm, forms the basis for algorithms to answer these many questions. Many distances can be measured, for instance the distance from the convex hull to the molecular surface while avoiding the interior of the surface is defined as Travel Depth. Alternatively, the distance from the surface to every atom can be measured, giving a measure of the Burial Depth of given residues. Measuring the minimum distance to the protein surface for all points in solvent, combined with topological guidance, allows tunnels to be located. Analyzing the surface from the deepest Travel Depth upwards allows pockets to be catalogued over the entire protein surface for additional shape analysis. Ligand binding sites in proteins are significantly deep, though this does not affect the binding affinity. Hyperthermostable proteins have a less deep surface but bury atoms more deeply, forming more spherical shapes than their mesophilic counterparts. Tunnels through proteins can be identified, for the first time tunnels that are winding or bifurcated can be analyzed. Pockets can be found all over the protein surface and these pockets can be tracked through time series, mutational series, or over protein families. All of these results are new and for the first time provide quantitative and statistical verification of some previous hypotheses about protein shape.

Table of Contents

List of Tables

List of Illustrations

Chapter 1

Introduction

The amazing property of macromolecules, especially proteins, is that they form precise three dimensional structures by folding. These three dimensional structures are the active forms which perform all the necessary biochemistry to maintain life in all forms. These structures have a specific shape, which along with other properties like charge determine the specific activity and function of each protein. This thesis examines new methods of analyzing the shape of these macromolecules and the results obtained from such methods.

Atomic Radii and Macromolecular Surfaces

There is a rich history of treating atoms as spheres and constructing surface models that model the solute/solvent boundary in structural biology. The van der Waals radius of an atom is a model that allows the size of atoms or molecules to be understood in terms of spheres that cannot overlap due to steric constraints. The intermolecular force that leads to this radius was postulated by Johannes Diderik van der Waals when he developed a model that showed liquids and gases could be made of the same matter, given that molecules existed and they had this finite size and some attraction to each other [1; 2]. For all work done in this thesis, the radii of the atoms involved (mainly the heavy atoms in biological molecules: carbon, nitrogen, oxygen, phosphorus and sulfur) were those previously shown to give good liquid and gas kinetic properties and critical densities and packing among other desirable properties [3].

The van der Waals radii are used to represent a macromolecule as a set of overlapping spheres in their specific position determined by how the macromolecule is folded. By choosing a probe to represent solvent, commonly sized between 1.2Å and 1.8Å, a surface can be constructed that represents the boundary between solute and solvent. The surface can be constructed from the center of the probe sphere, as it moves as close as possible to the macromolecule (the solvent accessible surface), or it can be constructed from the front of probe sphere (the molecular surface). An early review on the subject of these surfaces and the areas and volumes is by Richards [4]. Many other advances in surface generation and analysis have been forthcoming[5; 6; 7; 8; 9; 10; 11; 12]. As the probe radius varies, the position of normal protein atoms leads to a fractal surface [13]. Using these surfaces to analyze protein-water interactions has been reviewed by Levitt and Park [14] and Raschke [15]. Overall, any new analysis must be automated and fast to analyze the genomic scale data now present in the Protein Data Bank [16].

Many analyses have been done on various aspects of these surfaces, particularly examining the exposed surface area of the atoms. However, relatively little work has been done with methods relying on distances from these surfaces to other surfaces or features. This is likely due to the complicated nature of the molecular surface, it is like no other surface in nature as there are no straight lines, no flat shapes, and due in part to the fractal nature. Though some work has been done on measuring the distance of each atom to the surface (or to a surface atom) [17; 18; 19; 20; 21; 22], very little other work has been done owing to the complicated algorithmic nature of the problem. The distance of each atom to the surface is the simplest to implement, as there are no disallowed regions and it can be computed trivially by comparing the distance of each atom center to all surface atoms.

In this thesis, various different distances from and to this molecular surface are computed, using a grid representation and Dijkstra's shortest paths algorithm [23] to approximate the distances. This allows computation of the distance of the molecular surface from the convex hull while avoiding the molecular interior, a useful construction that allows computation of what is called Travel Depth throughout this work. This allows for the first time the depth of pockets on the protein surface to be computed. Also, the distance of each atom from the molecular surface can be computed within this framework. Finally, the distance from the molecular surface into solvent can be computed, leaving ridges of maximal distance in the solvent that can be exploited along with topological guidance to find tunnels all the way through these surfaces.

In the rest of this introduction, some background on the computational geometry and graph theory techniques used is given, followed by some background on the various application areas to be examined along with a brief preview of the methods and results.

Computational Geometry and Graph Theory

The exact methods of constructing surfaces used for this work uses one of two methods, either a gaussian approximation method designed to mimic the reentrant molecular surface[24] or a variation of the inkblot algorithm that colors grid points within the van der Waals plus probe radius and then erases those within the probe radius of the surface to model the reentrant probe surface. Both methods use a grid spaced at a resolution, typically 1Å, and produce a fully triangulated surface, something which not all methods do. The various algorithms present here work on these triangulated surfaces and their underlying grids, however the algorithms could be modified to run on any triangulated surface by imposing a grid or other structure to represent the volume.

In several algorithms used here, the convex hull surface of this molecular surface is also calculated. In three dimensions, the Qhull code, which is algorithmically optimal and also very fast in practice, [25] was used. The convex hull is the smallest surface with no invaginations or dimples that encloses the underlying surface or point set. In two dimensions it can be visualized by wrapping a rubberband around a set of points, in three dimensions the surface is that of a rubber ball stretched around points [26; 27].

From here, the general outline of the algorithms is to set a surface or set of points as the initiator, where all distances are set to zero. The next step is to set the allowed regions where the distance can propagate and the edges in the geometric graph that are traversable. Finally, an ending set of points or surface can be selected, however this is unnecessary. From here, the algorithm proceeds to compute the shortest paths from the initiation set to all other allowed points [23]. This is accomplished by using one data structure that holds the list of unseen points and another that is the tree of connections already made. The exact nature of these data structures changes the computational complexity of the algorithm but does not affect the results, for review see relevant chapters of the text of Cormen et al [28]. By keeping track of the closest points not yet seen and adding the closest point, the algorithm runs until all points have been seen or the termination surface has been reached. Since this problem has optimal subproblems, that is the shortest distance from A to C that passes through B is the shortest distance from A to B added to the distance from B to C, this algorithm can be completed quickly in terms of computational complexity as well as real computer time. This algorithm is referred to as multiple source shortest paths, Dijkstra's shortest paths or just shortest paths.

Note that the general problem of computing shortest paths in three dimensions with obstacles has been shown to be NP-hard [29], in other words it is likely that no polynomial time solution exists as it would mean polynomial time solutions exist to many other common problems thought to be exponential. However, as the construction of this proof involves creating obstacles of very fine complexity, we avoid this lower bound since proteins, while fractal in nature, have obstacles of a finite nature, the lower limit of size is that of the atomic radii involved.

Surface Depth

Many features of macromolecules are often referred to as deep or shallow. Grooves in DNA are often referred to as deeper or shallower and qualitative depths were assigned to the various canonical forms [30; 31 32]. No quantitative measure of this depth existed. Also, many binding sites in enzymes are called deep, or binding sites of protein-protein interactions are called shallow, again this qualitative description had little physical meaning and no quantitative method.

Chapter 2 is a description of the algorithm invented to quantitatively measure the depth of the protein surface including that of pockets, grooves and even tunnels. Briefly, this involves computing the distance from the convex hull to the macromolecular surface, while avoiding the molecule interior. This algorithm measures the depth to all points on the molecular surface and the entire intermediate volume between the molecular surface and the convex hull. Several applications are included, for instance examining a large set of protein-ligand co-crystal structures with experimental binding affinity data [33; 34]. Understanding the structural features of binding sites is important for many reasons. The structural basis of affinity between a protein and its ligand is a very important problem, since this could lead to the ability to design tighter binding drugs. Also importantly, the surface can be visualized, providing excellent graphics that aid in understanding and viewing complicated three dimensional surfaces in two dimensions. The Travel Depth computation, as this procedure is named, is completely automated once a molecular surface has been generated. This chapter is a based on previously published work [35].

Concurrent with this work on Travel Depth, a procedure to compute distances from a point of interest to the convex hull was published, called CAVER [36]. This procedure is different in several ways, first it requires the user to input a point of interest from which the distance to the convex hull is calculated. Second, surfaces are not explicitly constructed, instead a modified shortest paths algorithm finds the path that passes as far from the atoms as possible on the way to the convex hull. This procedure does not compute the distance to all surface and intermediate volume points, and cannot aid in visualization of the surface by coloring according to depth. Finally, no genomic scale analysis was completed. Any analysis possible with CAVER is possible with Travel depth, however the inverse is not true.

Ion Channels and Pores

Ion channels and pores are membrane spanning proteins that allow substrates to pass from one side of the membrane to the other. The process of membrane transport is extremely important biologically, and is involved in many processes like nutrient import or signaling. Though progress in determining their structures is behind that of soluble proteins, the number of structures is rising at similar rates now [37; 38], and numbers more than 200.

Finding the holes that allow these substrates to pass presents a challenge computational task even once the structure is known. Some ions are very small, smaller than the heavy atoms that make up the proteins themselves. These tunnels often vary in diameter as they pass through the membrane, for instance they usually have a narrow region that functions as the selectivity filter that specifically allows only one type of ion to pass through. These paths may not follow a straight line, though the original potassium channel structure does [39].

The previous work on finding and analyzing these holes is called HOLE [40; 41]. HOLE needs a starting point and direction, but proceeds from there by finding the largest circle that can be placed in each z-slice through the protein in the direction given. This procedure works well when very small steps are taken and when the starting point and direction are given correctly. However, it cannot identify paths that take a winding route and cannot deal with bifurcated paths. Also, it will attempt to identify a hole in the protein even when none exists, no topological checking is done to ensure that each path is through a hole.

In Chapter 3, the method called CHUNNEL is presented. The first step in CHUNNEL is to measure the shortest distance from the protein surface to all solvent points, which leaves a maximal ridge in three dimensions near the centers of all tunnels. This is combined with several topological procedures to guide the hole finding procedure and ensure that each hole is actually a hole and that each hole found is topologically distinct from all others found. This procedure works for all holes regardless of the path complexity through the protein and how many branches are encountered. Also, this procedure is completely automated, given a surface constructed it reports all the holes in that surface. Many analyses are completed, the most prominent being a complete catalog of all transmembrane proteins and their holes [42]. This chapter has been published previously [43].

Concurrent with CHUNNEL, several other methods were published and are discussed here. MOLE [44] is an extension of CAVER [36], which again computes distances from a point of interest provided by the user to the convex hull, along a path optimized to be far from the atoms. Instead of using a grid as before, the new path points are at Voronoi vertices [45] created from the protein atoms. Again, user input is required and no topology checking is completed, so paths are not guaranteed to be topologically distinct. MolAxis is similar in approach in that it uses Voronoi vertices instead of grid points, but again, user input is required to find the paths and no topological checks are done on paths found to ensure that they are tunnels[46]. Neither of these methods can perform the fully automatic analysis enabled by CHUNNEL, neither are run on the entire set of transmembrane protein structures for instance, neither find the complete set of topologically distinct holes.

Using Voronoi vertices created from atom centers is however an interesting technique. Since a Voronoi edge exists where any three atom centers are equidistant, an edge will be present throughout the length of any tunnel, connecting Voronoi vertices of atoms lining the tunnel. This seems possibly superior to using grid points, as very fine grids may be necessary to find the smallest tunnels of interest, for instance chloride channels. Combining the Voronoi methods with the methods to compute the distance from the surface into solvent and topological checking would likely be the a good combination approach, and is discussed in Chapter 6 with other future work.

Thermostability

The structural basis for thermostability of protein structures has been examined from many perspectives. Proteins from hyperthermophilic organisms maintain their stability even at temperatures as high as 80 degrees C. Understanding the structural basis for thermostability is important due to the many applications like protein design[47]. Examinations of the differences between these structures and those from mesophiles have commonly included analyzing the differences in exposed surface area [48; 49; 50; 51; 52; 53; 54; 55]. However, few if any studies have examined the distribution of residue burial, or distance from the atoms to the molecular surface. Also, no study had examined the number or depth of pockets, or more generally, the overall shape differences between hyperthermostable proteins and mesostable homologues.

Both Burial Depth, the distance of each atom to the molecular surface, and Travel Depth, the distance of the molecular surface from the convex hull avoiding the protein interior, were used to analyze a dataset of thermostable and mesostable pairs of proteins [51]. These analyses are presented in Chapter 4, leading to the conclusion that hyperthermostable proteins are more spherical, in that they have fewer pockets and fewer deep pockets, and they bury atoms more deeply from the surface. This work was published previously [56; 57].

Protein Pockets

Pockets, like depth, are an oft-discussed but ill-defined feature of protein surfaces. Finding potential pockets to evaluate their possibility for ligand binding is just one of many applications where a good pocket definition is necessary. The field of functional site location, similarity between sites and docking ligands into those sites is reviewed by Campbell et al [58].

In Chapter 5 the CLIPPERS method is introduced. Building on top of the Travel Depth analysis, CLIPPERS analyzes all pockets on the protein surface, using a very liberal definiton of pocket, which generates a hierarchy of nested pockets that completely cover the protein surface. After finding pockets, their shape features are easily computed, and pockets can then be compared and clustered. Pockets can be tracked throughout transition pathways with time, across mutations, with different binding partners, or across diverse families of protein structures. This work will be published as all other work in this thesis has been[59]. There are many other pocket finding methods, reviewed in Chapter 4, however CLIPPERS is the first to completely cover the protein surface with pockets and also to compare them based on shape alone, not alignments or by residues.

Summary

By using the shortest paths method on geometric graphs, distances between surfaces and/or volumes can be easily quantified. These distances, along with other techniques, allow algorithms that can measure the depth of an entire macromolecular surface, or the depth of all the atoms within the surface. Also, these distances form the basis for methods to automatically catalog and measure both tunnels and pockets in proteins. This thesis presents all these algorithms and applications, all made possible through consistent application of the shortest paths algorithm and additional supplementary algorithms.

Chapter 2

The bulk of this chapter was previously published [35].

Summary

Depth is a term frequently applied to the shape and surface of macromolecules, describing for example the grooves in deoxyribonucleic acid (DNA), the shape of an enzyme active site, or the binding site for a small molecule in a protein. Yet depth is a difficult property to define rigorously in a macromolecule, and few computational tools exist to quantify this notion, to visualize it, or analyze the results. We present our notion of *travel depth*, simply put the physical distance a solvent molecule would have to travel from a surface point to a suitably defined reference surface. To define the reference surface, we use the limiting form of the molecular surface with increasing probe size: the convex hull. We then present a fast, robust approximation algorithm to compute travel depth to every surface point. The travel depth is useful because it works for pockets of any size and complexity. It also works for two interesting special cases. First, it works on the grooves in DNA, which are unbounded in one direction. Second, it works on the case of tunnels, that is pockets which have no 'bottom', but go through the entire macromolecule. Our algorithm makes it straightforward to quantify discussions of depth when analyzing structures. High-throughput analysis of macromolecule depth is also enabled by our algorithm. This is demonstrated by analyzing a database of protein-small molecule binding pockets, and the distribution of bound magnesium ions in RNA structures. These analyses

show significant, but subtle effects of depth on ligand binding localization and strength.

Introduction

Depth is a term frequently applied to the shape and surface of macromolecules. For example, enzyme active sites are routinely described as shallow or deep. Small ligand binding sites on proteins are also frequently described in term of depth. Depth is just one facet of the property 'binding pocket shape' one would like to define quantitatively, to aid for example, in screening a large library of potential ligands, or in docking of a candidate ligand. Groove depth is one of the fundamental terms used to describe the differences in structure of the A, B and Z forms of DNA [30; 31; 60]. In spite of the common use of the term depth, it is a surprisingly difficult property to define rigorously in a macromolecule. Discussions of depth in the literature, although intuitively reasonable, are usually qualitative. The concept of depth is thus difficult to subject to rigorous analysis or to extract the most information from. A large part of the difficulty in analyzing depth is due to the complexity and range of shapes adopted by macromolecules. Protein surfaces are fractal in nature [13], adding to the difficulty. To illustrate some of the difficulties, consider first the issue of a reference point or level. In geodesy, mountain peaks and ocean depths are referenced to the mean sea level, providing a standard reference level (Although not without regional difficulties: mean sea level either side of the Panamanian isthmus differs considerably, for example). There is no equivalent to mean sea level in a molecule. Second, consider the case of deep pockets involving overhangs or that re-approach the molecule surface at some point away from their origin. Euclidean distance of the bottom of the pocket to the nearest surface, while easy to define and compute, will be a very misleading and grossly underestimating measure of depth. These difficulties are reflected in the fact that there are few computational tools to quantify the concept of depth, to visualize it, or analyze the results. To address this problem, we present here our notion of *travel depth*, simply put the physical distance a solvent molecule would have to travel from a surface point to a suitably defined reference surface. The concept of travel depth was designed to avoid the 'short circuiting' error described above, and also to solve the problem of a reference level. We first define the concept of travel depth, and the reference level used by it, then present a fast, robust approximation algorithm to compute travel depth to every surface point. Selected examples using very different molecular shapes are used to demonstrate that our definition of depth works for special cases, and that it conforms to our intuition, so confirming that we have introduced a 'good' definition for depth and that our approximate numerical implementation of it is reasonable. We then describe some applications of our algorithm, including a high throughput application to a small molecule binding database.

Theory and Methods

Definition of travel depth

Any measure of depth must start with the questions: Depth of what, and from what? In this work, we are concerned with the depth of any point on the molecule's surface. Two definitions of surface predominate for macromolecules, the solvent accessible surface [4], and the molecular surface [7]. In both cases a crucial parameter is the probe radius, which is almost universally taken to be that of water (usually values between 1.4Å and 1.8Å are used). Many algorithms exist for computing these idealized surfaces. Most, but not all, produce a triangulated form of the surface, primarily for display using standard computer graphic routines [10; 12; 24; 61]. Our algorithm assumes a simple closed triangulated surface. The surface must be orientable and connected, though these are not strong requirements; The latter disallows only cavities. For the broadest applicability of our method, we make no other assumption about how the surface was produced, or what it should look like. In practice we use the molecular surface as generated by the algorithm in the GRASP macromolecular graphics program [24] implemented as a stand-alone program [62] using a probe radius of 1.8Å and standard atomic radii [3]. Though we test only this surface generation scheme and the resulting triangulated surfaces, our definition and algorithm generalize to any triangulated surface generation scheme.

Our definition of travel depth is that for each point on the molecular surface, the travel depth is the minimum distance a solvent probe would have to travel through the solvent from that surface point to get to the reference level. A natural and parameter independent reference level is provided by the convex hull of the molecular surface. The convex hull is a standard construct in computational geometry. In three dimensions, the convex hull is the smallest volume convex polyhedron that contains all the surface points [25; 26; 27]. In terms of molecular surfaces, the convex hull is equivalent to the molecular surface produced by an infinite solvent probe radius. Algorithms and code for convex hull computation have been well studied and are fast and reliable [25; 26; 27].

The next step is to compute the minimal distance from every surface point to the convex hull while respecting the boundary of the molecular surface. In other words, the travel path along which the distance is computed must lie outside the molecular surface in the solvent. We note that computing such a minimal distance between two points while avoiding obstacles is exactly the shortest path planning problem commonly encountered in robotics, and that an exact solution to the problem is NP-hard. Our solution, described below, is to approximate this minimal distance in such a way that it was easy to code and run in a short time so that we could establish what the depth measure would look like on real examples, and whether it would be useful in structural analysis.

Calculation of travel depth: Preprocessing

The first step is to remove cavities, defined as completely enclosed solvent pockets in the molecular surface. The triangles that represent these cavities are removed from the surface and are not used in later calculations. Since there is no way for the solvent probe to travel from a closed cavity surface to the convex hull without passing through the macromolecule itself, travel depth does not apply to these surfaces. We note, though, that simple Euclidean distance to the nearest part of the external molecular surface would provide a satisfactory definition of the minimum depth of a closed cavity.

Two important pre-processing steps are done at this stage. First, the longest edge of any triangle in the surface is found and the length saved for later. Also, all the points on the surface are put into an two-dimensional orthogonal range search tree structure oriented along one grid axis [26]. This helps improve the running time, as described later, but it is non-essential to the algorithm.

Calculation of travel depth: Mapping onto Grid

The macromolecule and a region of the surrounding solvent are embedded in a cubic grid of dimensions K x L x M. For convenience, the grid extends to one grid cube beyond the minimum and maximum coordinate of the molecular surface in each orthogonal direction, so that the border is completely outside the surface. The default grid spacing used in our algorithm is 1Å, however the algorithm and code generalize to any spacing. The only consideration is for the spacing to be small enough to approximate well the topology of the given molecular surface. For instance, when a probe radius P=1.8Å is used, as in our surfaces, the maximum concavity of any section of the molecular surface is limited to that of the probe radius. From this, a maximum allowable grid spacing, G, can be calculated from the formula

$$G = 2P \Big/ \sqrt{3} \qquad (2\text{-}1)$$

This grid spacing ensures that any concave depression in the surface is represented by at least one grid center. Using the same formula with the smallest atom radius used to construct the molecular surface leads to another bound on the grid spacing to guarantee that any convex protrusion is represented by at least one grid point. Again, 1Å is well within this limit for most commonly used radii for heavy atoms (the smallest such atom commonly found in biomolecules is oxygen, with a radius of

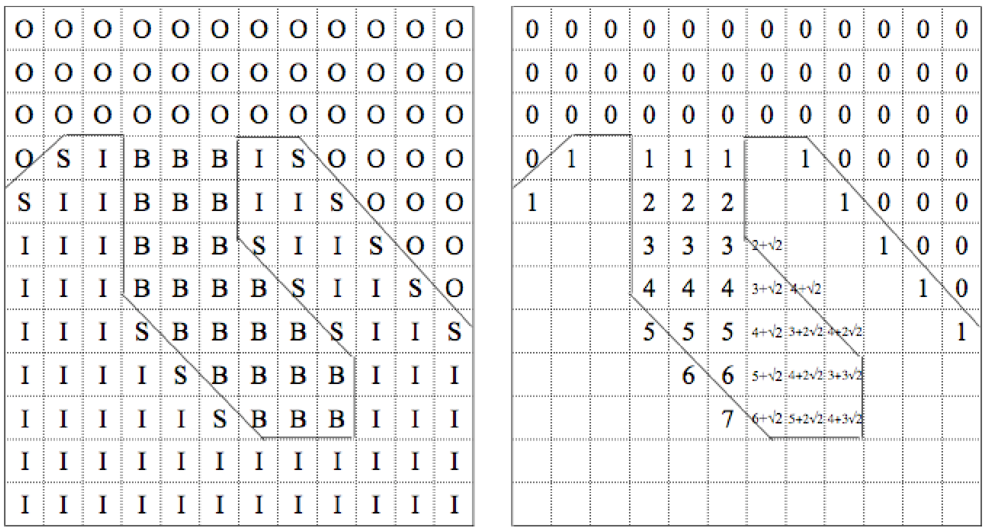
1.4Å). This assumption ignores problems caused by a very coarse surface, though this assumption is relaxed and a solution to problems caused by this in our algorithm are discussed later.

The next step of the algorithm is to find the convex hull of the molecular surface. There are many *O(n log n)* algorithms for computing the convex hull in three dimensions. We use available code from Qhull or Quickhull, an optimized and robust package [25].

Calculation of travel depth: Classifying Grid Points

After construction of the convex hull each point lying at the center of each grid cube must be checked to see whether it lies inside or outside the convex hull and inside or outside the molecular surface. The convex hull can be represented as a list of outward facing triangles. A sufficient check for being outside the convex hull is to check the point against each triangle and surface normal to see which side it is on. A point that is outside any convex hull triangle is outside the entire convex hull. Doing this check for each point in the grid is sufficient to determine which points are outside the convex hull and which are inside. Next, points inside the convex hull are assigned to either the outside or the inside of the molecular surface. This step is the most time consuming portion of the entire algorithm. The problem is that determining whether a point is inside or outside a general triangulated surface requires global information. It is not sufficient to check a point against every surface triangle. However, an appropriate geometric property can be used to solve this problem quite efficiently: Any line drawn completely through the molecular surface will intersect an even number of triangles. Lines are constructed in one orthogonal direction of the grid such that they each pass through a set of grid points. Moving from grid point to grid point along this line from one side of the grid until the first triangle is met assigns all those points to the outside. Each time a triangle is encountered, the inside/outside assignment switches. This procedure is continued until the opposite side of the grid is reached. In this manner, when a complete set of lines through the grid in one direction have been processed, the correct assignment has been made for all the points. In practice, since a line of grid points is used in this step, all their inside or outside checks can be done at once: Each triangle from the surface can be checked to see if it intersects this line, and to find the point of intersection if it exists. After this, the previously described procedure can be used to determine on which side of the surface each point on that line lies.

Naïvely, each triangle could be tested against each line. However, a more efficient procedure which drastically cuts down the number of intersection checks uses both of the preprocessing steps mentioned earlier. After picking a dimension along which the lines will be constructed, the other two dimensions are chosen as the orthogonal directions to construct a 2 dimensional orthogonal range search tree from all the surface points. This polynomial time construction allows queries that consist of any orthogonal, or grid-aligned, rectangle which return all the points in that area [26]. This is used in conjunction with the precomputed longest edge length, to quickly find all triangles that possibly intersect the line of interest, by querying the square centered around the line's axis plus and minus the longest edge length. Only triangles that have all three points in this square possibly intersect the line of interest, and this



O	O	O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O	O	O
O	S	I	B	B	B	I	S	O	O	O	O
S	I	I	B	B	B	I	I	S	O	O	O
I	I	I	B	B	B	S	I	I	S	O	O
I	I	I	B	B	B	B	S	I	I	S	O
I	I	I	S	B	B	B	B	S	I	I	S
I	I	I	I	S	B	B	B	B	I	I	I
I	I	I	I	I	S	B	B	B	I	I	I
I	I	I	I	I	I	I	I	I	I	I	I
I	I	I	I	I	I	I	I	I	I	I	I

0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	1		1	1	1		1	0	0	0	0
1			2	2	2			1	0	0	0
			3	3	3	$2+\sqrt{2}$			1	0	0
			4	4	4	$3+\sqrt{2}$	$4+\sqrt{2}$			1	0
			5	5	5	$4+\sqrt{2}$	$3+2\sqrt{2}$	$4+2\sqrt{2}$			1
				6	6	$5+\sqrt{2}$	$4+2\sqrt{2}$	$3+3\sqrt{2}$			
					7	$6+\sqrt{2}$	$5+2\sqrt{2}$	$4+3\sqrt{2}$			

Figure 2-1 Travel Depth 2D example

Two schematic two-dimensional examples of the Travel Depth Algorithm. Left: an example of a piece of a macromolecule, the grid superimposed on it, and class assignments made: outside convex hull (class O), inside the molecular surface but containing at least one molecular surface point (class S), inside molecular surface and not containing any molecular surface points (class I), and between the convex hull and molecular surface (class B). Right: the travel depths assigned to each grid square, note the diagonal paths lead to non-integer travel depths.

test quickly reduces the number of triangle intersection checks that must be done. Though these checks each take constant time, they can be very slow, as they involve evaluating several matrix determinants. To unambiguously determine inside and outside, our algorithm assumes that these lines will not intersect a triangle across its face, or through a single vertex. These special cases, if they occur, are easy to detect and the points can be slightly perturbed until the ambiguity no longer occur.

At this point each grid cube has been classified into one of four categories based on the location of its center and whether it contains any molecular surface points. Either outside the convex hull (class O), between the convex hull and molecular surface (class B), inside the molecular surface but containing a molecular surface point (class S) or finally inside the molecular surface but containing no molecular surface points (class I). A small example is shown in the left panel of Figure 2-1. Class I cubes are ignored in the rest of this work, as no depth needs to be calculated for them.

Calculation of travel depth: Assignment of Travel Depth to Grid Points

It remains to approximate the minimum distance that a probe sphere would need to travel to get from each surface point to the convex hull. This travel depth is assigned to class B and S points recursively, as follows. All grid cubes of class O are assigned a travel depth of zero. All cubes of class B and S are initially assigned an unreachably large value, e.g. KxLxM, indicating that no depth has yet been determined for those cubes. For each grid cube i of class B or class S, its travel depth d_i is set to the sum of the travel depth of its neighboring grid cube, d_j, plus the distance to that neighboring cube, dist(*i*,*j*) (*vide infra*). If the cube has more than one neighbor with assigned depth, which is usually the case, the neighbor that results in the minimum depth d_i is chosen. Symbolically.

$$d_i = \min_j(d_j + dist(i,j)) \tag{2-2}$$

where j ranges over all neighbors of i. This procedure is repeated until no new depth assignments are made.

A key requirement to correctly propagate depth with respect to the topology of the molecular surface is the appropriate definition of neighboring cubes in equation 2-2. For a class O or B cube, any of the 26 immediately adjacent cubes of class O, B or S are considered neighbors. Additionally molecular surface edges which have an endpoint in a class O or B cube and another endpoint in a class O, B, or S cube make those two cubes neighbors. For class S grid cubes, any adjacent cube of class O or B is a neighbor. However, for a class S cube only class S grid cubes that are connected to it by a molecular surface edge are considered neighbors, even if the two S class cubes are adjacent. There may be adjacent class S grid cubes that do not have a molecular surface edge between them, for example when two distant parts of the molecular surface approach each other very closely without meeting. It is important not to propagate the travel depth across this gap.

The neighbor distances in equation 2-2 are defined as follows: Each grid cube has 6 adjacent cubes that share one face, 12 adjacent cubes that share only an edge and 8 adjacent cubes that share only a vertex. The distances to these three types of adjacent cube are the Euclidean distances between cube centers, 1, √2, and √3 grid units respectively. Additionally, cubes of class S can have additional neighbors defined by edges of the molecular surface which have endpoints in the two grid cubes, i and j. Their distance is also the Euclidean distance between the grid cube centers.

Starting from the class O grid cubes with depth 0, the neighboring grid cubes are assigned a depth according to equation 2-2, then the neighbors of the neighbors are assigned and so on. In this way the depth propagates in towards the molecular surface, and into the class S cubes, but it does not propagate through the macromolecule since the depth assignment is not propagated into class I cubes. This is illustrated in the right panel of Figure 2-1. After the assignment phase terminates, the depth is converted from grid units into a physical distance by multiplying by the grid spacing. This results in a calculation of the shortest paths from the class O cubes to all class B and class S cubes, given the neighbor and distance definitions above.

The depth assignment phase of the algorithm is speeded up by using Dijkstra’s algorithm for shortest paths on a graph [28] and using available code that implements a key component of that algorithm, a priority heap. Dijkstra’s algorithm keeps track of the vertices in the graph (grid cubes) which have already been assigned a travel depth, and the shortest path from these assigned grid cubes to the rest of the grid cubes. The priority heap keeps track of the unassigned grid cubes that can be assigned a travel depth, and efficiently updates and finds the current shortest travel depth grid cube that has yet to be processed. In practice, we use a priority heap that has reasonable amortized performance and was compatible with the rest of our code [63]

At this stage, all that remains is to assign each surface point a depth based on the grid cube it is located in, resulting in a computed travel depth for each point on the surface. The travel depth is also computed for all the grid cubes B between the molecular surface and the convex hull as well as the grid cubes I that contain surface points. Although the travel depth assignment of points between the convex hull and the molecular surface is not used in the applications of travel depth described here, it is a property that may prove useful in future applications like docking.

Presentation of results

To visualize the results of our algorithm we used the macromolecular graphics package PyMOL [64]. The triangulated molecular surface can easily be read into this program, along with travel depth values, and a red-green-blue color gradient assigned to each point of the surface based on travel depth. Red represents a travel depth of zero, with increasing depth indicated as the color changes from green to blue. The depth represented by blue is set either to the maximum value for that molecule, or to a fixed value to compare of a set of molecules. Color values at each point along each edge and triangle are interpolated using the standard approach to produce a smooth visualization of depth [64; 65]. Further refinements, such as displaying only surface in a certain range of depth may be useful for particular applications, and are straightforward with our algorithm.

Robustness, errors and timing analysis

Depending on the size of the macromolecule and the resolution at which the molecular surface is generated, the input surface to the travel depth algorithm might be quite coarse. In this case regions of the surface may not conform well to the estimates of maximum concavity. This may result in small crevices or tunnels which violate the maximum concavity assumption. These errors are accounted for by the molecular surface edges that define grid cube adjacencies. The only level of coarseness that may cause a problem is where two parts of the surface approach each other very closely, less than the grid spacing. In these cases, the travel depth would propagate between these surfaces when it should not. However, to violate this assumption requires a violation of the maximum convexity assumption, which corresponds to a severe underestimation of the size of an atom or adjacent atoms forming such a barrier.

There are two sources of error in our approximation algorithm, each of which can be reduced at the cost of increasing the running time of the algorithm. The first source of error comes from the grid orientation. The approximate distance can be overestimated if a significant part of the path traveled is diagonal with respect to the grid axes. The worst case is when the actual distance should be down two grid units and over one grid unit in both other directions, the path length here is $\sqrt{6}$, while the approximation given is 1 grid unit down and then one diagonal step of length $\sqrt{3}$. This type of error leads to an error factor at most $(1+\sqrt{3})/(\sqrt{6})$, or roughly 1.11 times the actual shortest path length. Rotating the grid axes and re-running the algorithm and taking the minimum computed in either orientation would reduce this error, although we found that for the applications described here it has not been necessary.

The second source of error lies in the discretization of the distance, again from the use of the grid cubes to approximate the distance. Using smaller grid cubes, at a cost of increasing the running time, can reduce this error. In practice, there is little reason to get an extremely accurate measure of this distance, as there are already sources of uncertainty regarding the travel depth property, and indeed in the molecular surface construct itself. It would be hard to argue that differences of some small travel depth distance like 1Å had any real physical meaning.

Our algorithm has both a reasonable asymptotic running time when the complexity is analyzed, and a reasonable running time in practice. Also, following the philosophy of keeping the code as simple as possible, time spent coding and debugging was minimized, available pieces of code like PyMOL [64] and a priority heap [63] were used when possible.

We have highlighted the practical runtime issues throughout the description of the algorithm. The algorithm also has a reasonable running time when analyzed asymptotically [28]. Without the orthogonal range search tree speedup mentioned, the running time is

$$O\left(p\log p + cd^3 + (t+d)d^2 + (d^3+e)\log d^3\right) \tag{2-3}$$

where *p* is the number of points on the molecular surface, *c* is the number of triangles on the convex hull, *t* is the number of triangles on the molecular surface, *d* is the number of grid cubes in any dimension, and *e* is the number of edges, which is linear in terms of t and d^3. The first term in equation 2-3 comes from the convex hull construction, the second term from the checks for each grid cube to see if it is inside or outside the convex hull. The third term comes from the checks to see if each grid cube is inside or outside the molecular surface. The fourth term is the cost of the propagation step using the shortest path algorithm and amortized time cost priority heap.

With the orthogonal range search tree speedup in place, there are two additional components to consider, the $O(t)$ steps to find the longest triangle edge, the $O(t \log t)$ steps to build the orthogonal range search tree (faster algorithms exist, but are harder to code [26]). The $O((t+d)d^2)$ term to check each grid cube becomes $O((\log^2(t) + k + d)d^2)$ step to do a range search query and then k checks must be done, where k is the number of triangles returned from the range search. Also, it should be noted as was later revealed by our timing analysis that the orthogonal range search idea should probably be applied to the inside/outside convex hull routine, changing the $O(cd^3)$ time into $O(c \log c + ((\log^2(c) + d)d^2))$ as the time for the convex hull checks now outweighs the time for the molecular surface checks as we have it implemented.

At the heart of this analysis is the fact that if we halve the grid spacing used, our algorithm gets worse by a factor of 8, since there are twice as many grid cubes in each dimension. This is one reason grid distances smaller than 1 Å are never considered. Though they could be calculated they are impractical. Fortunately this analysis shows us that the overall speed of the slowest steps in practice, that is checking whether each grid cube is inside or outside the various surfaces, can be made to grow only with the squared logarithm of the number of triangles, plus the factor *k* representing how many triangles are returned from an average range query. Though an initial penalty must be paid, this provides an overall faster approach as the number of triangles increases. This allows us to use very fine triangulated surfaces and still maintain reasonable runtimes, or use very coarse triangulated surfaces to get good exploratory results.

To give some estimate of the processing time involved, we provide the following timing analysis, conducted using one processor of a dual processor machine (Intel 2.4 GHz chip, 4797 BogoMIPS, 1 gigabyte RAM) running RedHat Linux 9.0. Different parts of the algorithm were timed separately. Two test PDB files were used, s a representative small protein, cyclic bovine pancreatic trypsin inhibitor, PDB code 1K6U. To represent larger more complicated proteins, the 6 chain biological unit of pertussis toxin was used, taken from PDB code 1PRT. For these two samples we constructed molecular surfaces of varying fineness, the number of triangles in each is reported, along with times in seconds for each of the three main phases of our algorithm. All these results are shown in Table 2-1. It should be noted that while the orthogonal range search tree speedup was in place for the molecular surface, it was not in place for the convex hull code here.

We note that during even the largest test case examined, only about 200 megabytes of available memory were in use, suggesting that memory usage is not a limiting factor in our algorithm, even though no formal asymptotic space analysis was conducted. The runtime for the large example at fine granularity represents an extreme case, one which would typically only be undertaken for a figure. The

Table 2-1 Timing analysis of travel depth code

PDB Code	1K6U			1PRT		
Level of Detail	coarse	medium	fine	coarse	medium	fine
Number of Triangles	3644	8148	30896	5940	13780	57836
Inside/outside Convex Hull (s)	14	21	53	14	241	443
Inside/outside Molecular Surface (s)	4	5	12	5	48	153
Depth Assignment(s)	9	10	15	9	126	149
Total Time (min)	1	1	2	1	7	13

statistical analyses were done at a more medium granularity setting, which proved sufficient.

Results

The first tests of the travel depth algorithm were designed to see if the definition conformed to one's qualitative intuition about depth in macromolecules. In other words, is the definition of travel depth reasonable and useful? We used a variety of structures that had qualitatively different surface topographies. The first is duplex DNA, to which the term groove depth is commonly applied. We evaluated the depth of the major and minor grooves in A, B and Z canonical forms of DNA. 15 base pairs of A-T were generated with the routine NUCGEN [66] in canonical A form, crystal structures 1BNA[67] and 3ZNA [32; 68] were used for the B and Z forms respectively. It should be noted the structure 3ZNA was constructed by duplicating base pairs present in the crystal to achieve the length shown, and is therefore considered a theoretical model in the PDB. Our travel depth algorithm gives intuitively reasonable results, shown in Figure 2-2. All surfaces are colored from red (travel depth 0 Å) to green (travel depth 7 Å), then finally to blue (travel depth 14 Å). It is clear that the major and minor grooves of the B-form are nearly the same depth, whereas the major groove of the A-form is much deeper than the minor groove of A-form or either groove of the B-form. Also, what would usually be the minor groove has turned into a very deep groove in the Z-form, and the major groove has almost no depth. We summarize these results quantitatively in Table 2-2. This is in good agreement with the standard description of these grooves [30; 31; 60]. Specifically, "in A-DNA the helix axis passes by the major groove side of each base pair, making that

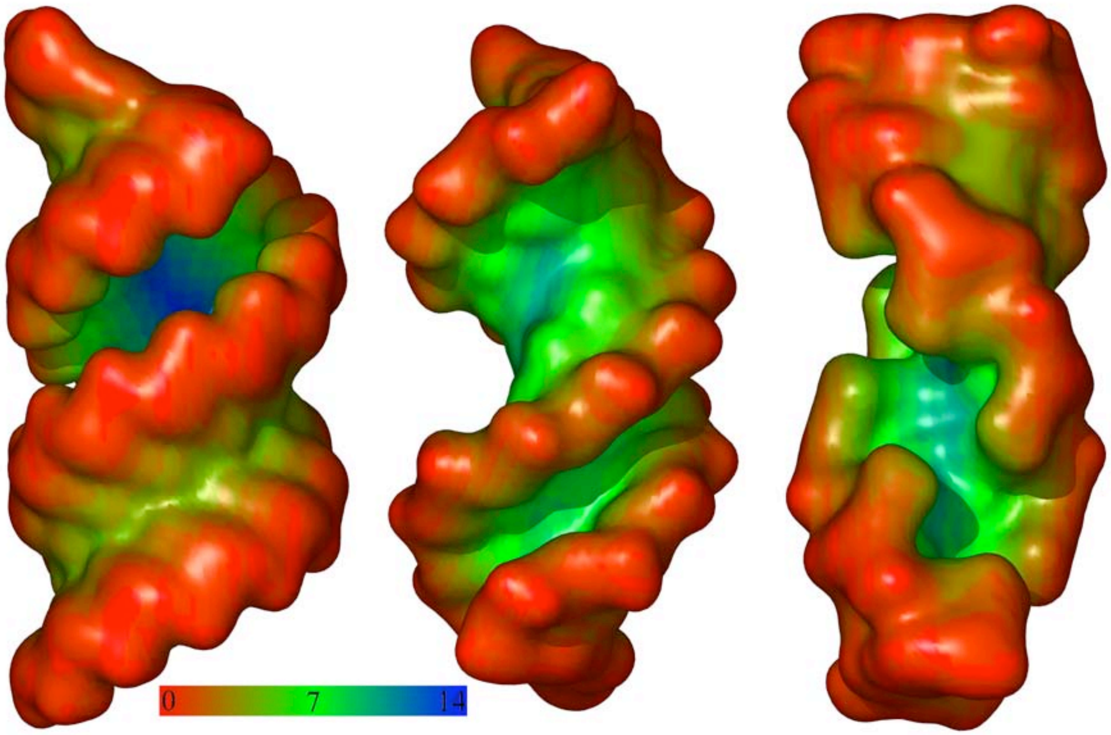


Figure 2-2 Travel Depth of DNA

Travel Depth coded molecular surface of the three canonical forms of DNA, from left to right, A, B, and Z. All surfaces are colored from red (travel depth 0Å) to green (travel depth 7 Å), then finally to blue (travel depth 14 Å) as indicated by the color bar legend.

Table 2-2 Travel Depths of Selected Macromolecular Features[a]

	Major Groove Max Depth	Minor Groove Max Depth	Major Groove Average Depth	Minor Groove Average Depth
A-DNA	13.7	5.0	8.6	3.6
B-DNA	10.0	9.1	4.6	5.6
Z-DNA	4.2	10.8	2.0	6.5

	Binding Site Average Depth	Binding Site Max Depth	Tunnel Depth at Center	Ring Around Tunnel Depth
Tunnel (1A0Q)	10.6	18.0	23.0	18.0
Horseradish Peroxidase (1ATJ)	18.8	24.0		
Streptavidin-Biotin (1MK5)	8.5	10.0		

[a] Depths in Ångstroms.

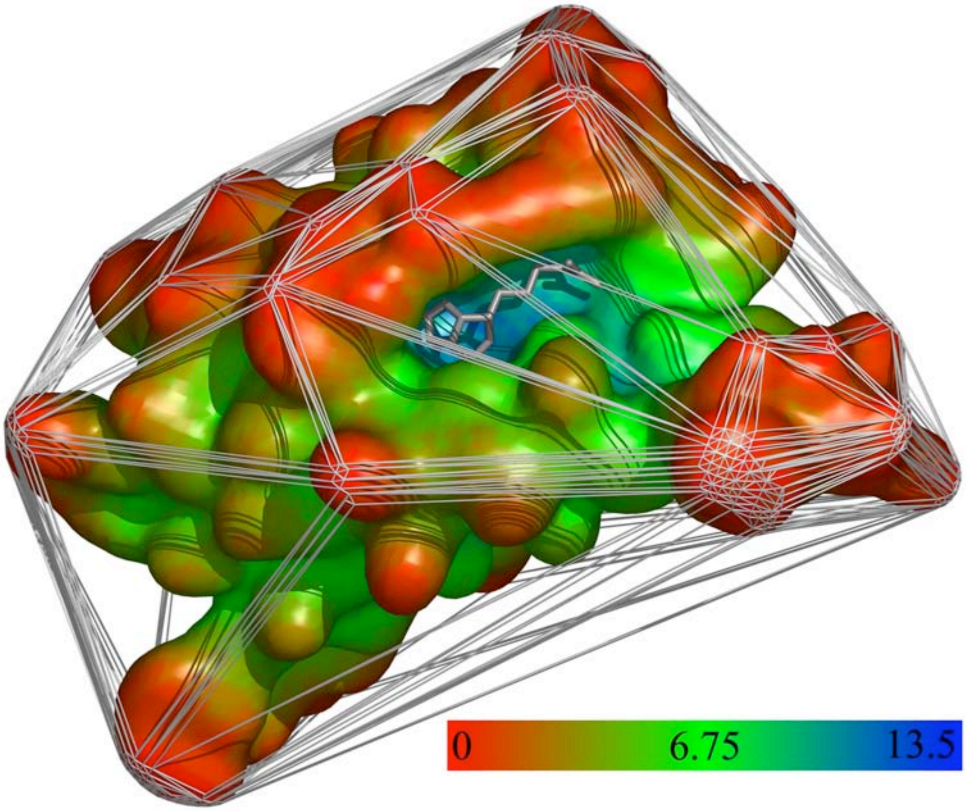


Figure 2-3 Travel Depth of Streptavidin

An example binding pocket color coded by travel depth. This example is PDB code 1MK5, a biotin/streptavidin complex, the biotin binding site has a maximum travel depth of 10Å. Additionally, the edges of the convex hull are shown.

groove very deep, the minor groove shallow…" [30]. Also, B-DNA is described: "This means that major and minor grooves are of comparable depth…" [30]. Finally, Z-DNA is described: "With the helix axis passing down the minor groove, that groove is extremely deep, whereas the major-groove edge of each base pair is pushed out to the perimeter of the helix, giving the groove zero depth" [30].

To further illustrate that our algorithm is intuitively correct, we show three other examples. First, a simple well-known pocket was analyzed, that of streptavidin bound to biotin (PDB code 1MK5). The result is shown in Figure 2-3. This clearly illustrates that travel depth can quantitate a pocket near the surface. Next, a tunnel is shown in Figure 2-4 from the FAB fragment (PDB code 1A0Q [69]). The travel depth algorithm works well in this case. Despite the fact the tunnel has no bottom the middle of the tunnel is correctly identified as the deepest point. Also, the tunnel in Figure 2-4 is additionally characterized by the maximum distance for which a solid connected ring of surface points exists all the way around the tunnel, which is 18 Å. Finally, horseradish peroxidase (PDB code 1ATJ) is shown, which has a very deep active site. Figure 2-5 shows the result, which illustrates a case where a purely Euclidean distance algorithm would fail, as the deepest part of the pocket is closer to the other side of the protein than the one the substrate must enter from. A summary of various features on these previous six examples is shown in Table 2-2.

As an example of a high throughput data base application of the travel depth algorithm, we examined the small molecule binding structural database PDBbind [33; 34]. All 900 structures were used from the 2003 refined set [34]. The proteins each bind a single small molecule ligand, and have binding data associated with the complex,

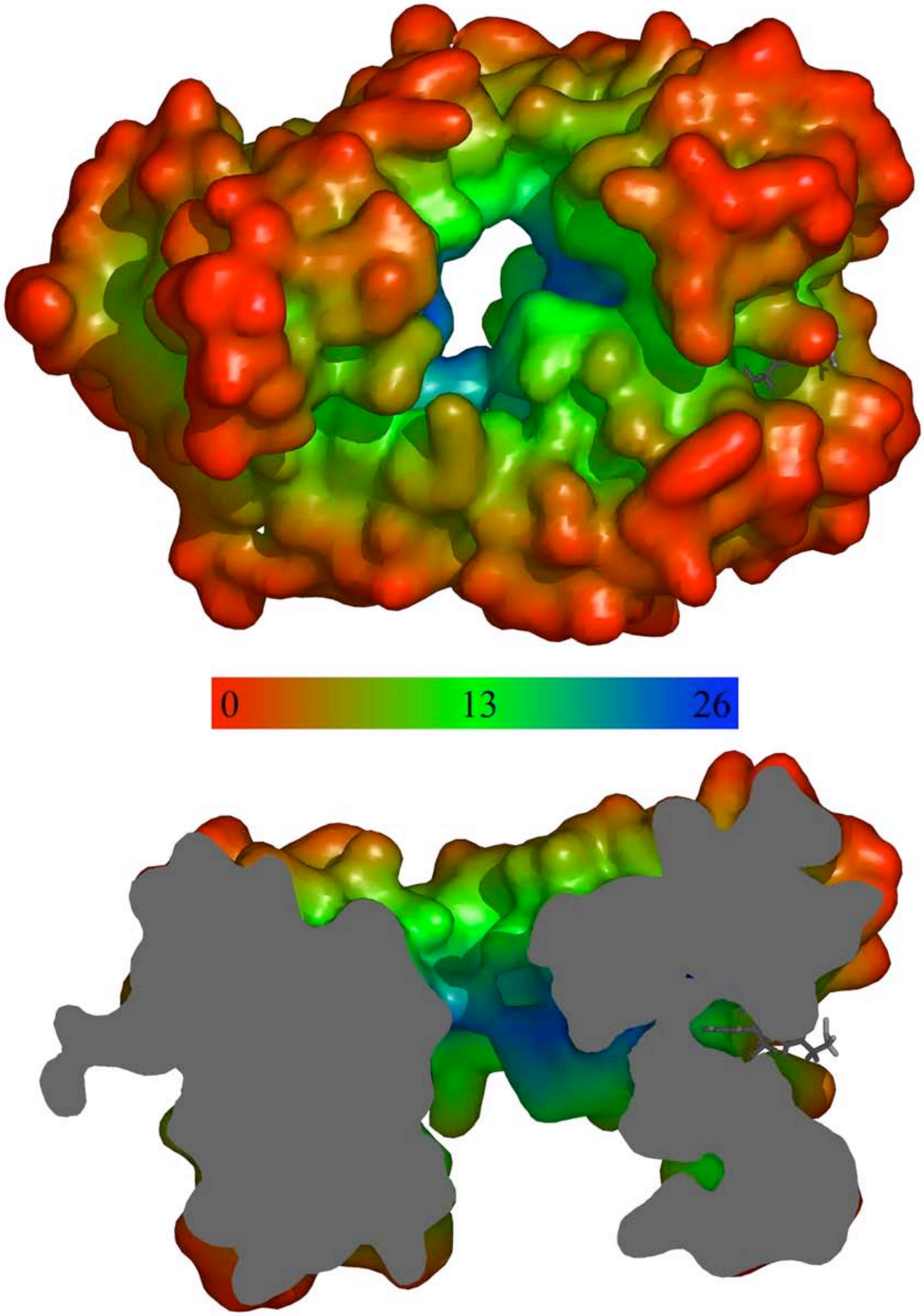
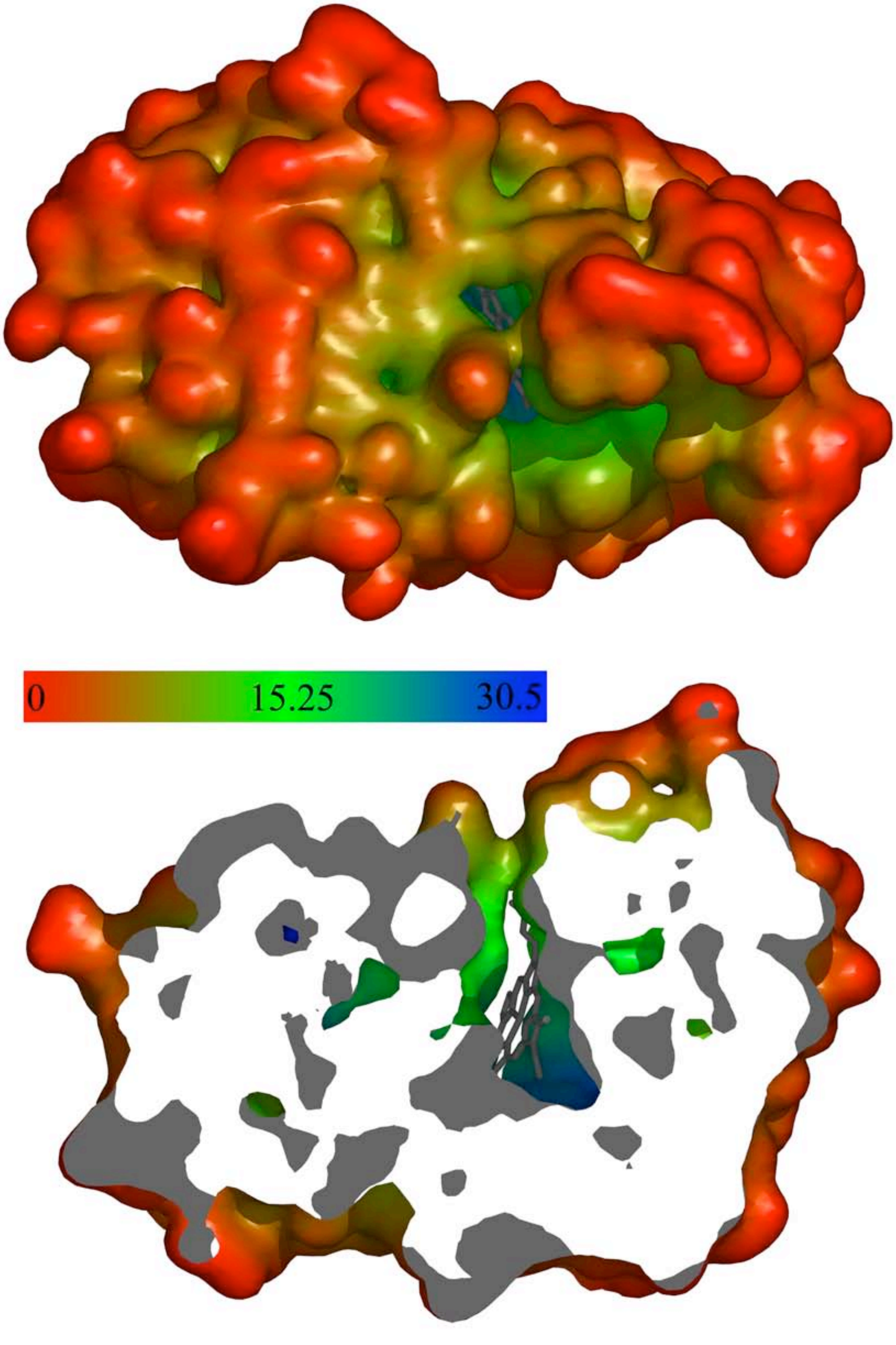

0
13
26

Figure 2-4 Travel Depth on a Tunnel

(Preceeding Page) An example tunnel color coded by travel depth. This is a FAB fragment from PDB code 1A0Q. The top view looks down on the tunnel, the bottom view is a side view that has been cutaway through the tunnel.

Figure 2-5 Travel Depth on a Deep Pocket

(Following Page) An example deep pocket color coded by travel depth. Two views of horseradish peroxidase, taken from PDB code 1ATJ. The bottom view is a cutaway showing one view of the pocket with a maximum depth near the ligand of 24 Å . A straight line Euclidean metric from the deepest point of this pocket would travel through the protein to the wrong side.


0
15.25
30.5

as well as separate files for protein and ligand. Binding data for this set is from either the dissociation constant (–log K_d) or competitive inhibitor concentration (- log K_i), both referred to here for brevity as –log K. This database was chosen over other available options because the structures and binding data had been hand checked and gathered from original sources, and the structure files were easily accessible, downloadable in modified, in clean form within one archive file. This allowed us to perform the analysis with only minor conversion of data formats, and no further editing or checking of input files. We note that 13 of these structures had ligands completely enclosed in cavities, inaccessible to solvent, and therefore only 887 structures were used whenever the ligand site was analyzed. The protein atom coordinates were used to construct the molecular surface at a medium setting of surface coarseness. We assume that sampling the travel depth at these surface points gives us an accurate and representative picture of the depth of the protein, or of a ligand binding site for instance. Under this assumption, averaging the travel depth across the surface points is an acceptable way to measure the overall travel depth of a protein, as is done later.

To test the hypothesis that ligands are in deeper pockets rather than shallower pockets, the protein surface points were divided into two classes, those near ligand atoms representing the binding site, and the rest. For each atom in the ligand, the single nearest surface point was found and included in the binding site if it was within an arbitrary threshold of 4Å. This method gave a simple way of partitioning the surface into the binding site and the rest of the surface, erring on the side of including too few surface points in the binding site. The results are shown in Figure

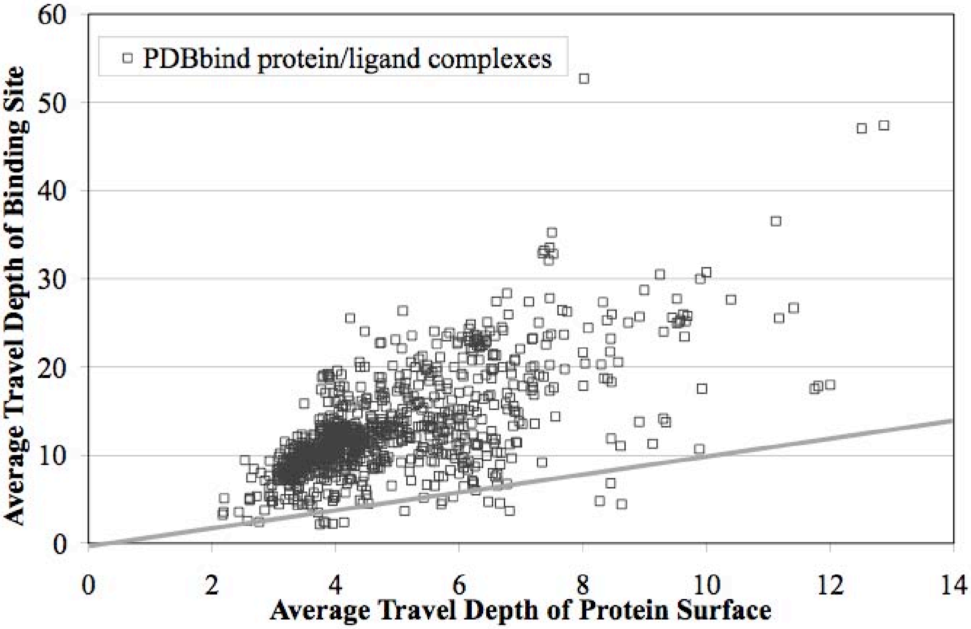


Figure 2-6 Travel Depths of Protein Surface and Binding Site

Average travel depth of entire protein surface plotted against average travel depth of just the binding site, d_b for each structure in the PDBbind dataset. The y=x line is shown on the figure.

2-6. In the figure, for each protein the average depth of the binding site surface, d_b, is plotted against the average depth of non-binding site surface, d_n. The vast majority of the points lie above the x=y line indicated on the figure, demonstrating that the binding site is almost always a pocket, as expected.

Next, the distribution of depths of ligand binding sites was compared to the distribution of the overall surfaces, across the entire 887 complexes. For comparison, a small control dataset with proteins not known to bind any ligands was analyzed as well [70]. We note that 1TGF was left out of the control dataset since it is no longer in the PDB. We removed the waters, ions and buffers found in these control structures for the analysis. The histograms showing the depths of the surface points in each category over the entire dataset are shown in Figure 2-7. The since the number of surface points in each set is so different, the data has been normalized so that the area under each curve is equal. The figure shows there is a clear but not complete preference for deeper points to be near a ligand binding site. Interestingly, the width of the histogram for proteins that bind a ligand is greater than that for the control, 'non-binding' proteins. This indicates that binding proteins tend to have a rougher, or more corrugated surface. This raises the possibility that some proteins are intrinsically more ‘bindable’ than others due to the kind of surface topography they have.

Finally, to calculate the statistical significance that the ligands had some bias to be near deeper surface points, a permutation p-value test was conducted. For each protein, the complete set of surface points was assembled, including both the ligand binding site and the rest of the surface. From this set, a random selection of points

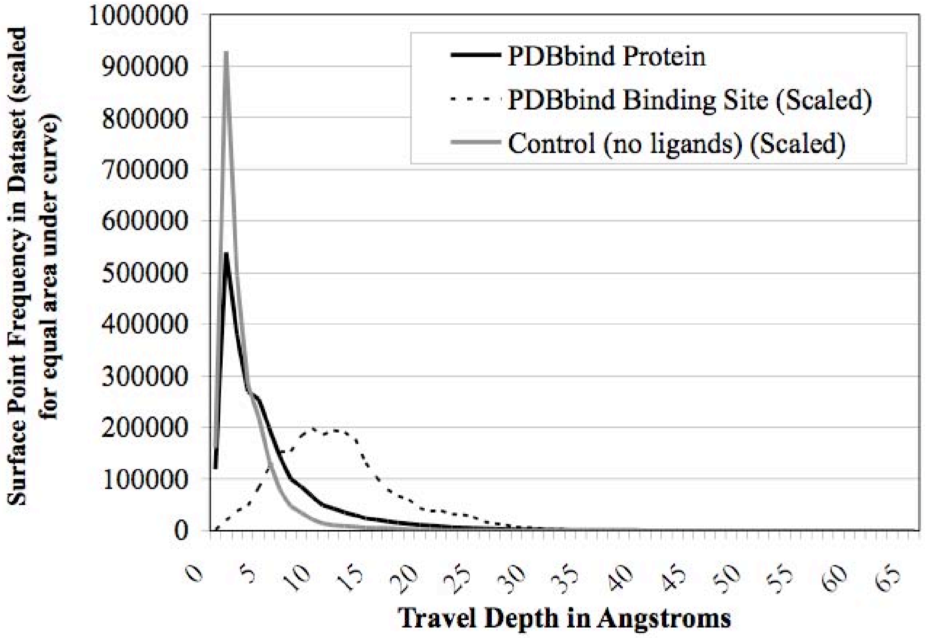


Figure 2-7 Travel Depth Histogram

A histogram comparing travel depths of different structure subsets: A control set with no known ligands, the PDBbind set, and just the binding pockets of the PDBbind set. The control and binding pocket curves have been normalized so the area under each curve equals that of the PDBbind set.

equal in number to those in the ligand binding site was taken, and the average depth found. This random selection was repeated five million times. The p-value is the number of times this selection had greater than the average depth of the true ligand binding site d_b divided by the number of random sets. With five million random permutations, the lower bound on the possible p-value is $2x10^{-7}$. This test gives a good measure of whether the ligand bound to each protein is bound in a deep pocket more often that random. A more complete estimate would use all the potential ligand binding sites on the surface, and calculate the average depth for each. However, generating all the possible ligand binding sites is a rather complicated problem, one which is usually solved by only sampling some of the possible binding sites [71; 72].

The complete results of the permutation tests on the PDBbind dataset are given as Table A-1, along with the average depth of the binding site. To summarize the results, 13 of 900 structures contained ligands that were completely enclosed in cavities, inaccessible to the outside solvent. Excluding those in cavities, 48 of 887 structures had a p-value greater than 0.05, so the remaining 839 structures had ligands which were in significantly deep pockets under this criteria. Under the strictest requirement tested, that of having a p-value less than $2x10^{-7}$, 688 of 887 structures had ligands buried in these significantly deep pockets.

In Figure 2-8 we examine the relationship between protein size and average surface depth using the PDBbind data set. As a robust measure of protein size that could easily be computed for the entire data set we used the total number of heavy atoms. Assuming very similar packing densities for all proteins, number of heavy atoms should be proportional to protein volume. Depth data were plotted against the cube

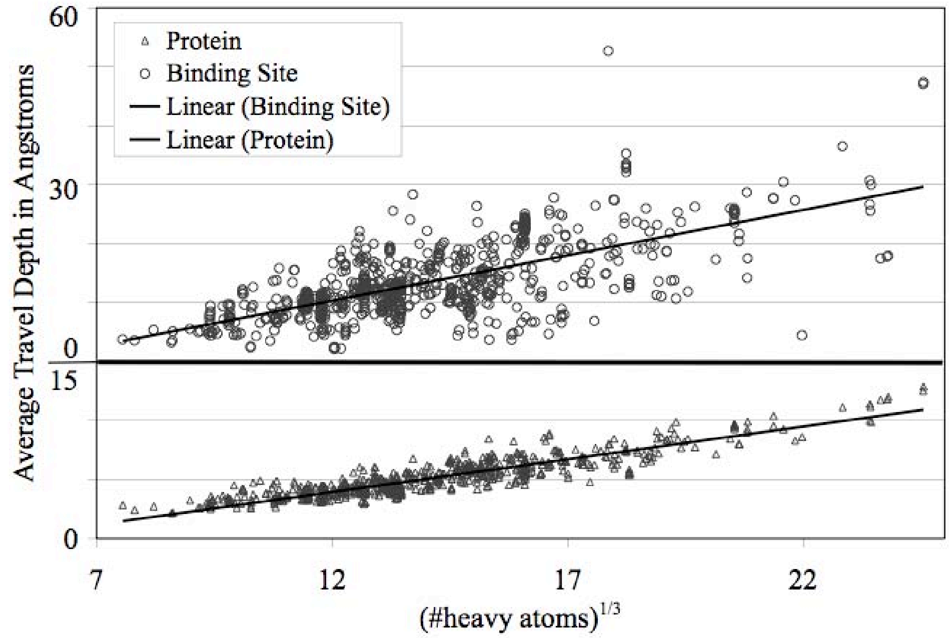


Figure 2-8 Travel Depth Correlated with Size

Average travel depth of the entire protein (Δ) and just the binding site (o) plotted vs. the cube root of the number of heavy atoms, for proteins in the PDBbind dataset. Lines show linear least squares fits for the entire protein (y=0.55x −2.72, R^2=0.84) and for the binding site (y=1.54x−8.13, R^2=0.47).

root of the number of heavy atoms since for a largely globular set of proteins this metric should scale well with the linear dimension of the protein. Indeed, for the mean surface depth averaged over the entire protein surface there is an excellent linear correlation (R^2=0.84). Thus average depth increases linearly with protein size. Not surprisingly, larger proteins can have deeper pockets, but for the *average* depth to increase with protein size larger proteins must also have more pockets of significant depth, i.e. be rougher. The scaling law indicates that average travel depth is an indicator of overall surface roughness, and a good reflection of the fractal nature of the protein surface, as was discovered previously by analysis of surface area[13]. The fact that the fractal nature of the protein surface also emerges from a quite different analysis based on depth provides additional validation of the concept of travel depth. Looking at depth data from just the ligand binding sites, there is still some correlation with protein size, but the significantly smaller variance in travel depth is explained by protein size (R^2=0.47). This may include effects from the smaller amount of averaging involved in using a small subset of the protein surface.

A straightforward question to answer with the binding affinity data from the PDBbind dataset is whether binding affinity of the ligand (-log K) correlates with the average travel depth at which the ligand is bound. *A priori*, one might expect deeper pockets to have greater affinity, based on the idea that a deeper pocket would make more interactions with the ligand. On the other hand, the amount of surface area one could bury or interactions one could make when binding a small ligand is limited by the ligand size. For a given amount of surface burial or number of interactions, one might expect deeper pockets to be less favorable as the long range electrostatic

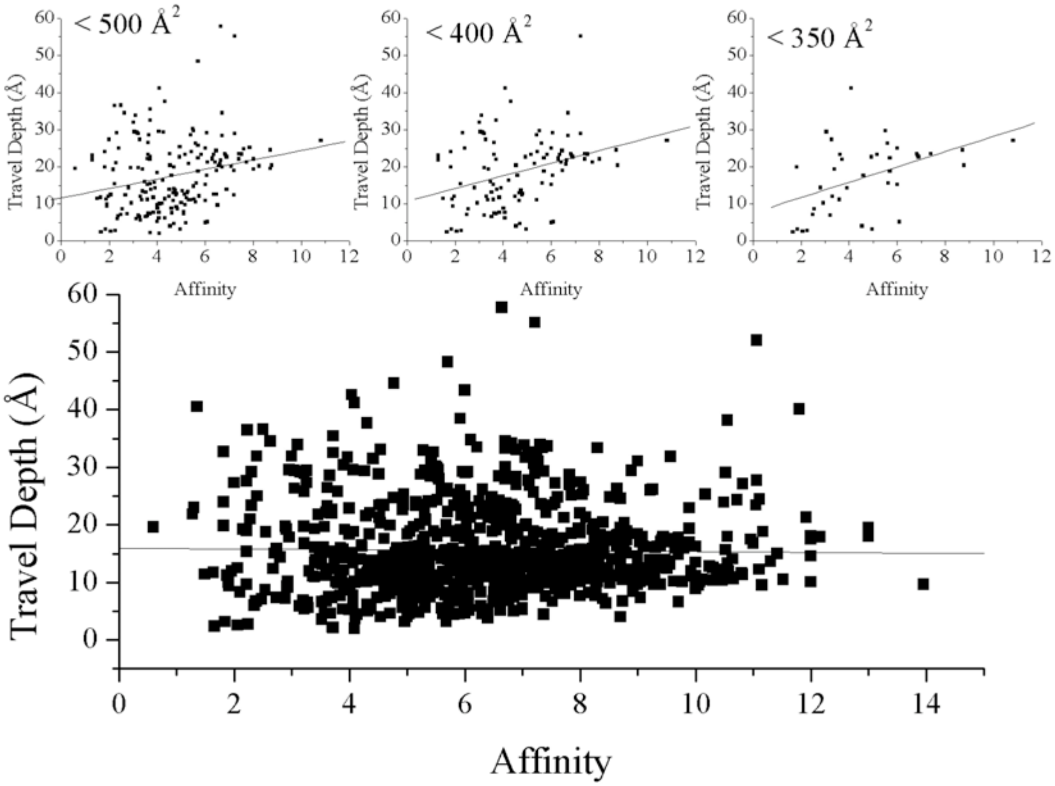


Figure 2-9 Travel Depth versus Affinity

Mean ligand binding site travel depth, d_b, plotted against experimental binding affinity for the PDBbind dataset. Only ligands bound significantly deep (p-value < $2x10^{-7}$) are shown in this analysis. Line shows the linear least squares fit, with (y=-0.39x + 16.5, R^2=0.0199). Inset shows the same plot for complexes that bury less than 500 Å^2, 400 Å^2, and 350Å^2 respectively.

desolvation penalty would be greater. In anticipation of these effects, we computed the change in solvent accessible surface area upon ligand binding, dA, for each of the 887 complexes in the database, in addition to travel depths. The change in surface area was obtained from the surface area of the entire complex minus that of the protein and the ligand alone, calculated using the program SURFCV [62]. Binding affinity, buried surface area and travel depth data for the set of 887 complexes is given in Table A-1. The average travel depth of the binding pocket, d_b, is plotted against the binding affinity in Figure 2-9 for the entire binding data set. A linear fit regression was conducted, and the R^2 values were very close to 0, indicating that none of the variation in binding affinity can be explained by the depth. Even using only those ligands binding in a very significantly deep pocket as judged from the p-value being less than $2x10^{-7}$, no clear relationship is seen. However, if the data is restricted to those ligands that bury less than $500Å^2$ of surface area, there is a significant positive correlation between depth and affinity of R=0.23. Restricting the area burial still further to $<400Å^2$ and then $<350Å^2$ increases the positive correlation to R=0.34 and R=0.47 respectively, with a positive slope of about 2.5 (Inset, Figure 2-9). Although the amount of data at lower areas is sharply reduced, the trend is clearly that affinity depends on depth when the area buried by the ligand is low. This indicates that there is no simple dependence of binding affinity upon depth, because of factors such as surface area burial and no doubt other factors alluded to above.
To extract the broad trend in binding affinity, the affinity data was modeled as a two variable function of buried area and depth, $-\log K = f(dA,d_b)$. Using the 2-dimensional data smoothing function in Origin [73] a piecewise linear approximate $f(dA,d_b)$ was constructed using a 10x10 interpolation matrix with weighted

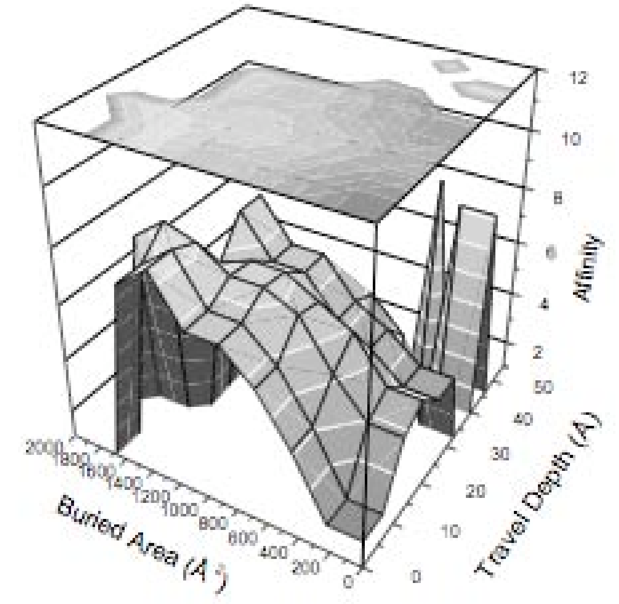


Figure 2-10 Travel Depth, Buried Area, Affinity

The binding affinity from the PDBbind dataset, -log K, described as a function of two variables, average travel depth of the binding pocket, d_b and surface area buried upon binding, dA as –log K = f(dA,d_b). f(dA,d_b) is plotted as both a grey scale colored surface in 3-D and as a grey scale 2-D contour plot in the upper part of the figure. Determination of f(dA,d_b) is described in the text.

averaging. The resulting function f(dA,d_b) is plotted as a surface in the three dimensions of –log K, dA and d_b in Figure 2-10. For additional clarity the figure also depicts f(dA,d_b) as a gray scale contour plot in the upper projection of the figure. Considering first the effect of buried surface area, at small to medium depth the affinity shows an initial approximately linear increase, followed by a plateau. This general trend follows closely that seen in earlier broad surveys of the effect of ligand size [74]. At the greater travel depths, there are very few compounds, and the range of observed buried areas is sharply restricted to small values so no trend is discernable. Considering now the effect of travel depth, for small burial areas greater travel depth does appear to increase the binding affinity. For larger amounts of buried area, the affinity is insensitive to the average binding pocket depth. Overall, the highest binding affinities occur at comparatively low buried surface area and high travel depth, though there is not much data in this region. These results are of course broad trends which 'average out' the effects of specific interactions, shape effects, etc. in each complex, but the analysis demonstrates the kind of questions one can now examine quantitatively with a good measure of depth. As a final interesting note, we show just the mean ligand binding site travel depth against the number of ligand heavy atoms in Figure 2-11.

Considering further the argument that binding pocket depth would primarily affect the polar desolvation contribution to binding, this contribution would be larger for charged ligands such as ions than for neutral ligands. For charged ligands, the desolvation term would be larger for more highly charged ligands. This implies that if such an effect of depth on binding affinity exists, it would be more important for

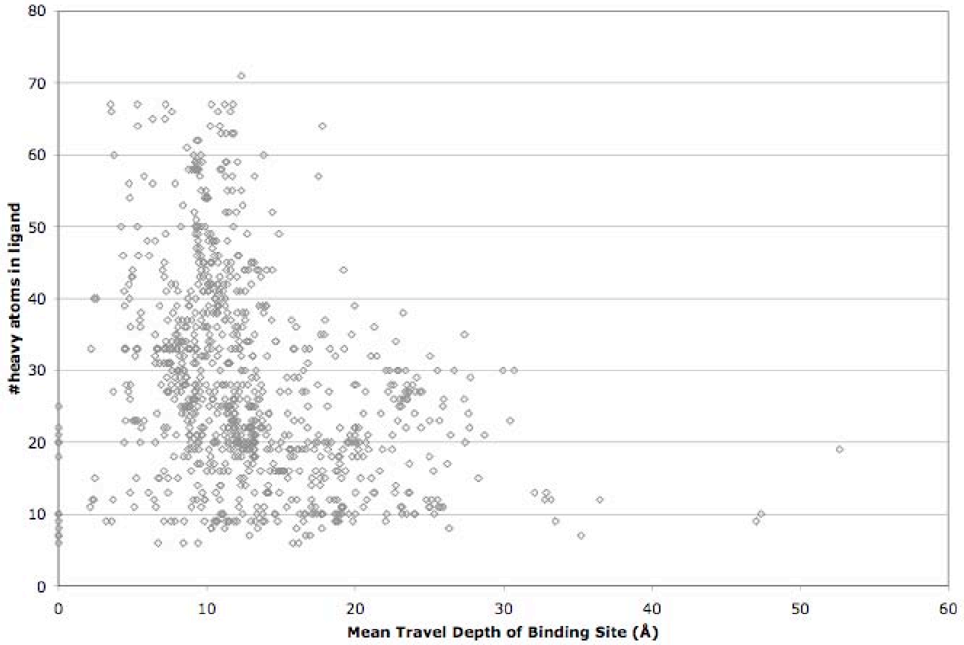


Figure 2-11 Binding Site Travel Depth versus Ligand Size

The mean travel depth of the ligand binding site is graphed against the number of heavy ligand atoms for all PDBbind structures.

divalent ion ligands than for monovalent ion or neutral ligands. Most RNA structures require the divalent ion Mg^{++} to fold and maintain a stable structure, and the growing number of RNA structures with bound magnesium ions allows one to analyze where the magnesium is binding, and how such binding might be related to surface depth. For these reasons we next investigated magnesium binding to RNA structures. We first extracted all RNA PDB entries with magnesium bound. After careful checking of the structures, and eliminating cases where the magnesium ions were bound to non-RNA molecules in the complex, we were left with a set of 29 structures. No pruning by similarity was performed: there are several tRNA structures, several dimerization site initiation points, and several pseudoknots, for instance. The following PDB codes were in our final set: 1EVV, 1F27, 1FIR, 1I7J, 1I9V, 1K9W, 1KXK, 1O3Z, 1TN2, 1TRA, 1XP7, 1XPE, 1XPF, 1Y73, 1Y95, 1Y99, 1YKV, 2B8R, 2B8S, 301D, 310D, 3TRA, 430D, 462D, 468D, 469D, 470D, 471D, 4TNA [16]. These structures contained 249 magnesium ions that were bound to RNA, or closer to RNA than to other molecules in the PDB structure. We broke down surface points on the RNA structures into five types, based on the nearest atom. The five types are phosphate, sugar, and three nucleotide groups, the major groove, minor groove, and other. Non-standard nucleotides found commonly in tRNA were grouped in with the other group in the analysis, since they do not have standard hydrogen bonding patterns and usually do not conform to the major/minor groove distinction.

Figure 2-12 shows an example tRNA structure with one magnesium bound, PDB Code 1FIR. The color scale on this runs to 17.6 Å in depth at the deepest blue. Figure 2-13 shows the distribution of surface depths where magnesium ions are bound

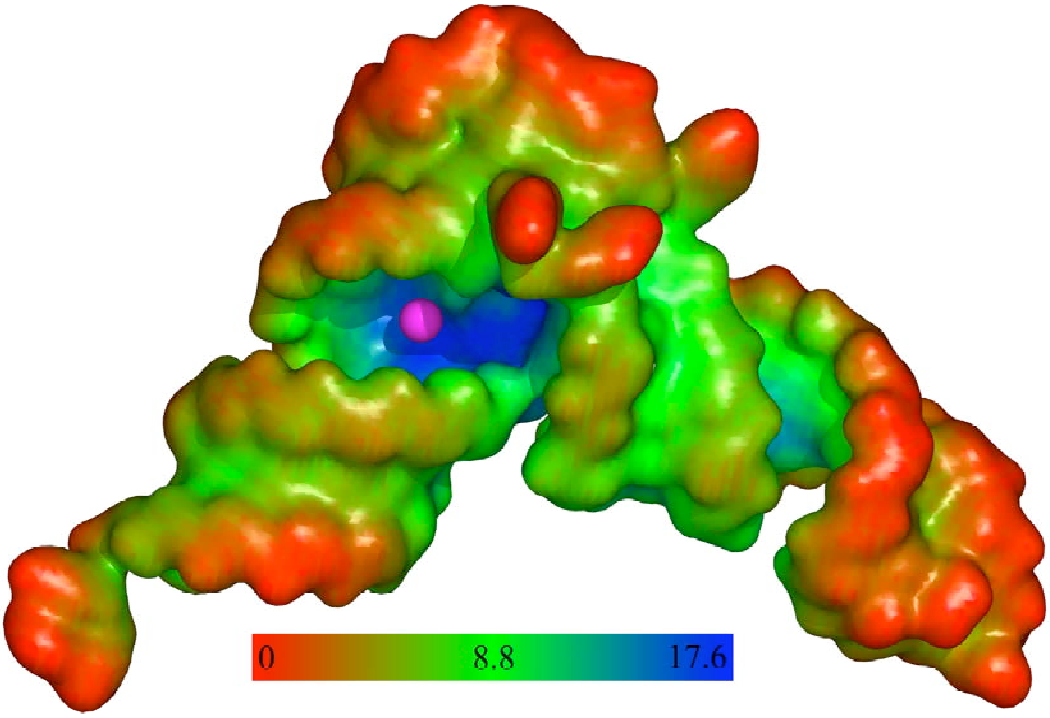


Figure 2-12 Travel Depth of tRNA

An example tRNA structure (PDB code 1FIR). A magnesium ion is shown as a purple sphere.

within 4Å of surface points, broken down into the five categories. Figure 2-14 shows the relative frequency of binding at each depth and category data by normalizing the curves in Figure 2-13 by the number of surface points at each travel depth in each of the five categories. The normalized data is cut off at depths > 13 Å since more than half the category/structure combinations either had no data representing that level, or the number of points was so small as to be statistically insignificant. The major feature from this analysis is the significant amount of Mg^{2+} binding near major groove atoms at a depth of 9 Å. This is apparent even without normalization. The binding of magnesium to the major groove at a travel depth of 9 Å is present in 18 of the 29 structures in our sample. These 18 structures represent a variety of structures in our limited test set, as do the 11 structures that do not contain magnesium bound at that travel depth.

The relatively high frequency of magnesium binding to the phosphate backbone category at depths ≈ 4-12Å seen in Figure 2-14, appears somewhat significant in the context of the RNA structural database available at this time (Figure 2-13). Looking at the overall frequency/depth distribution without regard to category, one can see a fairly uniform distribution of depths from 0-12Å, though the three peaks, two for phosphate regions and one for the major groove seem significant. The relatively uniform distribution, with ions occurring quite frequently at depths up to 10Å indicates that there is little depth dependency to the desolvation penalty. This probably follows from the fact that RNA structures tend to be quite open and highly solvated compared to protein structures, even in deep pocket or groove regions, as exemplified by tRNA in Figure 2-12. As described previously, the

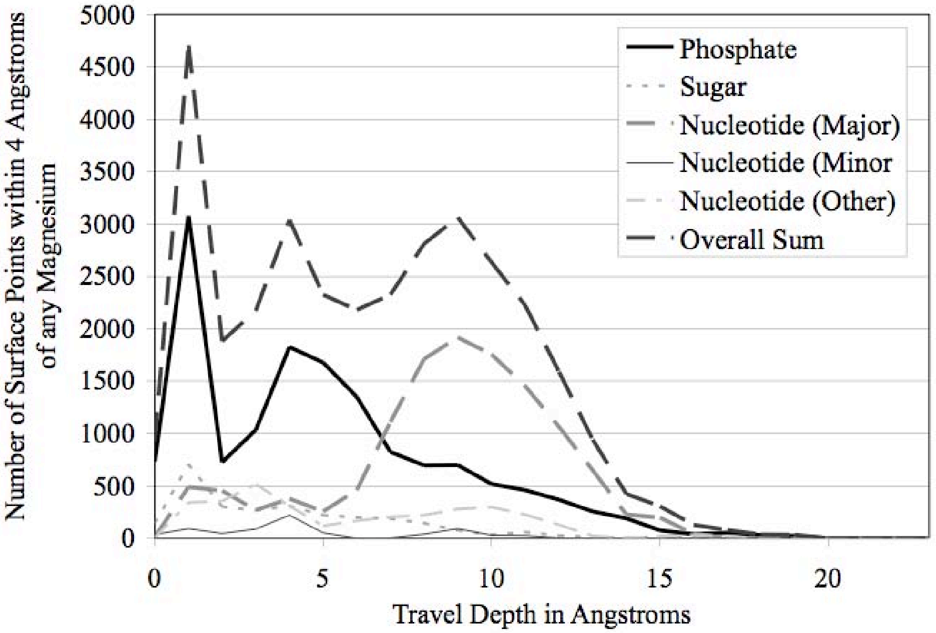


Figure 2-13 Travel Depth of magnesium binding surface in RNA structures

Frequency of each of 5 classes of surface points at given travel depths. Points are counted if within 4Å of a magnesium ion.

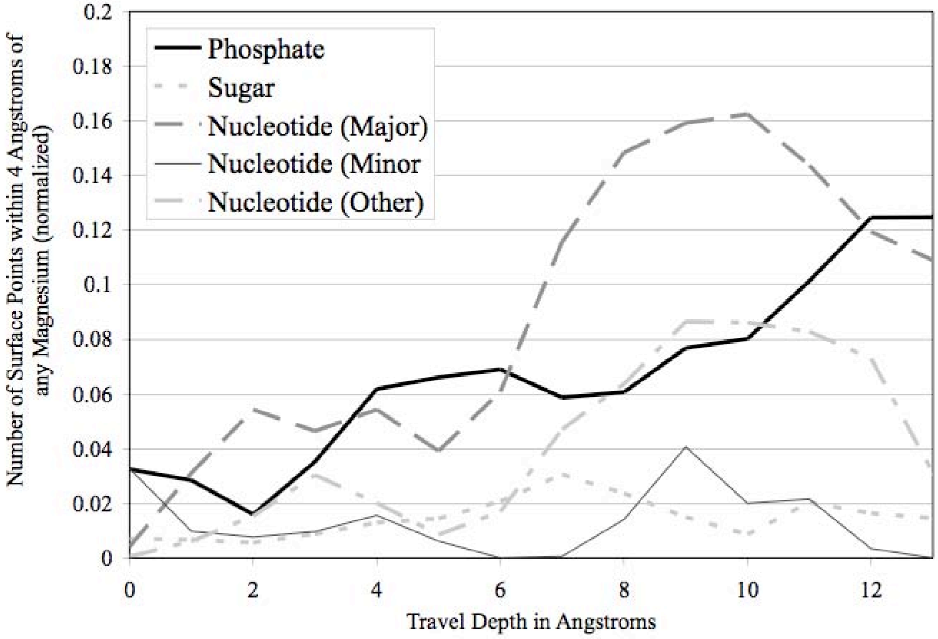


Figure 2-14 Normalized depth of magnesium binding surface in RNA structures

Points are counted if within 4Å of a magnesium ion. Frequency of each of 5 classes of surface points at given travel depths, each point was normalized by the overall number of surface points of that class at that travel depth. Data is only shown where at least half the classes/structures had data at that depth, in other words equal to or below 13Å.

irregular shape of the molecular surface influences the electrostatic potential, creating pockets away from the immediate vicinity of the phosphate groups where cations are likely to bind [75]. Our analysis supports the idea that these pockets occur relatively often in major grooves, and they are distributed around a specific travel depth of about 10Å.

As a final example of the use of travel depth we consider the problem of automatically identifying ligand binding pockets. This is a difficult problem, and the latest methods, which rely to a large extent on pocket volume, still encounter problems identifying extremely buried pockets [76]. Deeply buried pockets are often not the largest by volume. A different way of solving this problem involves clustering surface points within some distance of the centroid of the protein atoms [77], which is another way of incorporating depth information. The centroid method requires careful selection of the appropriate domain or subset of atoms to give sensible results, and so it is not straightforward to apply in large, multi-domain proteins. We examine the case of the FS4 cluster ligand binding site in PDB entry 1H2R, which is reported as problematic [76]. There are five separate ligands: three different iron-sulfur clusters, a nickel-iron active center and a magnesium ion. This protein's overall average depth of molecular surface is 8.7 Å. Examining this depth in terms of our earlier analysis of protein size against average travel depth, this is more than an angstrom deeper than the trendline. This means, for a protein of this size a rugged surface. Our travel depth algorithm gives average depths of the ligand binding sites as 27.6 Å, 22.9 Å, 38.5 Å, 18.6 Å, and 42.1 Å respective to the five ligands indicated in the structure by the abbreviations: FS3, FS4_1, FS4_2, MG, NFE. Figure 2-15

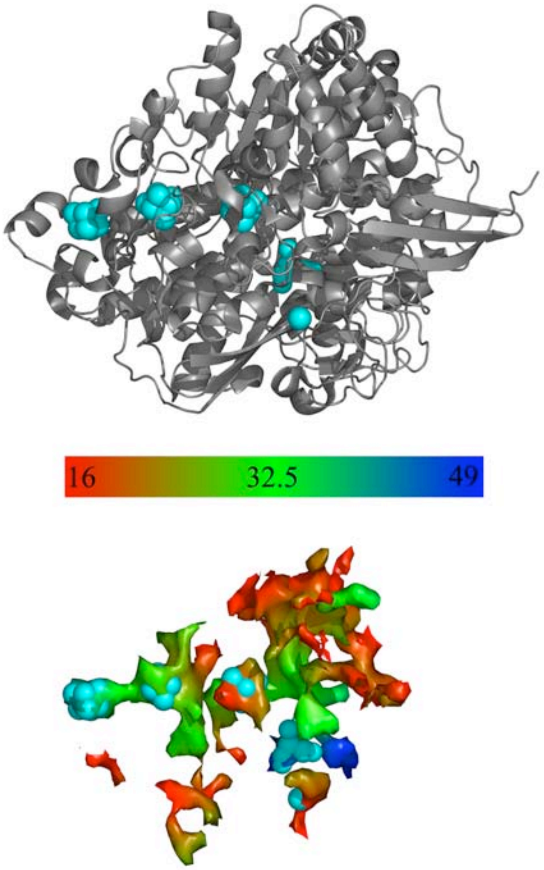


Figure 2-15 Travel Depth of Very Deep Pockets

A sample case where pocket volume does not correlate well to ligand binding sites [76]. PDB code 1H2R, NiFe hydrogenase is shown in two views. In one, gray ribbons are shown with the 4 ligands and magnesium ion in light blue. The other view, from the same perspective shows the surface colored by travel depth, only the surface with travel depth greater than 16Å is shown, cavities have also been removed for clarity.

shows this structure and the travel depth colored surface. All ligands except magnesium are significantly deeper than the average depth, and they would clearly demark these regions in a blind search. However it is notable that the deepest pocket contains no ligand. This example suggests that combining depth and volume would significantly improve binding site identification in cases where pocket volume alone fails.

Discussion

We have introduced here a quantitative, robust and useful definition of the depth of any region of a triangulated surface of a molecule. We have also implemented this definition with an approximate, though sufficiently accurate and fast algorithm. This implementation is suitable for quantitatively analyzing individual molecules or large databases of molecules. The algorithm satisfactorily quantifies binding pockets in proteins as intended. Interestingly, travel depth also works for two difficult cases for which it was not specifically designed. The first is for grooves in DNA, which present an interesting case since the grooves are unbounded in one direction. Second, our algorithm works in the case of tunnels, that is pockets that have no 'bottom', but go through the entire macromolecule.

Our definition of travel depth differs significantly from depth measures used in previous work. Other definitions have been proposed based on the difference between molecular surfaces of varying smoothness. GRASP [24] has a macro called Molecular Elevation which produces the difference between the normal molecular surface and a molecular surface generated with a probe radius of 10 Å [78]. APROPOS used a smoothed Euclidean difference between two alpha-shapes to locate binding sites [70; 72]. To define the reference surface, we use the limiting form of the molecular surface with infinite probe size, the convex hull. Both the GRASP macro and APROPOS program compute a simple Euclidean distance, ignoring the complicated surface structure of the macromolecule. The travel depth defined here, in contrast, uses a non-Euclidean, macromolecule-avoiding distance. While a Euclidean distance agrees with our definition in cases where the paths to the convex hull are simple non-macromolecule intersecting straight lines, it differs when the macromolecule contains overhangs and narrow tunnels to interior binding sites. This occurs frequently in proteins: over all surface points in the PDBbind dataset used in our analysis, about 52% of the surface points had a higher travel distance than Euclidean distance, and 5% of the surface points had a difference above 5Å. Moreover, most of the large errors occur in pockets, which are the regions of most interest. The APROPOS definition of depth is also not taken from molecular surface points, and has been highly tuned to detect binding sites. Our method is more general than these; it can calculate depths for any molecular surface, it works for pockets of any size and complexity, it can also calculate depths for the volume between the convex hull and the molecular surface.

Additionally, our definition is quite different from the notion of Extreme Elevation [79]. The extreme elevation is a height distance between any two points on the surface, and the algorithm finds all points that are local maxima of such a function. These pairs of points that maximize the elevation could be used in some applications, however it does not define a general notion of depth for every surface point, as a point could be in several pairs. The extreme elevation method also has a high asymptotic complexity, but has been shown to help generate possible poses in docking applications [80].

Our algorithm for solving the shortest path problem is significantly different from previous work. The related shortest path planning problem in three dimensions was shown to be in the class NP-hard with respect to the obstacle complexity [29], however some approximation algorithms have been proposed [81; 82]. The previous approximation algorithms usually subdivide each edge of the polyhedral obstacle into smaller pieces, then compute visibility maps among vertices. The code required to compute visibility maps in three dimensions is complex. Moreover, the computation time is large. Other previous approaches, which we did not use, are conformal or constrained meshing, fast marching methods, and proximity depth. If a reasonable quality conformal mesh could be generated between the molecular surface and the convex hull it would be easy to apply multiple source shortest paths to compute minimum travel distances [83]. There is a large amount of previous work on fast marching methods, algorithms to grow expanding boundaries. Again, the algorithms and code to implement these approaches are complex [84]. Also, producing the intermediate conformal mesh given an arbitrary triangulated molecular surface may be difficult unless constraints are applied to the latter. Such constraints may impose undesirable compromises on the type and resolution of molecular surfaces that could be handled. In this context, our minimal travel distance algorithm can be viewed as a discretized fast marching method, or a approximation to a conformal mesh generator constrained to a cubic lattice, although for our purposes it need not fully implement either of these intermediate constructs. Finally, our work is most similar to the notion of proximity depth, which has been developed for various proximity graphs where edges exist between close points [85]. In contrast we generate our points and weighted connectivity in different and explicit ways, and specifically exclude certain edges, those passing through the macromolecular surface.

In its actual implementation, our travel depth approximation algorithm has several advantages: It works on any orientable and connected triangulated surface, it is relatively easy to code, it has polynomial complexity and it has practical computation times. The result is at most a constant multiplicative factor worse than the true travel depth and this constant can be controlled to a degree. In application to several different kinds of macromolecular surface, including analysis of DNA groove depth, our measure agrees with previous qualitative descriptions, and one's intuition when looking at structures.

The significance of having only polynomial complexity and practical computation time is that the algorithm is practical for high-throughput analysis for large macromolecules. Our analysis of the PDBbind database of 900 protein-ligand complexes required less than a week of computation time on a single computer, encompassing all aspects of the computation: surface generation, travel depth computation, and all statistical analyses.

In analyzing the protein database, the travel depth algorithm revealed two important features. First, proteins with known ligand binding sites have a different depth distribution profile than those without known binding sites. Those with ligand binding sites have a wider profile at low depths with a higher tail at high travel depths. Second, protein size as a function of the cube root of the number of atoms explains a lot of the variation in overall travel depth from protein to protein; A clear linear correlation between depth and size is present.

The travel depth analysis was used to show that ligands tend to bind in deeper pockets. Moreover, when this analysis is combined with surface area calculations, it shows that binding in deeper pockets has a significant effect on binding affinity when surface area burial is low.Though the picture is not yet completely clear, this specific two factor effect has not been suspected before. As is apparent from the contour depiction of f(dA,d_b) in Figure 2-10 there are no complexes with both large buried surface area and great depth. Again this is unanticipated, and it may reflect some intrinsic constraints for good ligand binding in proteins. Although the PDBBind database is quite large, with diverse ligands and protein sizes, the lack of large dA/d_b could also reflect limitations of the PDBbind dataset. If complexes with this combination of dA and d_b, are discovered, along with more structures with low buried surface area and high travel depth, the relation between dA, d_b and binding affinity could be better understood. As confirmation, the mean Travel Depth against ligand size calculated as number of heavy atoms is shown in Figure 2-11, which confirms that large ligands do not bury deeply, at least in this dataset.

In addition, we analyzed the influence of travel depth on magnesium binding to RNA structures. Our preliminary conclusion is that there is little effect of desolvation in deeper ion binding sites. Also, perhaps surprisingly, a significant number of Mg ions bind closer to the major groove than to the phosphate groups, in contrast to one's naïve expectation based on charge complementarity. Since the RNA structural database at present is rather small compared to that of proteins, more data is needed to make the picture clearer.

It has been well established that the substrate binding and enzyme active site of a protein is commonly located in the pocket with the largest or second largest volume [58; 72]. However, there are some cases where neither of the largest pockets by volume contain the enzyme active site [86]. Some of these cases are peptide recognition sites that are commonly spread across the surface of the enzyme. However, other cases where the ligand lies in a deep, small pocket may benefit from taking depth of the pocket into account [86]. Our general analysis showing that many structures have ligands that bind in significantly deep pockets reinforces this conclusion. These issues are part of the larger problem of automatic identification of ligand binding sites. We considered one difficult example in this area [76] as an illustration of how the travel depth analysis can help. In addition, travel depth may help in the further problem of discriminating between different kinds of active sites.

Having a quantitative definition of travel depth also now allows one to combine this property of the surface with other features for analysis. Other surface features include volume, surface area [62; 72], curvature [87], and chemical features like electrostatics [88]. Together with sequence properties like conservation [76; 89], these combined analysis of all these features may allow for excellent overall prediction of ligand binding site location. A recent example of this kind of analysis shows the importance of having a good fast quantitative definition of depth. [90]. Depth analysis would also be useful in structural genomics projects where little functional information is known about a new structure.

Future Directions

The algorithm we develop here for measuring depth of macromolecular surfaces is a discretized approximation of a multiple shortest paths (MSP). The paths are initiated at the convex hull, and terminated at the molecular surface. We chose the convex hull as a natural and practical reference point to initiate the MSP, but we note that other choices are possible. Among the class of convex surfaces, an ellipsoidal or spherical surface completely enclosing the molecule is another possibility for a reference level. Some choice must be made of how far outside the molecule this surface, however, which introduces another arbitrary parameter. One expects that an ellipsoidal initiation surface suitably chosen and aligned to the molecule's axes of inertia would provide much the same rank ordering of depths as the convex hull. Alternatively, as a non-convex shape, one could use a molecular surface created with a large probe radius as the reference level.

Additionally, varying the probe radius used to generate the molecular surface would be another straightforward variant of our method. Our algorithm works for any triangulated surface with reasonable constraints. A larger probe radius might mimic, for example the larger size of a ligand molecule groups compared to water. Indeed, the default use of water-sized probes to create the molecular surface is a standard caveat in this area. For instance, minimal distance paths may travel through water tunnels to active sites, though the path a larger ligand may travel would be different, and probably longer.

More generally, our travel distance implementation of MSP could be used as a measure of shortest avoiding distance in a variety of applications to macromolecules. For example:

1. In cases where a particular protein pocket is accessible by more than one pathway or 'tunnel', traveling back along the steepest descent of travel depth values, or simply recording the last step taken to arrive at each grid cube, would provide the shortest 'escape' route and its length.

2. Examining the union of all such escape routes from a ligand binding site could also give interesting information, for instance, examining the number of grid cubes with varying depth from 0 upwards could yield information about the steepness or width of the tunnel.

3. Taking the molecular surface as the initiation surface, and propagating the travel distance inside the molecular surface until it self terminates (when all the grid points inside the surface have been assigned) the algorithm would assign a depth, with respect to the nearest surface point, to every part inside the molecule. Applications of this depth include quantifying the depth of burial of side chains in a protein core. This analysis would be similar to the notion of atom depth [18] and likely yield similar results. Identification of the peaks and ridges of this burial depth would provide an approximation of the

medial axis of the molecule surface. The medial axis is a powerful, but hard to compute construct and descriptor of surfaces [27].

4. Picking one part of a molecular surface as the initiation, the other parts of the molecular surface as terminators, and propagating the paths outside the molecule, the travel distance would provide the shortest distance between two sites on a protein. Application to the analysis of substrates that must diffuse between different sites on multi enzyme complexes [91; 92] suggests itself. Elaborating further, by choosing every site of the protein in turn as the initiator, a two-dimensional matrix of minimum avoiding distances between every pair of surface patches can be built up, providing a detailed descriptor of the surface topology.

5. In the case of ion channel proteins and other pore or tunnel containing proteins, if the entire molecular surface is used for initiation, and the paths propagated outside the surface, paths will either self terminate in the tunnel, or can be truncated a suitable large distance from the surface. The 'ridge' of maximum distance (essentially an everted equivalent of the medial axis with the role of inside and outside exchanged) will run through the tunnel or pore. This ridge identifies both the locus of the center of the pore, and its width. This could provide an alternative to the standard algorithm to automatically characterize the ion channel pores [40; 41].

In summary, we have introduced a quantitative measure of molecular surface depth called travel depth. Depth, though an intuitive concept, is in fact hard to define and calculate *prima facie*. We show here it can be calculated in an efficient manner for many types of macromolecules (DNA, RNA, and proteins), and that it works on a variety of surface topographies (channels, tunnels, pockets). The ability to quantify a key surface property, depth, allows us to address several interesting questions about macromolecule shape. These include a quantitative analysis of groove depth in DNA and RNA, the relationship of pocket depth to binding affinity, the relationship between protein size and average surface depth, and the automatic identification of binding pockets. In conclusion, we hope that this measure of travel depth will be a useful tool in many areas of structural analysis of biomolecules.

Chapter 3

This chapter was previously published [43].

Summary

We describe a new algorithm, CHUNNEL, to automatically find, characterize and display tunnels or pores in proteins. The correctness and accuracy of the algorithm is verified on a constructed set of proteins, and used to analyze large sets of real proteins. The verification set contains proteins with artificially created pores of known path and width profile. The previous benchmark algorithm, HOLE, is compared with the new algorithm. Results show that the major advantage of the new algorithm is that it can successfully find and characterize tunnels with no *a priori* guidance or clues about the location of the tunnel mouth, and it will successfully find multiple tunnels if present. CHUNNEL can also be used in conjunction with HOLE, using the former to prime HOLE, and the latter to track and characterize the pores. Analysis was conducted on families of membrane protein structures culled from the protein databank as well as on a set of trans-membrane proteins with predicted membrane-aqueous phase interfaces, yielding the first completely automated examination of tunnels through membrane proteins, including tunnels that exit in the membrane bilayer.

Introduction

Proteins adopt three-dimensional structures with complex shapes and surface topography. These topographical features, such as clefts, flaps and tunnels often have important functional roles. We define here the term tunnel or pore to mean a hole that goes completely through the protein, thus having two entrances or mouths. Many proteins contain tunnels or pores that are of physiological importance, the primary examples being membrane protein ion channels, pumps, porins and transporters. While some channels have a single simple tunnel structure, there are also more complicated structures, for example the mechano-sensitive channel of small conductance (MscS)[93]. Also, proteins like the ring clamp protein [94], the ribosome [95] and other proteins involved in transcription have topological features including pores that are important for interactions with DNA strands. Spastin has a central pore which is involved in microtubule severing by pulling the end of the tubulin polypeptide through the pore [96]. Some enzymes like rubisco also have tunnels through them [97]. At least one enzyme, acetylcholinesterase has a tunnel observed under simulation with distinct exits for the two products [98]. Photosystem II has three tunnels leading to the active site, theorized to be pathways for water, oxygen and protons [99]. Finding, cataloging, and measuring these tunnels is important in understanding their function. The ability to do this automatically is an important step towards automation of structural genomics, or characterizing new protein structures. While less than 400 high-resolution structures of trans-membrane proteins are currently known, and of these only about 150 are unique [38], many advances in techniques should increase this number [100], particularly as membrane proteins become targets of large scale structural genomics projects [101] Comparisons to the growth of globular proteins in the PDB suggest around 2200 membrane protein structures will be deposited by 2025 [37]. Additionally, as new examples of subclasses of membrane protein structures are found, accurate homology modeling studies become possible [102]. Tunnel analysis will increasingly be needed as these structures will no doubt include many new pumps, pores, channels and transporters.

The seminal work in characterizing protein tunnels was the development of the HOLE algorithm [40]. The algorithm has been applied very successfully to analysis of ion channels, in which the position and orientation of the pore (normal to the membrane) is known *a priori*, and can be used to 'prime' the HOLE search algorithm. The algorithm is less able to deal with arbitrarily positioned tunnels or multiple pores, and it is difficult to automate since it needs some initial user guidance. Additionally when multiple tunnels are present, HOLE or variations of HOLE were not able to find the 'correct' tunnel among several in some ribosomal structures [95]. There has been some work in calculating cavities and their volumes or volumes of portions of tunnels [95; 103]. Additionally, CAVER functions like a 3D version of HOLE in some respects, but it still needs a starting hint to find a tunnel, and it is primarily geared towards finding paths out from a pocket, not tunnels all the way through proteins [36; 104]. However, no further work in automatically identifying tunnels has taken place since the introduction of HOLE. This attests to the difficulty of developing a completely automated, general tunnel finding/measuring algorithm. We present such an algorithm, which we call CHUNNEL, then describe the principles of both topology and geometry on which it works. We then test CHUNNEL on a set of proteins with artificially generated pores of known path and width, and on various membrane proteins with tunnels from the PDB database [38; 42]. Tests of the HOLE algorithm were also performed on the same test set in order to compare the two algorithms, and show that CHUNNEL has a markedly improved ability to find tunnels automatically. We also show that CHUNNEL can be used to prime HOLE, which can then trace and characterize the pore. We also use CHUNNEL to find qualitatively new tunnels, for instance those that exit within the membrane bilayer, which have not been found or examined previously.

Methods

General outline of the approach

The procedure developed here for finding and characterizing tunnels is an outgrowth of our previous work characterizing depths of pockets, grooves, tunnels and other surface features in macromolecules using a measure known as Travel Depth [35]. The Travel Depth of a point on the molecular surface is defined as the shortest path through the solvent to that point from a reference surface (specifically the convex hull of the protein). The shortest paths algorithm [23], specifically the generalization we call multiple source shortest paths (MSSP) [28], is used to compute the travel depth, and it is implemented by discretizing space on a cubic grid. Following the application of the MSSP algorithm all surface points have been assigned travel depths [35]. In addition, the travel depths of all solvent grid points lying between the convex hull and the molecular surface are known.

The impetus to develop a tunnel-characterizing algorithm from this work had two sources. First, although the Travel Depth algorithm was designed to characterize pockets and clefts, an unexpected benefit is that it also measures the depth of both the lumen and the surface of a pore [35]. Second, the MSSP algorithm proves to be a general-purpose algorithm for calculating volume avoiding, shortest distance pathways. If the MSSP algorithm is started at the molecular surface, and the distances are propagated outwards in the solvent, then the 'Travel Out' distance assignment will self terminate in tunnels, forming a 'ridge' or everted medial axis in 3-dimensions. These two observations suggested that by starting at a maximum in Travel Depth and Travel Out distance, and following ridges in Travel Out distance of decreasing Travel Depth in two 'opposite' directions one would trace out the path along the center of a pore. The Travel Out distance along this path gives the radius of the pore at each point. In practice, using just these two distance functions it is difficult to automatically distinguish the difference between the bottom of a pocket and the center of a tunnel. It is also difficult follow a ridge of distance in three dimensions, especially with the discretization of space required to implement any algorithm. This problem, sometimes referred to as thinning, shape skeleton or medial axis, is complicated even in two dimensions [105; 106; 107] and can only be approximated in three dimensions [108]. Hence to implement this approach it is necessary to first ensure the starting point is in a pore, and then correctly follow the pore out in both directions. In addition, if there are multiple pores, one needs to reliably identify starting points and propagation 'directions' for all of them. We achieve this through topological and geometric analysis of the molecular surface.

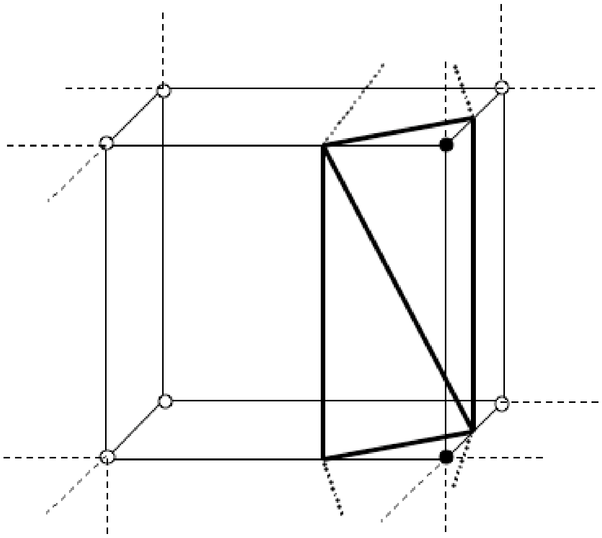


Figure 3-1 Surface Triangulation

Part of the triangulated surface passing through a grid cube. Of note is that all surface points lie exactly between two grid points.

Generation and preprocessing of the surface

We start with the generation of the molecular surface (MS), using the algorithm in the GRASP macromolecular graphics program [24] implemented as a stand-alone program. Standard atomic radii [3] are used to generate the MS with a probe radius of 1.2Å. This is a somewhat smaller probe radius than used previously, in order to treat ion channels: The permeant ions can have radii less than the standard probe radius of 1.8Å used for water. The modified GRASP surfacing algorithm first maps the molecule onto a cubic grid. It then produces a closed triangulated surface, for which the vertex coordinates, vertex connectivity, triangle normals and triangle connectivity are known. All cavities, defined as smaller disjoint sets of connected triangles, are discarded. In addition, because of the way this surface is generated, the volume inside and outside the molecular surface is already discretized on a cubic grid whose vertices are labeled as in or out (Figure 3-1). The vertices of the surface triangles also lie on edges joining inside and outside vertices of the volume grid, while triangle edges cross the surfaces of grid cubes or lie completely within a single grid cube (Figure 3-1). This well defined relation between surface and volume discretization is key to the successful implementation of the tunnel finding algorithm, as the latter uses both surface and volume properties. The final step in the surface generation/preprocessing is to generate the Convex hull, using the Qhull algorithm [25], which also generates a closed, triangulated surface.

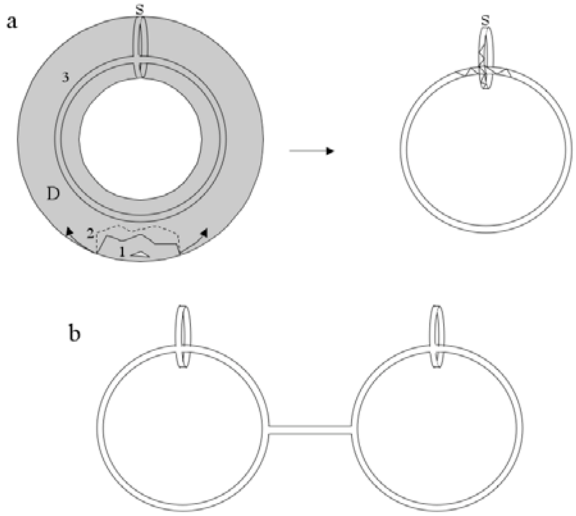


Figure 3-2 Expanding Discs

A) A 1-torus showing the original starting triangle for ‘disc’ region D (1), a partially region D (2), and the final maximally expanded region D and the corresponding leftover minimal strip S (3). The minimal strip S is also shown separately for clarity. B) A minimal strip S for a 2-torus.

Enumeration and localization of pores

Triangulation of the molecular surface (after discarding cavities) immediately provides the number of tunnels or handles present, through the Euler relation:

$$V + F - E - 2 + 2N = 0 \tag{3-1}$$

where V, F, E are the number of triangle vertices, face, edges respectively, and N is the number of handles, so the surface is an N-torus. Although the number of tunnels is known from this topological invariant, there is no indication of their location. With a complex protein surface, it is often difficult to find them even using 3-D modeling graphics.

The first step to localization of the tunnels is to 'remove' from the surface a maximal region of triangles, D, that is topologically equivalent to a disc. A triangle is picked at random to start D, and neighboring triangles are removed until it is impossible to remove another triangle and have the boundary of D remain a simple, closed, non-intersecting path, Figure 3-2a. The remaining triangles form a closed strip of triangles, S, one triangle wide with 2N loops. The loops come in N pairs of which one runs around each pore (an A-loop), and one runs through each pore (a T-loop). Figures 3-2a and 3-2b shows a residual strip S for a torus (1-torus) and for a 2-torus.

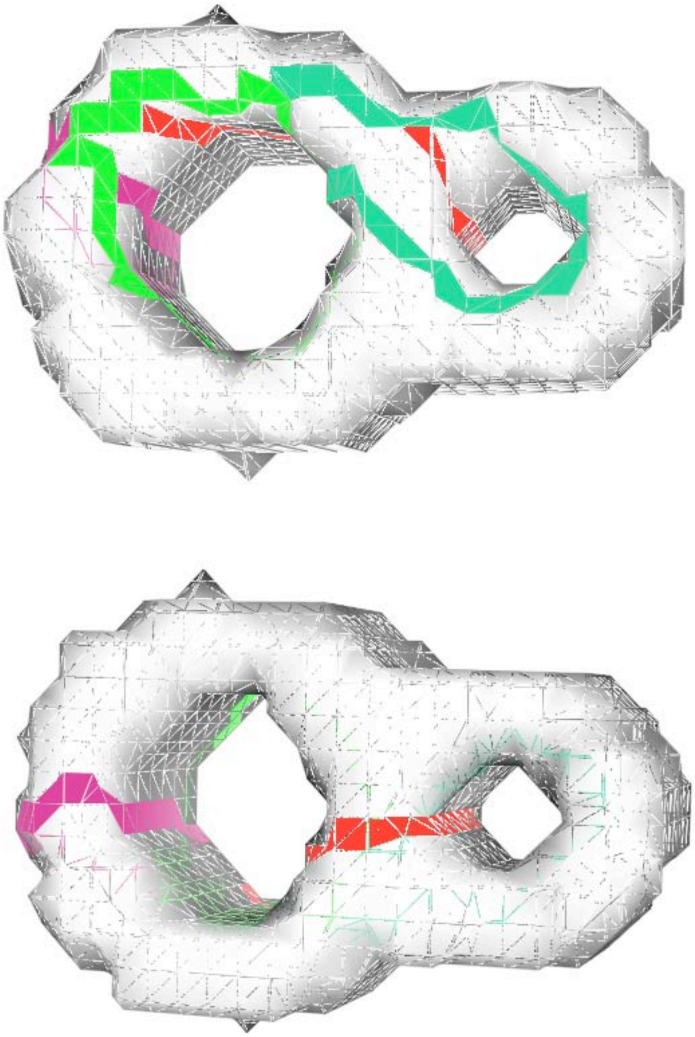


Figure 3-3 Example 2-torus

Front and back views of a real 2-torus and the resulting strip S broken into 4 colored loops, showing how the loops meander over the surface.

On a complicated protein surface, the path of S is usually very irregular and far from minimal in length (Figure 3-3). This divagation is usually great enough that one cannot at this point reliably categorize a loop as A or T just from the coordinates and orientation of the constituent triangles. In particular, there is no requirement for the A-loops to be anywhere near either the center or the narrowest part of a pore.

Obtaining a ‘tight’ loop of triangles around a pore

The next step is to regularize or 'tighten' S around the pores, and then find a set of N A-loops that are topologically distinct and go around each pore in the surface. A careful combination of topology (to ensure that the A-loops found are distinct) and geometry (to ensure that such loops are tight) must be employed to accomplish this goal as neither approach by itself would work. First the triangles of S are decomposed into 2N sets S_L, L=(1...2N) one for each loop. (Some triangles may be part of more than one loop). Using the MSSP algorithm, neighboring triangles are sequentially added to a loop S_L (it is 'fattened up') until its edges wrap around and meet at some point (Figure 3-4a). Because triangles are added in order of minimum neighbor distance from the original strip one can trace back neighboring triangles from the meeting edge along the shortest path to S_L. The set of trace-back triangles form another one triangle wide strip S'_L which is the complement of S_L: If S_L is an A-loop, then S'_L is a T-loop, and *vice versa*. At this point one can automatically and reliably classify such a loop as A-type or T-type from its triangle surface normals, by checking whether they point toward each other (A-loop) or away from each other (T-loop). A ‘regularized’ A-loop runs around the narrowest part of a pore because of the shortest paths property of the MSSP and so it more tightly delineates a pore.

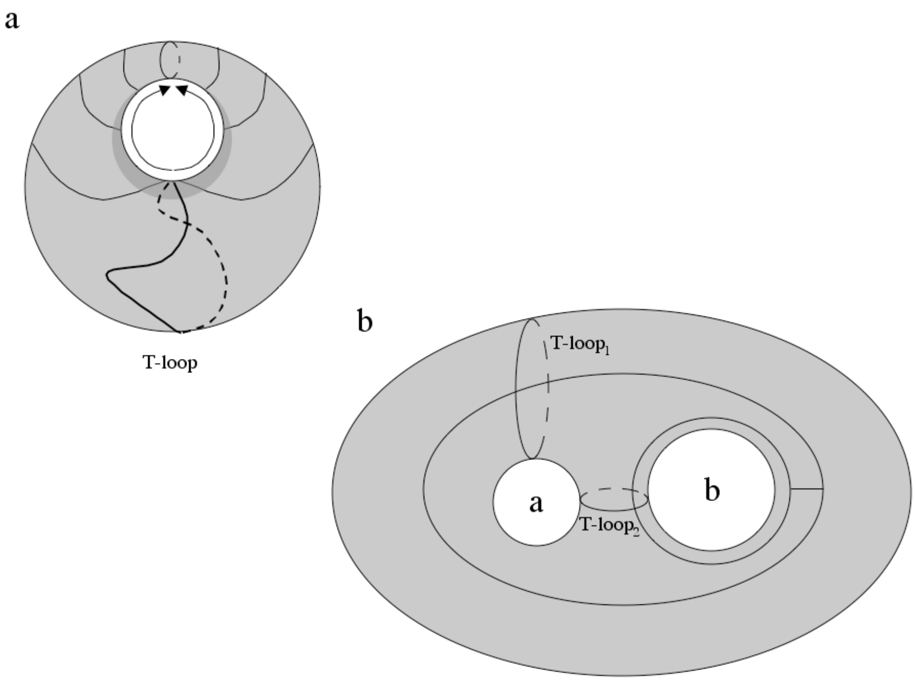


Figure 3-4 Expanding Loops

A) A T-loop (bold line) whose two boundary are sequentially advanced across the surface (light lines), to eventually meet (at arrows). Traceback according to the shortest paths algorithm (along arrows) yields a regular A-loop. B) Two T-loops which both regularize to form A-loops around the same pore *a*. No A-loops are formed around pore *b* in this case, so pores must be processed and capped one at a time.

Identifying two distinct directions in a pore

Having generated and identified a regularized A-loop, the next step is to unambiguously define the two distinct 'directions' from the A-loop out to the two tunnel mouths. We achieve this by building a 'plug' in the A-loop starting from the strip of triangles S'_L forming the regular A-loop. This strip has two edges G and H (Figure 3-5a). We collect two sets of grid points *G* and *H* such that any point in *G* is closer to a vertex in G than any vertex in H and *vice versa* for members of *H*. Additionally, any grid point *g* in *G* has at least one neighboring grid point *h* in *H*, and *vice versa*. The sets *G* and *H* are defined as the opposite sides of the plug. This procedure constructs an oriented, ‘leak proof plug’ across the pore circled by the regular loop S'_L. It is leak proof in the sense that there is no way to pass from one side of this region of the grid to the other staying in the solvent without passing through at least grid point from either side. It is oriented because we know from which edge of S'_L a plug point derived. Thus the plug separates one side of the pore lumen from the other (Figure 3-5a).

In some cases, a regular A-loop will produce a plug that extends out beyond the convex hull. This interferes with the later path-finding procedure but this is easy to correct by generating new loops and new corresponding regular loops using a different random initial triangle. Plugs that do not extend beyond the convex hull will be referred to as valid.

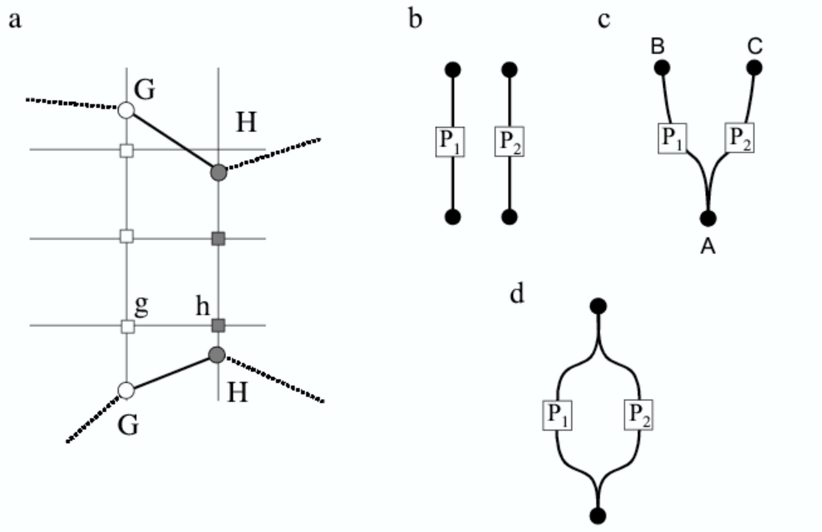


Figure 3-5 Plug Example and Topologically Distinct Paths

A) 2-D representation of a plug. Shown are the surface (dotted and heavy lines) and the volume grid (light lines). (O) Bounding vertices G and H respectively of the regular A-loop. (□) The final plug vertices, with fill indicating sides. B-D) All possible topological cases for a 2-torus: b) two completely separate pores, c) two pores that share one endpoint, d) one pore that bifurcates in the middle.

Ensuring a complete and non-redundant set of A-loops

Since it is possible for an unregularized T-loop to pass through two pores, or for two such T-loops to pass through the same pore, it is possible that the regularized A-loops derived from them would not completely and non-redundantly girdle the N pores. This possibility is illustrated for the simple case of a 2-torus with one narrow pore and one wider pore. If both the loops around the handles ‘find’ a regularized loop around the narrower pore, the wider pore will not have a corresponding regularized loop (Figure 3-4b). The solution is to apply the regularization procedure recursively, ‘masking’ off each pore as it is identified and plugged. A pore is masked off by removing the triangles S'_L of its regularized A-loop and creating two caps of new surface triangles joined the boundary edges *A* and *B*, updating the connectivity information of the surface triangulation as necessary. The remaining surface is now an (N-1)-torus. The procedure of residual strip generation, A-loop regularization, plug generation, and masking off is repeated until all N pores have been processed. We note that in practice this recursive step is the slowest step of our algorithm, as it has a quadratic dependence on the number of handles in the surface and a linear dependence on the number of grid points and surface triangles.

This set of N regularized A-loops with valid plugs contains one loop around each pore in the original surface, and one valid plug in each pore. Additionally, simple checks are done to ensure that all loops are in the original surface, that is they do not contain triangles that were added or removed in the pore masking step.

Generating a path through a pore

Each plug is used in turn as the starting point to generate two 'half' paths out of the pore, one in each direction, terminating at the convex hull. The two 'half' paths start from plug points on opposite sides. This ensures that the complete path really traverses the pore (i.e. can't double back and emerge from the same end it started from).

First the MSSP algorithm is used to assign a Travel Depth and Travel Out distance to each solvent grid point between the convex hull and the molecular surface. The initiating surfaces for this are the convex hull and the molecular surface respectively.

Next the plug point on one side with the maximum Travel Out is identified. Starting from this point a branch-and-bound search algorithm [28] is used on the Travel Out distance, with higher distances taking precedence, leading to a path that passes as close to the center of the tunnel as possible, following the ridge of maximal Travel Out distance. The path is terminated at the first grid point encountered outside the convex hull. In cases where multiple plug grid points have the same maximum value, each path is traced out and the one with the highest minimum value of Travel Out is kept, i.e. the one with the widest choke-point. This procedure is repeated on the other side of the plug. To connect the two half-paths, the two plug grid point maxima (one from each plug side) are connected in a branch-and-bound search, since this again gives a path that maintains the highest minimum Travel Out distance. We note that maximizing some minimum metric has been successfully applied to finding topological paths before [109]. Our approach here is similar to the approach of CAVER

[36; 104]. Our concept of Travel Out distance is the same as the r_{max} function from CAVER, though the methods used to compute them are different. However, in contrast to a branch-and-bound search to maximize the minimum radius of the path, CAVER uses a modified shortest paths search to find a path out, which would seem to maximize the total radius passed through, this differs from our paths.

Building all topological paths through the pores

In cases where there is more than one pore, the set of half-paths generated by the branch and bound algorithm may be combined in different ways to form alternative full paths (Figure 3-5b). For example a Y-shaped or branched tunnel has three entrances, A, B, C and one can define three full paths A-B, A-C and B-C, which share segments (Figure 3-5c). Finding one path per entrance/exit combination is not sufficient to get all topologically distinct (non-looping) paths. A path is defined uniquely only by the entrance, exit, and plug maxima through which it passes. Therefore all plug-to-mouth half-paths are added to a tree, which is then re-processed to get individual full paths. This reprocessing attempts to connect all combinations of points in the tree by all possible non-cycling paths. This gives all the possible topological paths of interest. The potential number of such pathways grows exponentially with the number of pores in the protein surface, however most structures do not have the maximum number of pathways, in fact many have only one pathway per pore, for instance when none of the pores intersect.

Checking that paths traverse pores

An important final step in the path generation approach is to check each potential path to ensure that it passes through an actual topological pore in the protein. This prevents false positives. This is accomplished by using the set of tight A-loops, S'_L. If a path passes through at least one of these loops then it passes through a pore in the protein. Starting with the loop of connected triangle vertices forming one border of a loop strip, *A* (Figure 3-5a), it is triangulated by arbitrarily selecting one point as the common base point, creating triangles using the other points, and then checking whether each path segment intersects with any of these triangles. An odd number of intersections means this path goes through this loop, and therefore through a pore of the protein surface. We note that in theory a path could pass through more than one pore before encountering the convex hull. Currently only one passage is reported, though all passages could be reported with slight additional processing.

In summary, the above procedure results in a complete list of topologically distinct paths. Multiple paths can then be prioritized based on several geometric properties described below.

Test set of protein pores

Having a set of protein structures with realistic and known pores created in them is desirable for two reasons. First, to check the accuracy of the algorithm. Second, to test the algorithm without accessing the limited number of real pore and ion channel structures in the training phase. For this purpose, we created a set of 'punctured' or drilled structures. Starting with larger structures (> 100 residues) from the PDBbind database [33; 34], pores were punctured from one side of the protein to the other by moving a sphere in a biased random walk (using a Von Mises Distribution [110]) from one side of the convex hull to another, removing all atoms overlapped by the sphere at any point. The radius of the sphere at each step was picked randomly from a Gaussian distribution and restricted to be between 2Å and 4Å. The bias for the Von Mises Distribution was set to either 2/3 or 2, which creates relatively straight or a somewhat winding paths, respectively. This procedure was conducted a few times for each protein, then the resulting punctured structures were examined by hand to weed out some pathological cases. 86 relatively straight, and 55 winding, punctured structures were produced. Of this total of 141 known pore cases, a randomly chosen set of 100 were used during the development of the algorithm to identify errors and make improvements. The remaining 41 were reserved until the final version of the algorithm was developed, in order to provide an unbiased estimate of accuracy.

It should be noted that these structures have a reasonable exterior and a reasonable channel through them, but the composition of the interior side chains is severely disrupted by this puncturing process. Characterizing the pores using residue identities or other structural motif methods would not make sense. As the algorithm presented here relies only on gross topological and geometric features and uses atoms, not residues, to create the surface, it is acceptable to use these punctured structures for training and testing. A probe radius of 1.2Å was used when making the molecular surfaces for these structures. This is much lower than the minimum radius of the created pores, to ensure that some additional pores would be present. Also 1.2Å should be small enough for most real ion channel use, so this value was used throughout training and testing, and in all further analyses except where noted. However, this radius could be changed in future applications as nothing in the training or testing procedure is materially dependent upon this parameter.

Quantifying and checking pores

A pore is fully characterized geometrically by the locus of the pore center and the width at each point (the maximum radius sphere that can be placed at this point). Other properties that are of interest include the length, the minimum radius over the entire length, the first minimum radius found from each end, and the maximum radius between the latter two minima. [40]. Additional geometric metrics are also computed as different properties may play roles of varying importance depending on the physiological function of the protein. To get some estimate of the uniformity of the path, the number of local minima is determined. The maximum travel depth is also computed, providing an alternative measure of path length. To estimate how direct a route the path takes, its length is divided by the distance 'as the crow flies' between the ends, which will be 1 for a perfectly straight route and higher than 1 for a route that takes a more circuitous path. This is called the winding metric. Given the path and its radius at each point it is straightforward to identify residues lining the path, or any particular subsection such as a choke point by identifying residues within the pore radius plus some additional distance threshold. The threshold of 4Å was used for all analysis presented here, but this cutoff is under user control. CHUNNEL calculates and outputs each of these metrics for each pore, along with a listing of each tunnel's entrance and exit, and the plug(s) each path passed through, which together uniquely identify the tunnels in a multiple tunnel structure. Finally, CHUNNEL sorts the list of tunnels in order of decreasing minimum radius.

For test cases with known pore paths and radii we designed several measures to check how closely a computed path matched the known path. Since paths are drilled and found by independent algorithms, each with a finite path point resolution, there isn’t necessarily a one-to-one mapping between points on the known and computed paths. In the following measures, for any pair-wise comparison each computed point is mapped to the closest known point.

1) Root mean square deviation between known and computed paths. This was computed using either equal weighting (Prms) or weighting by one over the radius of the known path (Wrms). Wrms weights the narrow sections of the tunnel over the typically wider mouths, as the former are usually more important to get right.

2) Span. We first determine all the points on the known path that are mapped onto by at least one computed path point. The two extremal mapped points are identified, and the span is defined as the fraction of the known path that lies between these two points.

By examining these measures, we can show how closely the paths computed by CHUNNEL are to the known paths in the drilled training and test structures, additionally we can compare the performance of CHUNNEL to the performance of HOLE.

Table 3-1 Representative Timings and Algorithm Statistics

	Sample A	Sample B	Sample C
Number of Atoms	388	2148	4380
Number of Handles	1	7	15
Number of Triangles	5564	37520	63832
Number of Nodes	16943	207703	479422
Number of Paths Found	1	11	156
Count Handles (s)	0.001	0.005	0.008
Travel Out (s)	1.1	58.5	222.7
Get Loops and Plugs (s)	2.1	249.0	1454.5
Find Paths (s)	0.001	0.2	2.0
Total including I/O (min)	0.6	16.1	239.4

Computational requirements

The overall algorithmic complexity of finding tunnels is quadratic in terms of the topological complexity, linear in terms of the number of grid points and quadratic in terms of the number of triangles. Outputting all possible paths is exponential in terms of the topological complexity since there are potentially that many possible paths, however in most cases there are far fewer paths than this. To give an estimate of the practical runtimes involved, we performed some timings using one processor of a dual processor machine (Intel 3.06 GHz chip, 6094 BogoMIPS, 2 gigabytes RAM) running GNU/Linux Fedora Core 4. The results are shown in Table 3-1. The relationship between topological complexity and total processing time can be seen. Though no formal computational space analysis was performed, many hundreds of megabytes of RAM were often in use. Our code currently writes output files compatible with PyMOL, though customization for other programs is possible.

Results

Verification and Accuracy of the Algorithm

The CHUNNEL algorithm was developed on the drilled training set of proteins with known pores. The goal here was to reserve all real structures and the drilled test set of known pores for analysis only after the algorithm was completely developed and we could successfully identify the known pores in the training set. We note that of the 100 training cases, only 10 had a single tunnel. Multiple tunnels commonly arise during drilling when, as atoms overlapping a drill sphere are removed, an additional

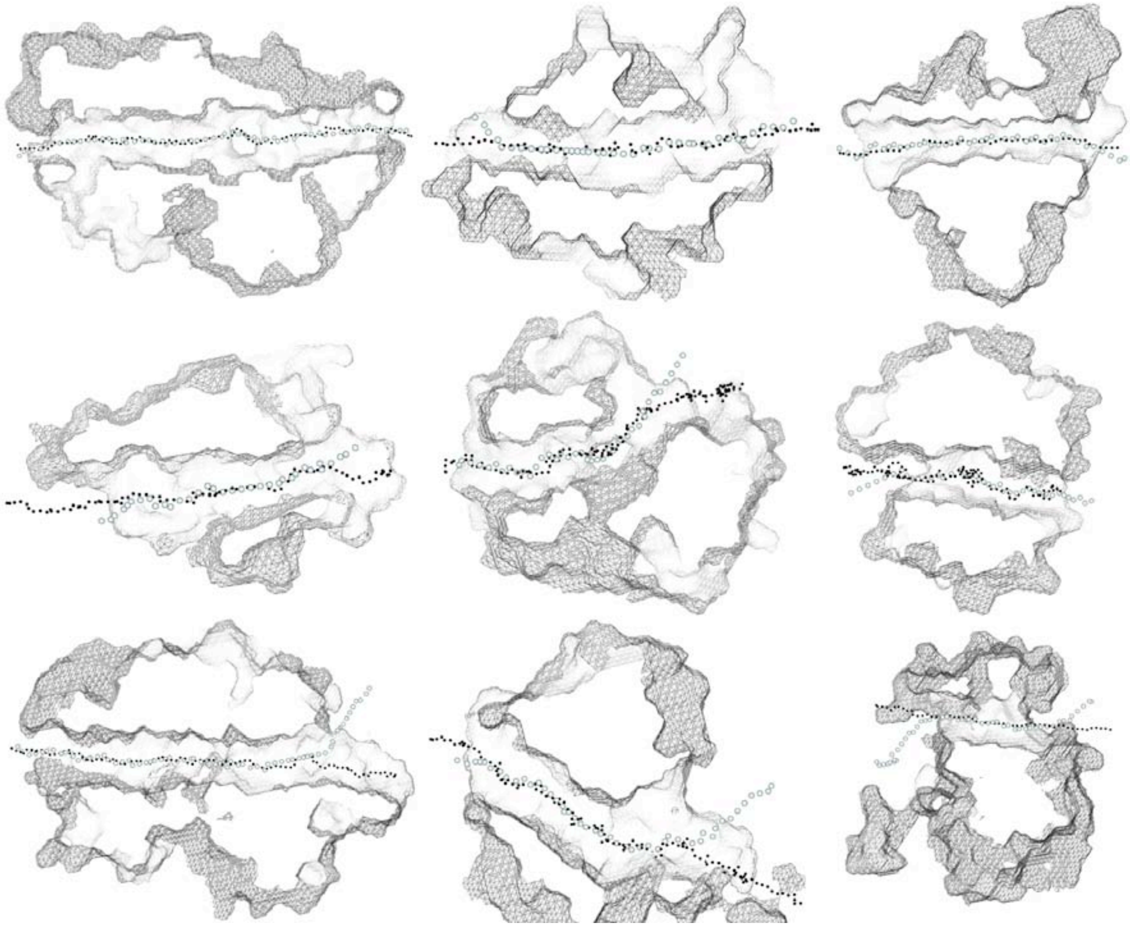


Figure 3-6 Finding Training Set Holes Montage

A montage of 9 (of 100) sample training set cases. The known path is shown in black, the best path according to the lowest Prms is shown in light grey (almost white) spheres, the surface is shown in cutaway. The top 3 cases have Prms < 1Å. The second row of 3 all have Prms of 1.9 Å. The third row shows two examples with Prms of 4.7 Å and then (on the right) a Prms of 6 Å. Figures were produced using customized PyMOL [64].

exit is created. These extra mouths are no different qualitatively from the known tunnel, except their exact path is not known. The CHUNNEL algorithm finds all tunnels but for purposes of testing the algorithm we focus on how accurately the single known path is found. Identifying which of the computed tunnels is the correct one for comparison with the known tunnel is straightforward from either visual inspection, or by its significantly lower Prms.

To interpret the accuracy of CHUNNEL is necessary to know what different values of the measures described previously (Prms, Wrms, Span) actually mean in terms of deviations between computed and known paths. In Figure 3-6, a montage of 9 examples from the training set is shown. In each image, the tunnel is shown via the surface, which has been clipped for visibility, along with both the known and calculated path. The examples were chosen to represent three ranges of Prms values. The first row highlights computed paths that are essentially perfect, they are very close to the known paths from end to end: The Prms values are less than 1Å. In the second row, three examples with Prms values of about 1.9Å are shown. In the leftmost of these, both ends are slightly incorrect, in the other two examples, one end is moderately incorrect. However, these inaccuracies are in the mouths of tunnels, where the lack of a well-defined pore makes it harder to completely and accurately follow the entire length of the path. In the bottom row are examples chosen from the worst performance on the training set. The leftmost two examples have Prms values of 4.7 Å and in both cases the computed path deviates from the known path in one mouth. Again these inaccuracies can be attributed to wide mouths and since the paths are still in the correct mouth they are not a cause for concern.

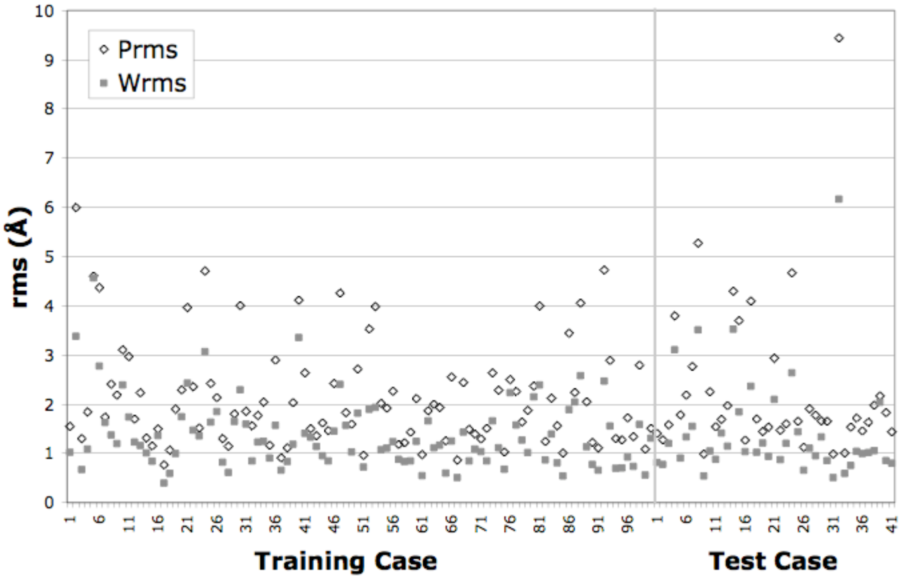


Figure 3-7 Performance on Training/Test Sets

The best Prms (◇) and Wrms (□) found by CHUNNEL for the 100 training cases and 41 test cases in the known pore set.

The rightmost example on the bottom row has a Prms of 6Å and there are inaccuracies in both mouths. The Wrms for all these examples is lower than the Prms, the top row is in the range of 0.4Å to 0.7Å, the middle row's range is 1Å to 1.7Å, and the bottom row's range is 2.5Å to 3.4Å. The range of values for Span on these examples goes from the nearly perfect upper left example with 0.97 to the middle left example with a value of 0.60. The Span values for the bottom row are all greater than this worst case value of 0.60.

With an understanding of the meaning of specific values for the various measures we can examine the performance on the training and test sets of known paths. In Figure 3-7 we show the best Prms and Wrms values for the training and test set. Most Prms values are less than 2Å and most Wrms values are less than 1.5Å, indicating that they have almost the entire path correct. There are however, a number of cases where wide mouths cause the computed path to have a high Prms and Wrms from the known path. In Figure 3-8 the Span values across the training and test sets are shown. Again, most paths are found with high accuracy. Those that are less accurate have inaccuracies in one or two mouths, but the central part of the path is found correctly in all cases, indicated by Span values > 58% in all cases. There are no significant differences in average Prms, Wrms and Span between the training and test sets for our method, indicating that CHUNNEL was not over-trained to perform well only on the training set.

In Figure 3-9 we compare the performance of our method with that of HOLE [40]. Considering first the performance of HOLE in many cases it does poorly, often giving Wrms values of 6-10Å, and even Wrms>10Å, values that indicate partial or complete

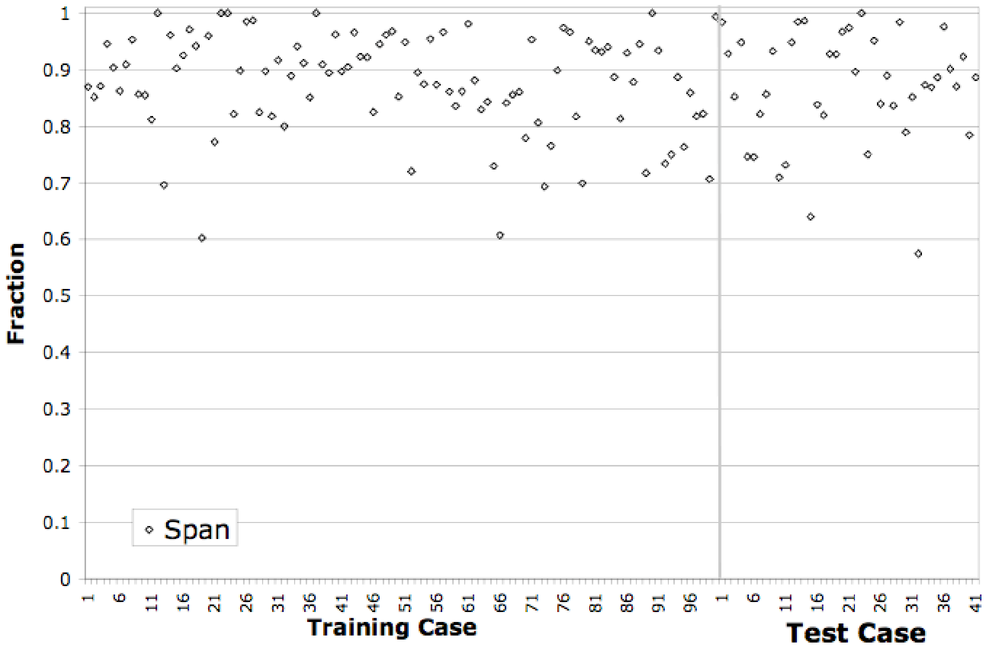


Figure 3-8 Span across Training/Test Sets

The best Span (◇) found by CHUNNEL for the 100 training cases and 41 test cases in the known pore set.

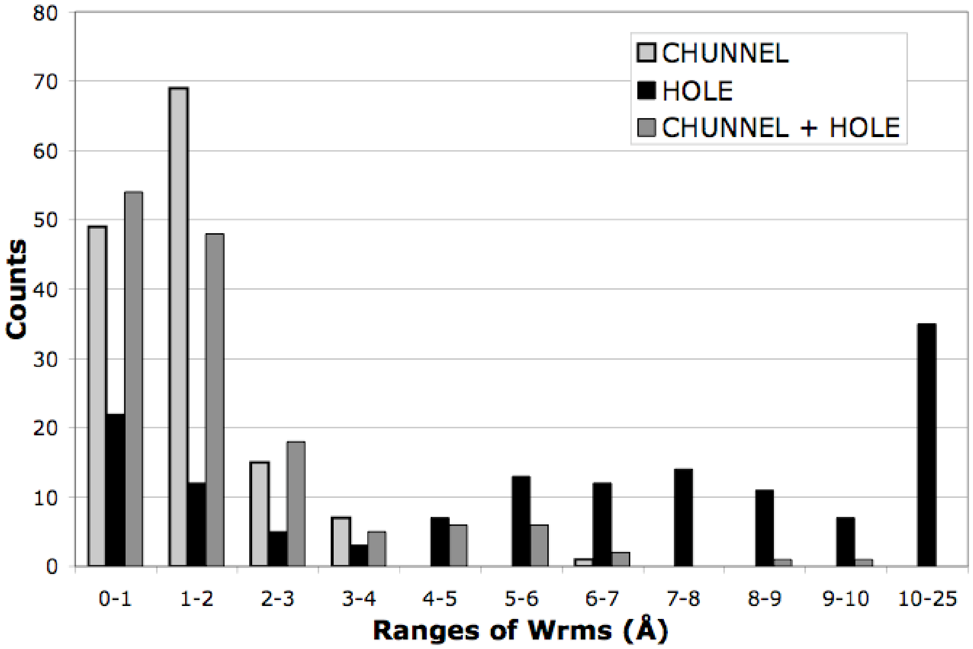


Figure 3-9 Comparison of CHUNNEL and HOLE

A histogram of weighted pore path error, Wrms, between CHUNNEL and HOLE using the combined known-pore training and test sets. (light gray) minimum Prms path from CHUNNEL. (black) HOLE, no hint. (medium gray) minimum Prms path from HOLE given several plug points with maximal Travel Out Distance found using CHUNNEL. Note that above 10Å the results are put into a single bin.

failure to find the path respectively. In contrast, CHUNNEL gives Wrms≤2Å for the majority of cases, indicating the entire path is correct, or there is at most a small error in one of the mouths. In all other cases CHUNNEL gives 2Å<Wrms≤7Å, usually from an error in following the wide mouths. In the process of running CHUNNEL, it identifies the plug points with the maximum travel out depth, i.e. point in the middle of a narrow part of each tunnel. Illustrating a possible way to combine both CHUNNEL and HOLE, these plug positions were used to initialize the latter. With this hint, HOLE produces values of Wrms≤3Å for most of the paths. However, the results are no better than using CHUNNEL for both initiation and generation of paths. In summary, HOLE can perform well when given a hint from the plug generation from CHUNNEL, but in fact getting to this point is really the bulk of the CHUNNEL algorithm. Once a good starting point is found for the tunnel, HOLE and CHUNNEL follow the paths out with similar accuracy.

Application to the Porin membrane protein family

A likely use for our method is to predict the paths of tunnels in membrane proteins. The number of structures of membrane proteins determined through experimental methods, like those of the PDB database in general, is on the rise. The difficulties in obtaining structural data for membrane proteins are being overcome by various methods and membrane proteins will likely become the focus of future structural genomics projects [111]. We used part of a hand-collated database of membrane proteins [38], which on October 1st 2007 had 278 structures representing 132 unique proteins. In this database structures are broken down into groups based on fold and known function, which aids closer analysis. One such sub-group contains the Porins, which provide the molecular basis for membrane permeability. These porins are found in bacteria, and allow promiscuous or specific transport through the outer membrane [112]. CHUNNEL was used to analyze the porin family, as defined by the beta-barreled porin fold [113]. We examined the subset comprised of homotrimers plus structurally related monomers. In each case the complete biological unit was examined. Overall we examined 17 structures [38], including two structures which were analyzed with bound ligands and then again with the ligands removed, for a total of 19 cases [114; 115; 116; 117; 118; 119; 120; 121; 122; 123; 124; 125; 126]. In five of these cases, the physiologically relevant tunnel was blocked either by a structural rearrangement, a peptide or a ligand. Either no paths were found by CHUNNEL or non-physiological paths were found with a very small minimum radius and length, instances where small adventitious pores are created by particular side chain conformations near the surface of the protein. In the other 14 cases the path with the largest minimum radius, ranked first by CHUNNEL, was the physiologically relevant and significant tunnel. Most of the structures are homotrimers, so there are 3 ‘correct’ tunnels, which are all found by CHUNNEL.

It is interesting to note that when viewing the van der Waals representation of the homotrimeric Porins, there is a small gap in the middle of the trimer interface which appears to be a tunnel. However, due to the size of the solvent probe there is no tunnel in molecular surface surfaces and therefore CHUNNEL does not find any paths through this middle region. The first tunnel found in each of the 14 successful cases has a minimum radius of between 1.4 Å and 4.3 Å. The low end of this range is PDB code 2O4V, a porin adapted to phosphate transfer, with the bound phosphate

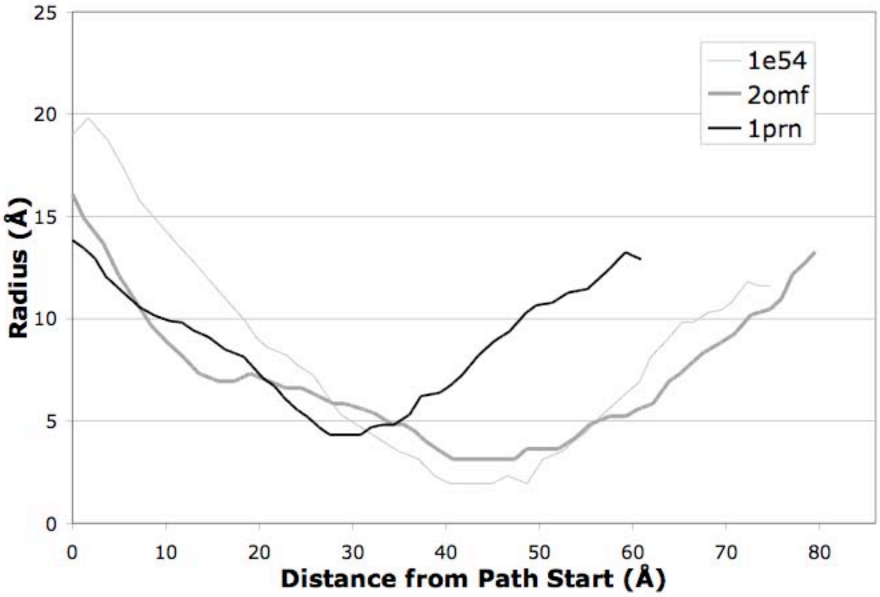


Figure 3-10 Radius Change Along Porin Paths

A graph showing the path radius profile for the first found path from three homo-trimeric porins, PDB entries 1E54, 2OMF, and 1PRN.

removed [126]. This makes sense as phosphate is the smallest specific ligand bound to any of the porins that are not promiscuous transporters. The other bound ligands once removed have paths with larger minimum radii of 1.93 Å for glucose in 2MPR [123], 1.93 Å for malate in 2FGR [118], and 2.4 Å for sucrose in 1A0S [125]. Of course the minimum tunnel radius is obviously not the only factor contributing to specificity in these cases, as many other nonspecific porins have tunnels with similar radii. The two cases of PDB codes 2IWV and 2IWW represent a pH-dependent folding change that blocks the pore [122]. When unblocked the minimum radius is 2.25 Å, when blocked 2 paths formed by side chains on the exterior are found, but no paths are found through the pore.

To further illustrate the ease with which our code allows paths of related proteins to be compared, we compare three of these homotrimeric porins with a small minimum radius (1.9 Å) [117], a medium minimum radius of 3.1 Å [119], and a large minimum radius of 4.3 Å [115]. The first found path for each is shown in Figures 3-10 and 3-11. In Figure 3-10, the radius is graphed against the distance from the beginning of the path, and the minimum point is easy to recognize. In Figure 3-11, the structures with the found paths are shown in increasing size of minimum radius from top to bottom.

As a final example from the porin set we analyzed the makeup of residues lining the entire tunnel and each choke point, using the 14 non-blocked structures. A distance threshold of 4Å from the radius of the pore was used to define lining residues. The enrichment factor for each residue was calculated as the fractional occurrence of that residue lining the path divided by its fractional occurrence over the entire 14 porin

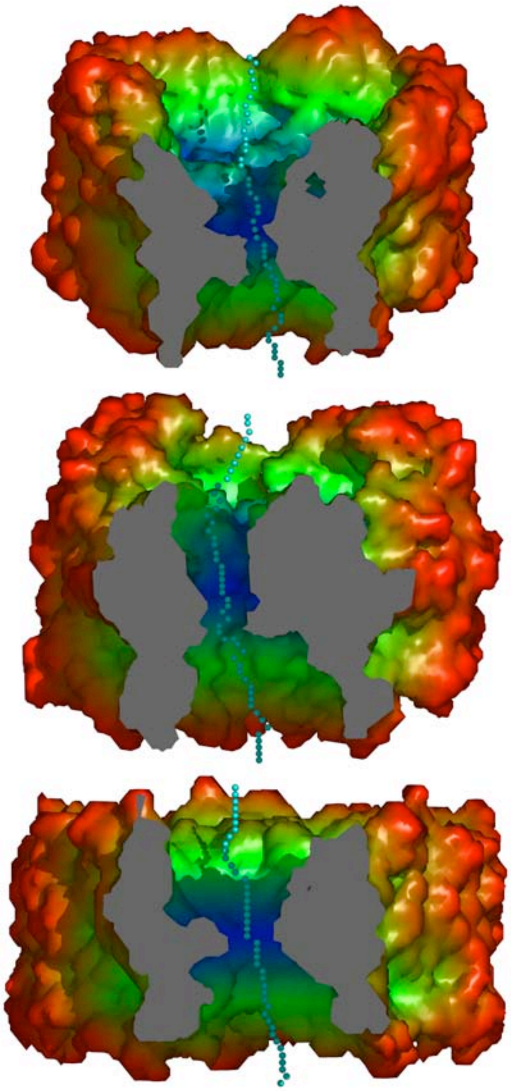


Figure 3-11 Porin Paths

Pore paths (light blue spheres) of three homo-trimeric porins. The molecular surface is color coded according to Travel Depth. The minimum pore radius for each protein, from top to bottom, is 1.9Å (1E54), 3.1Å (2OMF) and 4.3Å (1PRN).

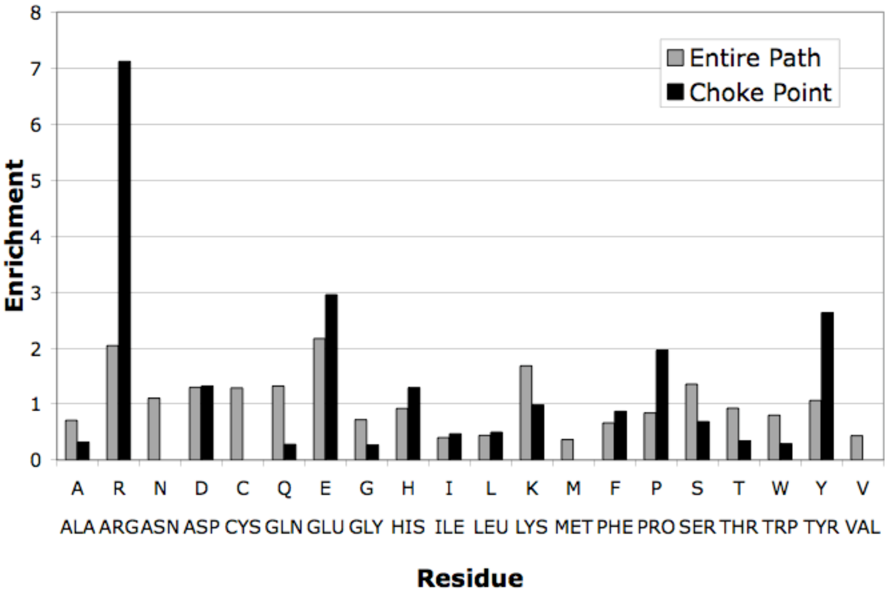


Figure 3-12 Residue Enrichment for Porins

The residue makeup of 14 porins. Shown is the enrichment for either the entire path or the choke point, where enrichment is calculated as the percentage of each residue in the path or choke point divided by the percentage of each residue in the entire 14 porin set.

set. This is shown in Figure 3-12. There is the expected enrichment of polar residues lining the pores, along with a notable enrichment of Arg, Tyr, Glu and Pro residues at choke-points.

Application to Aquaporin

We also examined the integral membrane protein Aquaporin, (which is not a member of the porin family) using CHUNNEL, as this protein presents a challenge for structural analysis of this type due to its complexity, and the small width of the water pores. Each of the 4 units has a tunnel and there is a central tunnel created between them [127]. It is debatable whether or not the central tunnel has physiological importance, so it is important to catalog and compare all the tunnels. We used the aquaporin structure, PDB code 1J4N [128]. In the analysis we found that since the water channels are very small they are missed using the default CHUNNEL probe radius of 1.2Å for surface generation. Hence we used a smaller probe radius of 1.0 Å. However, this creates many small adventitious tunnels where side-chains just barely touch, particularly on the cystoplasmic face of the structure, and a surface with 37 pores results. Many of the 37 pores result from the alternate mouths for all 5 important tunnels on the cytoplasmic side of the protein. Due to the hole-ridden cytoplasmic face of the surface and the different exit/plug combinatorics one can generate hundreds of alternative pore-transiting paths from the half-paths produced by CHUNNEL. The central channel, formed by tetramerization, has a minimum radius of 1.97 Å. The 4 water channel paths found by CHUNNEL have minimum radii of 0.74 Å. Note that this minimum radius is lower than the probe radius used to

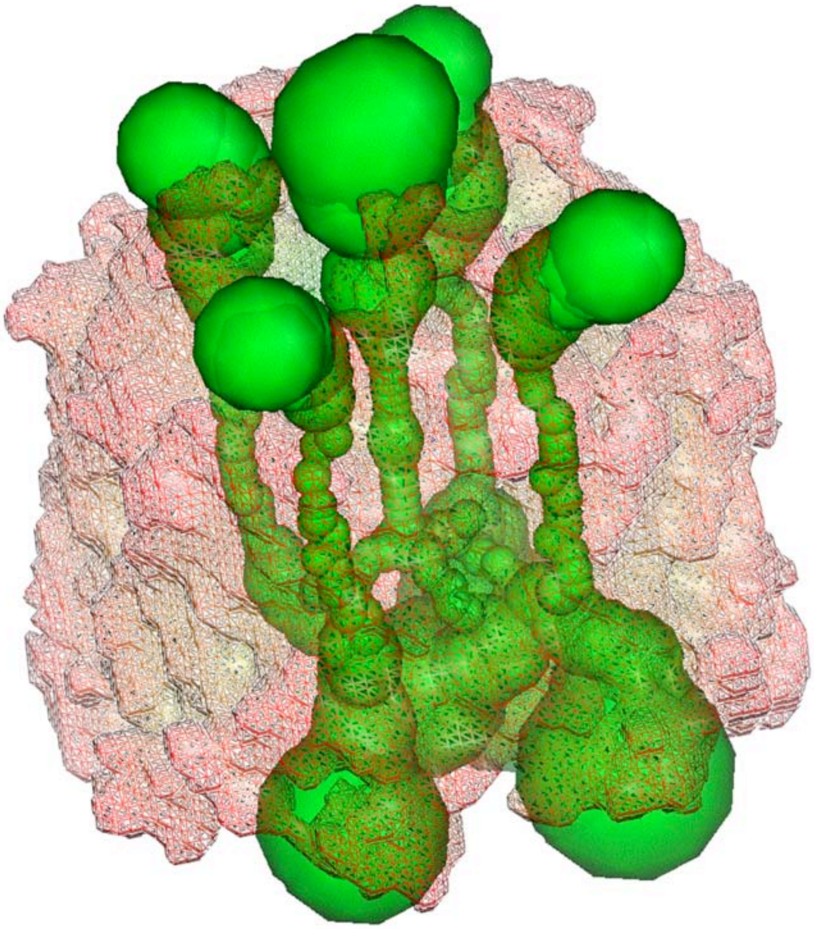


Figure 3-13 The 5 Paths in Aquaporin

The 4 water channels and central channel of an aquaporin, PDB code 1J4N. Each path shown as a series of spheres. The molecular surface is shown in wireframe. The extracellular side of the protein is facing up and towards the viewer. At the bottom, some of the many alternate mouths on the cytoplasmic side of the protein can be seen.

construct the surface, due to the finite resolution of the surface and volume discretization. These five paths are shown in Figure 3-13.

Application to other transmembrane proteins

As a final application for CHUNNEL, we analyzed a larger set of trans-membrane proteins. To do this, we used a set of 192 structures from the OPM database [42]. These trans-membrane structures were gathered from the PDB and their positions within the membrane bilayer were calculated computationally and compared with experiment when possible [129]. We chose the OPM database and methodology since it included not just alpha helices but beta barrels as well, unlike some metrics which were designed for helical trans-membrane proteins only [130]. We accessed this database and used the 192 trans-membrane structures available on January 28, 2008. We removed waters and hetero atoms from the PDB files, which contain complete biological units [42]. Our goal was first to generate all pore paths using CHUNNEL. Second, to identify the subset of CHUNNEL paths which pass exactly once through the membrane bilayer, using the bilayer boundary information of Lomize et al. Third, to identify tunnels that exit within the membrane bilayer. We presume that the bilayer transiting pores would be of greatest physiological importance. The OPM data set also contains many structures for which no physiological path is expected to be found using the CHUNNEL method, including those involved proton channels or proton pumps, as well as GPCRs.

No information on the placement of these structures in the lipid bilayer is used in the CHUNNEL algorithm. This information is used to sort the found paths only after

Table 3-2 Numbers of Tunnels of Various Types in the OPM database

		Entire OPM	Alpha-helical	Beta-barrel	NR25A[d]	NR25B[d]	Entire OPM, Radius > 1.8	Alpha-helical, Radius > 1.8	Beta-barrel, Radius > 1.8
	Total Structures	192	140	52					
Putative Physiological[a]	# of Paths	284	173	111	121	51	82	40	42
	# of Structures	52	26	26	19	14	35	19	16
One Side Exit[b]	# of Paths	1232	1199	33			284	274	10
	# of Structures	73	69	4			30	29	1
Two Side Exits[b]	# of Paths	446	415	31			87	84	3
	# of Structures	51	49	2			19	18	1
Side Branch[c]	# of Paths	108	86	22			63	55	8
	# of Structures	13	12	1			10	9	1

[a] Membrane-transiting

[b] One or both ends of tunnel exit within bilayer.

[c] Branch off a membrane-transiting path that exits within the bilayer.

[d] Nonredundant set with maximum 25% sequence similarity of proteins with alpha (NR25A) or beta (NR25B) motif.

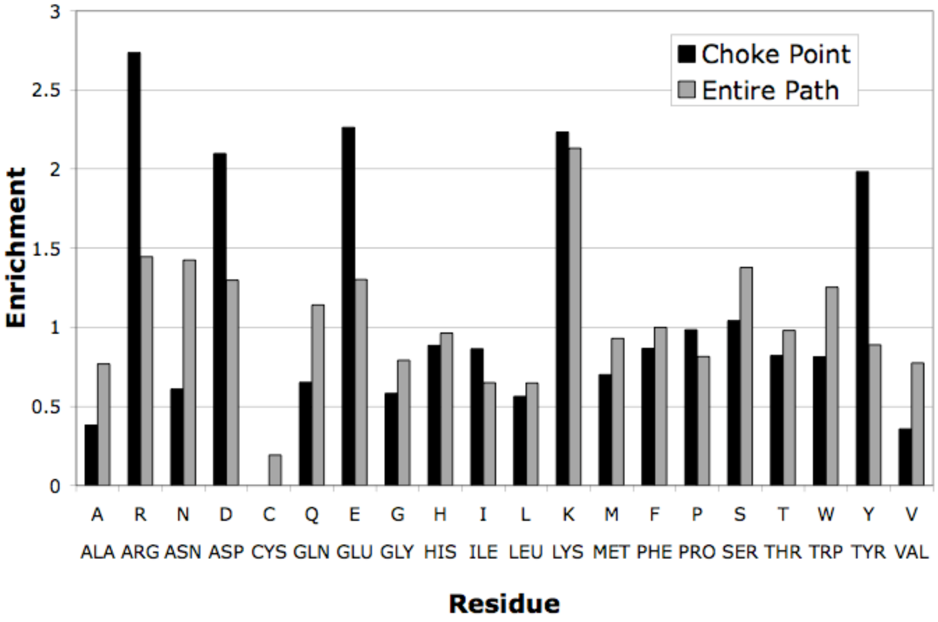


Figure 3-14 Residue Enrichment of Transmembrane Paths in OPM

The residue makeup of the putative physiological paths in the trans-membrane part of the OPM database[42]. Shown is the enrichment for either the entire path or the choke point, where enrichment is calculated as the percentage of each residue in the path or choke point divided by the percentage of each residue in the entire trans-membrane set.

processing is complete. Note that while the OPM methodology is limited to flat symmetric membranes, our analysis could be repeated for more general definitions of membrane barriers, for instance by the use of elastic theory to define the lipid/water interface [131].

After processing the OPM database with CHUNNEL, 284 membrane-transiting, putative physiological paths were found in 52 unique structures, indicating that multiple tunnels are the rule rather than the exception (Table 3-2). Accounting for degeneracy of paths due to multimeric proteins, there are 175 unique membrane-transiting paths in 52 unique monomers/proteins. In 28 of these structures, there is a single unique path per monomer. The mean length of these putative physiological paths is 126±51Å, much greater than the width of the membrane bilayer (usually 25-30Å). There are two reasons for this. First, the paths must pass through not just the lipid barrier, but the whole protein, to reach the convex hull of the protein. Second the paths are usually not straight, the data set having a mean winding metric of 1.68±0.5. The path width minima over the set have a mean of 1.35±1.8Å which is within the expected physiological range considering that 1.2Å probes were used to construct these surfaces. Enrichments for residues found near the choke point and near the entire path were calculated relative to the residue composition of the entire OPM trans-membrane database. These enrichments are shown in Figure 3-14. There is an overall enrichment of the charged amino acids, particularly Arg, Glu and to a lesser extent, Lys, and an enrichment of the polar aromatic residue Tyr. For a finer analysis, the structures were split into either alpha-helical or beta-barreled classes, and pruned to a maximum of 25% mutual pairwise sequence identity using

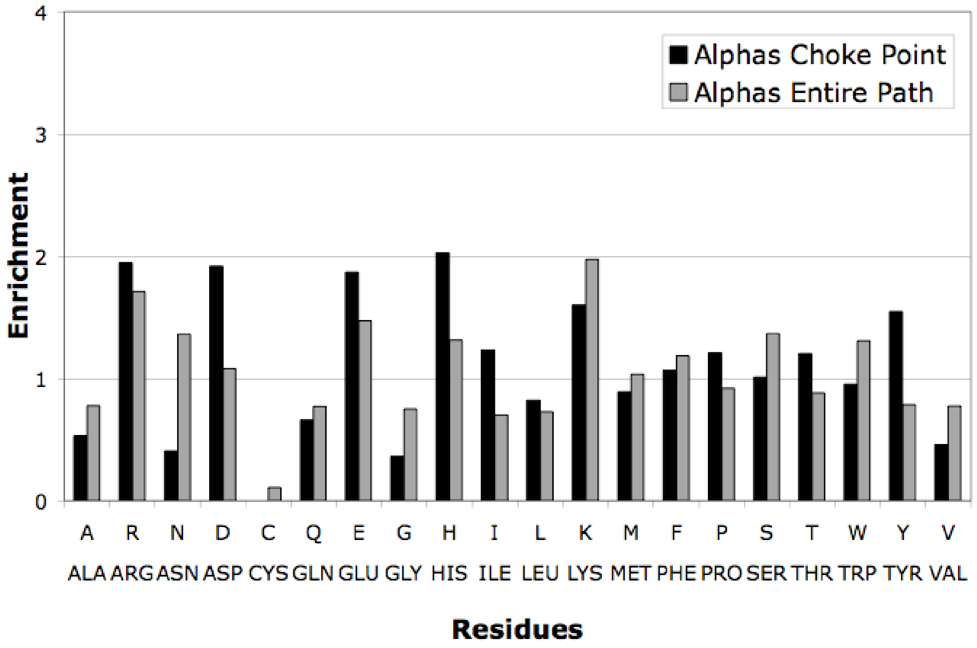


Figure 3-15 Residue Enrichment in Alpha Helical OPM

Residue enrichment for pores and choke points of alpha helical motif proteins of the OPM database[42], pruned to 25% sequence similarity using PISCES [132].

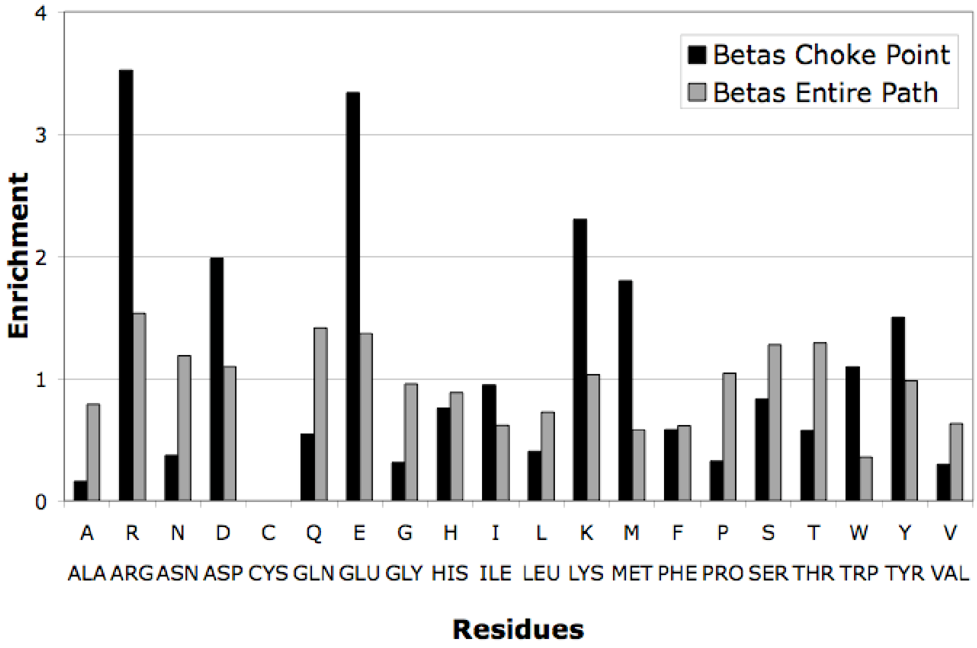


Figure 3-16 Residue Enrichment in Beta Barrel OPM

Residue enrichment for pores and choke points of beta barrel motif proteins of the OPM database[42], pruned to 25% sequence similarity using PISCES[132].

PISCES[132]. The results are shown in Figures 3-15 and 3-16. Removal of sequence homologous duplicates insures that these graphs reflect real pore amino acid preferences, not just sequence conservation. The same 4-5 residues show enrichment, but interestingly, the degree of enrichment is much greater in the beta-barrel class than the alpha-helical class.

Additionally, there is a surprisingly large number of paths, 4879, that do not pass through both membrane barriers once. This shows the power and importance of the membrane barrier data of Lomize et al. [42] in analyzing membrane protein pores. From this set of paths, we analyzed three interesting subsets: 1) Those that start on one side of the membrane bilayer and emerge within the bilayer. 2) Those that start and end within the bilayer. 3) The branches of membrane-transiting putative physiological tunnels that terminate within the bilayer. Other classes of paths, such as those that lie entirely within a region on one side of the membrane, were not analyzed. Since we are also interested in paths that could potentially contain water, we separately identified tunnels whose minimum radius is greater than 1.8Å, the commonly accepted upper limit on the size of a water. The numbers of such tunnels and what kind of structures they are found in (alpha-helical or beta-barrel) are summarized in Table 3-2. When examining the data graphically we notice that when side exits lie very close to the membrane surface they may exit the protein outside the membrane but reach the convex hull at a point inside the membrane, in which case they are classified as exiting inside the bilayer. The reverse situation also occasionally occurs. This introduces some ambiguity into the classification of intra-membrane side exits, and some degree of uncertainty in the numbers tabulated in

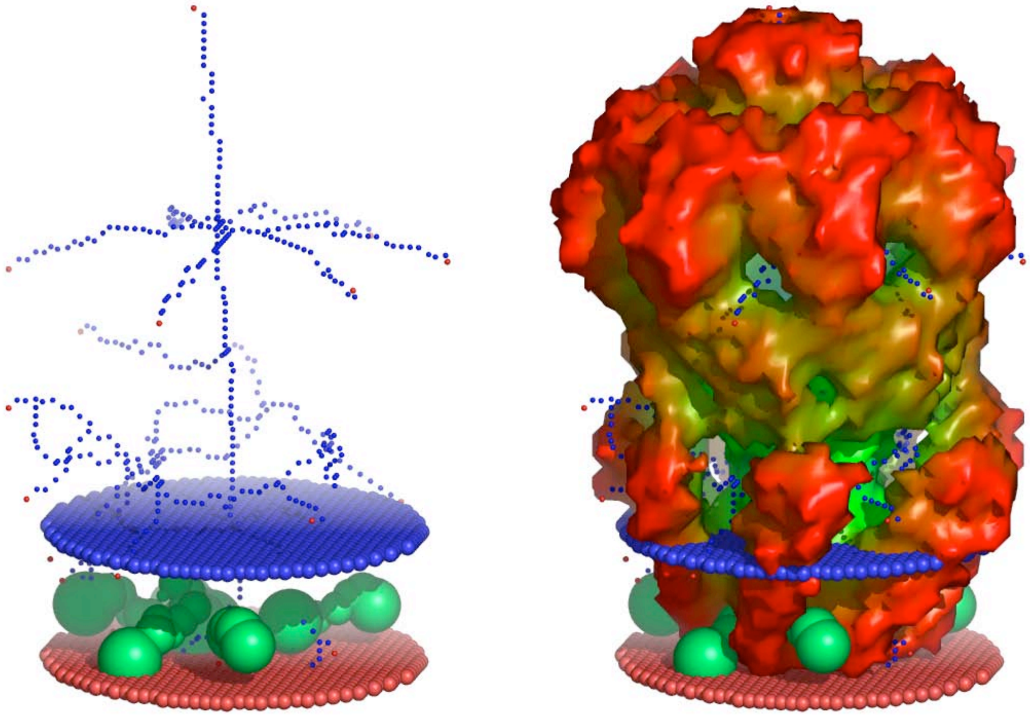


Figure 3-17 Paths in a Complicated Membrane Protein

The mechanosensitive channel of small conductance (MscS), PDB code 2OAU, shown with the membrane barriers in red and blue disks. The complete tree of paths is shown in blue spheres, the end points in red spheres. Some of the branched tunnels shown in green. At left, no protein is shown for clarity, at right, the Travel Depth surface is shown.

Table 3-2. In specific proteins of interest, the ambiguity is easily resolved using graphical analysis.

The overall message from the data in Table 3-2 is that complicated tunnel topologies, defined as multiple membrane transiting paths, paths with intra-membrane exits, and branches with intra-membrane exits are not rare. For example side tunnels and branched tunnels, although not ubiquitous, are quite common. Of particular interest is that they are much more common in alpha-helical domains than in beta-barrel domains. As a good example of a complicated tunnel structure, we show the Mechano-sensitive channel of small conductance (MscS)[93] in Figure 3-17, showing the complete tree structure of the tunnels and some of the intra-membrane branched tunnels as well.

Preferences for residues lining intra-membrane exiting and side branching tunnels were also examined. The most interesting case appears to be the paths and choke points of the tunnels that branch off of physiological tunnels that exit inside the membrane. Strikingly, a strong, five-fold enrichment for Trp is shown (Figure 3-18). Even using the residue composition of the protein regions just within the membrane barriers, the enrichment of Trp in these branch paths is still over 2-fold, and near choke points is still almost 3.5-fold. It has been noted that in many membrane protein structures tryptophan is often found near the polar head group, and head-group/acyl chain interfaces regions of bilayers [130; 133]. Together these data imply that side branches preferentially exit in this polar/apolar transition region of the membrane. Significant amounts of water within the membrane are also observed in

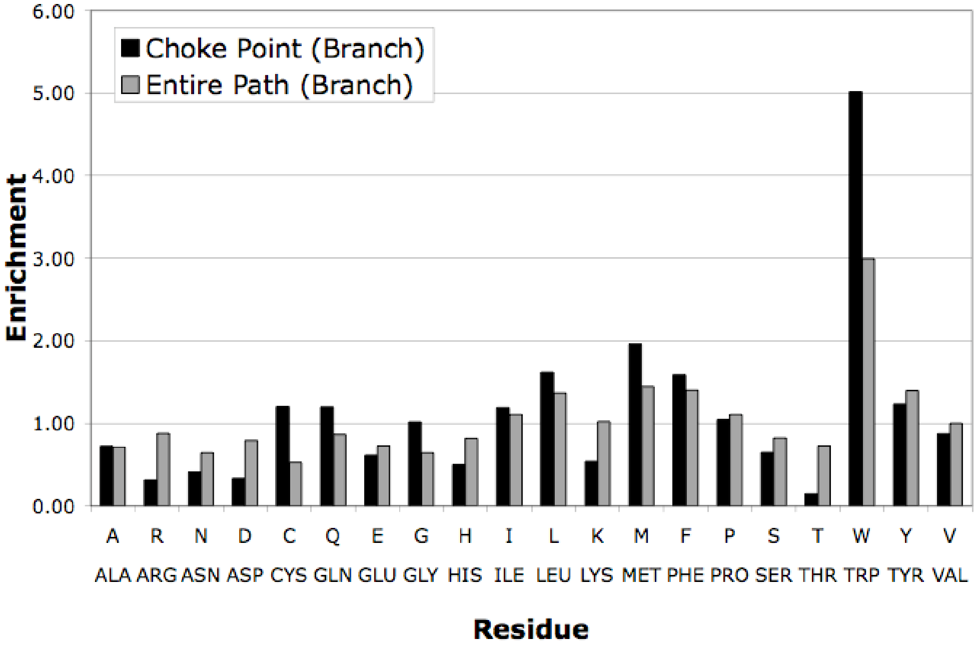


Figure 3-18 Residue Enrichment For Branched Side Tunnels

Enrichment of residues near the branches of putative physiological tunnels that exit into the membrane bilayer.

the head-group/acyl chain interface region [134]. It is thus likely that these side branch exits are accessible to some water.

Discussion and Future Work

We have presented here the implementation and testing of a new algorithm, CHUNNEL, to automatically find and characterizes pores in proteins. The main contribution of CHUNNEL is the ability to identify and catalog all the tunnels through a given surface, which neither HOLE [40], CAVER [36; 104] nor other work [95; 103] could accomplish automatically. Though CHUNNEL is markedly slower than HOLE due to complicated geometrical and topological computations, the results are worth it for various applications. Moreover, complete automation is necessary for analyzing more than a handful of structures, and for the membrane protein databases. These databases are growing at a steady pace due in part to structural genomics projects [101]. Our analysis of the trans-membrane portion of the OPM database [42] is the first large-scale, automated analysis of channels that pass through the membrane barrier.

A second contribution of CHUNNEL is the ability to easily analyze structure and residue composition of the pores. Some studies on smaller classes of trans-membrane proteins have been conducted, for instance on aquaporins and related proteins [102; 135]. These studies highlighted the arginine/aromatic selectivity filter. Our results on a much larger OPM data set confirm this pattern of residue enrichment: Both arginine and tyrosine are highly enriched at choke points in the larger set, shown in Figure 3-14. Arginine is also highly enriched in the choke points of the unrelated outer membrane porin family, shown in Figure 3-12. We also partitioned membrane proteins of the OPM trans-membrane database into the two alpha helix and beta barrel motif subsets. The analysis of the residue enrichment shows significant differences between these two motifs (Figures 3-13 and 3-14 respectively). The beta barrel motif has a less uniform distribution, showing stronger preferences for Arg, Glu, Lys and Met than the alpha helix motif. In other words the alpha helix subset seems to favor a wider variety of amino acids in choke points than the beta barrel subset. The reasons for this marked difference in amino acid preferences with structural motif are unknown at this time. Possible factors include evolutionary and environment constraints, since the beta-barrel trans-membrane proteins are only found (so far) in the outer membrane of bacteria. Since there is still a small number of non-homologous proteins with trans-membrane paths in either class (14 beta barrels 19 alpha helices), these difference may be due in part to normal statistical fluctuations. As the database expands in future, this question can be easily revisited, due to the automated nature of CHUNNEL.

Another striking finding is the sheer number of tunnels and tunnel branches in membrane proteins, both membrane transiting, and non-transiting. While additional channels in the extra-membrane portions of membrane proteins have been noted, to our knowledge, the analysis here is the first to draw attention to and analyze the multitude of intra-membrane exiting channels. In part this is a consequence of HOLE's intrinsic design for finding linear tunnels: These side or branched tunnels would not be found with previous methods. Regarding the physiological importance of these additional tunnels and branches, this can be systematically evaluated based on the tunnel type:

1. Both exits in the aqueous phase, and transiting the membrane once. This is the 'classical' tunnel of putative physiological function, subject of numerous analyses. Presumably at least one such channel must exist in the 'open' state for the protein to function. The exception is for proton or electron transport across the membrane, which can occur through 'wires' or chains of donors and acceptors. Here, due to the small size of the permeant entity, no actual tunnel may exist.

2. Both exits in the aqueous phase, not transiting the membrane, i.e. confined to the extra-membrane region on one side of the membrane. This is not likely to have any functional importance.

3. A branch off a membrane transiting tunnel, with the exit in the aqueous phase. If the selectivity filter, or highest energy barrier controlling the flux is in the common part of the tunnel, before the branch, then the extra mouth is likely to have a small effect, otherwise an extra branch would create a 'short-circuit' The extra entrance may however increase the probability of the substrate finding the channel, which at low concentrations could increase the rate. Multiple entrances may also play a role if multi-substrate interactions, such as ion-ion interactions, are important in conduction [136].

4. A branch off a membrane transiting tunnel, with the exit in the membrane interior. For an ionic or polar substrate, presumably the solvation penalty for exiting in the membrane, compared to the aqueous phase, is so high that conductance is minimal. Effectively the apolar part of the membrane 'plugs'

such leaks. This may explain why such tunnels are relatively common, as there is little evolutionary pressure for a protein to evolve a structure that is completely leak-proof alone. However, there is a propensity for such tunnels to exit in the transitional region between acyl tails and head groups, where there is a significant amount of water. Thus a sufficient degree of hydration to allow leakage currents cannot be ruled out. The existence of such water filled side tunnels also has implications for the interpretation of membrane structure probing experiments such as cys-labelling and spin labelling mapping of water accessible and inaccessible regions [137; 138; 139]. Regions may be accessible to the probes, but inside the membrane. In a solubilized form of KcsA, waters can be seen to exit and enter through these side tunnels under molecular dynamics simulations [140]. Finally, since any such tunnels with a minimum radius of 1.8Å or greater are presumably filled with water, this may play a role in the energetics and dynamics of substrate permeation. First, by providing an additional reservoir of water in the interior of the channel that could help hydrate ions. Due to the long range nature of the electrostatic interaction, this water need not actually be touching the ion, or even in the main channel to be energetically significant. The energetic effects need not be limited to the permeant ion. Voltage sensing of channels require that charge elements be moved in the membrane, and the energy of this would be affected by nearby water [141]. Second, in allowing water to flow in or out in response to substrate movement. In many cases the main channel is narrow enough that substrates and waters must move in file, requiring concerted movements and limiting conductance [136]. Additional water passages ahead or behind the substrate could facilitate motion.

5. One or both exits inside the membrane region. These could play a role in allowing the interaction between membrane soluble carriers and channel permeant species. Examples of the former include the apolar quinones that interact with the bc(1) complex [142].

Clearly more analysis of such epiphytic channels needs to be done in specific cases to investigate their functional importance.

Future work in this area includes calculation of additional metrics and pore properties, with the aim of possibly distinguishing non-physiological tunnels from ion channels and pores from the structure in the absence of relevant experimental data. While the influence of some geometric properties on various properties of tunnels, particularly ion channels [143], has been conducted, there is still much work to be done in this area, in part because the databases are still developing, in part from lack of fully automated, reliable pore finding. A single metric used here, the largest minimum radius, correctly identified the physiological tunnels in the porin set. However, a complete set of geometric features, as well as other physical features will no doubt be necessary if we are to identify physiological tunnels of other classes of protein. In this regard, we point out that CHUNNEL, like HOLE and CAVER, does not provide much assistance in finding the paths of proton channels. Proton channels function in a different manner than ion channels in that the proton is not necessarily transferred through an open tunnel [144]. Thus reducing the probe radius is of no help.

CHUNNEL uses a set probe radius, chosen in advance to be smaller than the smallest permeant specie of the channel(s) being analyzed. An interesting alternative is to use methods taken from the alpha-shape filter idea [11]. This would allow one to find a probe radius where the first topological tunnel emerges. However finding additional tunnels would require complete recomputation of the CHUNNEL procedure as the alpha-shape filter changes, effectively decreasing the probe radius and changing the entire surface. This is currently beyond practical computational capabilities. For this reason as well as reliance on previous code for surface generation we currently implement a fixed, user controlled probe radius parameter, rather than an automated method.

Further work in both the algorithm and the implementation remains to be done. The quadratic dependence of the algorithmic complexity on the number of holes is acceptable, but should be improved as the program can take hours to run if the surface has many holes. The worst combination is an extremely large complicated structure and a very small probe radius, these prove to be impractical to run on desktop workstations. Improvement here may also make the automated probe radius option discussed above feasible.

The methods developed here to find a topologically complete and geometrically distinct set of loops could prove useful in other applications. The ability to remove the handles from an n-torus and turn it into a topological sphere is a powerful method in many fields of computational geometry, for instance to use spherical harmonic methods [145]. Since our removals are done to cap tunnels roughly at their narrowest point, the caps are geometrically well placed. For other applications it may be better to remove handles by cutting the handles at their narrowest point, or possibly a mix of cutting handles and capping tunnels. For instance, removing each handle by doing the smallest amount of changes would result in the closest thing to a topological sphere for a given protein surface, which would be useful for algorithms that only work on topological spheres, for instance mapping complicated topological spheres to geometric spheres [146].

In summary, we introduce a method, CHUNNEL, that automatically finds starting points and paths for all possible topological tunnels through a macromolecular surface. This improves upon the mostly, but not completely automated methods of HOLE [40] and CAVER [36; 104]. Starting points found using our method can be used by these other methods as well, in fact a hybrid approach may be advantageous for some applications. We show that we can find all known paths in a constructed data set of drilled tunnels and show examples and some overall analysis from a set of trans-membrane proteins [42], including automatic identification of residues found near the tunnels or in the choke points.

Chapter 4

This chapter is based on previously published work [57]

Summary

Organisms evolved at high temperatures must maintain their proteins' structures in the face of increased thermal disorder. This challenge results in differences in residue utilization and overall structure. Focusing on thermostable/mesostable pairs of homologous structures, we have examined these differences using novel geometric measures: specifically Burial Depth (distance from the molecular surface to each atom) and Travel Depth (distance from the convex hull to the molecular surface that avoids the protein interior). These along with common metrics like packing and Wadell Sphericity are used to gain insight into the constraints experienced by thermophiles.

Mean Travel Depth of hyperthermostable proteins is significantly less than that of their mesostable counterparts, indicating smaller, less numerous and less deep pockets. The mean Burial Depth of hyperthermostable proteins is significantly higher than that of mesostable proteins indicating that they bury more atoms further from the surface. The Burial Depth can also be tracked on the individual residue level, adding a finer level of detail to the standard exposed surface area analysis.

Hyperthermostable proteins for the first time are shown to be more spherical than their mesostable homologues, regardless of when and how they adapted to extreme temperature. Additionally, residue specific Burial Depth examinations reveal that charged residues stay unburied, most other residues are slightly more buried and Alanine is more significantly buried in hyperthermostable proteins.

Introduction

It seems likely that hyperthermophilic archaea occupy positions near the root of the phylogenetic tree of life. However, there is still some debate as to whether life originated in hyperthermophilic conditions [147; 148; 149]. Nevertheless, life has adapted to many niche temperatures. Of these, the high temperature niche is the most puzzling to explain from a thermodynamic perspective, due to the increased thermal disorder that favours denatured or unfolded states. In addition to insights into fundamentals of protein stability, the discovery of thermostable variants of many enzymes has led to many practical applications [47]. Understanding how these variants achieve thermostability could lead to new ways to design proteins for greater thermostability, among other applications.

Inspired by recent work examining protein structures from a range of environmental temperatures from mesophiles to hyperthermophiles [51] we wanted to examine the overall shape and structural features of these proteins using recent advances in protein shape analysis. Additionally, we wanted to perform a more detailed analysis of structure at the residue level. With the ongoing determination of structural data on many homologues from various thermophiles and mesophiles, such overall structural differences can now be examined with increasing statistical resolution.

Many structural features that could lead to increased thermostability have been examined previously, and diverse factors have been found to differ between thermostable proteins and mesostable proteins with varying degrees of significance. The picture is further complicated when considering the extremity of the temperature (thermophiles vs. hyperthermophiles) [51; 52] and the evolutionary background of the organism- ancient (original?) thermophiles vs. acquired themophilicity [150]. Structural factors that have been studied include increased hydrogen bonding in thermostable proteins [48; 53; 55; 150; 151; 152], an increase in the frequency of ion pairs and electrostatic contributions in thermostable proteins [48; 51; 52; 53; 55; 152; 153; 154; 155; 156] and an increase in the amount of certain apolar contacts [157]. Differences in unfolding have been studied by various methods [150; 158] including differences in rotamer states [51; 156]. Also the differences in solvent exposed surface area have been examined [48; 49; 50; 51; 52; 53; 54; 55]. Van der Waals interactions, the amount of packing, and the number and size of cavities have been examined but lead to conflicting conclusions [53; 150; 152; 156; 159]. This is only a brief review of the structural features examined on multiple sets of protein pairs; Many other features have been examined, but only on single pairs of protein structures or by sequence based analysis.

Surprisingly, there has been no definitive study of overall shape and geometry differences, such as sphericity, arising from environmental temperature differences. This work addresses this by examining the overall geometric structure of thermostable proteins. In addition, we perform a finer resolution analysis of surface exposure. Previous work [48; 49; 50; 52] examined the solvent exposed surface area changes per atom type, residue, or residue group. Some studies examined just the nonpolar exposed surface area changes [51] or the differing counts of residues that were exposed or buried according to a cutoff of solvent accessible area [53]. Only one previous study accounted for the overall shape changes by correcting the surface areas examined [55]. In this study, we not only identify residues that are buried, but we examine how deeply they are buried, using the distance to nearest point on the protein surface, or 'Burial Depth.' We also use Travel Depth [35] to examine the overall structure of the pockets and clefts of the proteins. Combined, these two depth measures provide complementary measures of how spherical the proteins are, if they are closer to ideal spheres or if they have more indentations, dimples and clefts.

We use a collated data set [51] which contains homologous structures from both mesophiles and several kinds of thermophiles. Both moderate thermophiles (45° C to 80° C) and hyperthermophiles (above 80° C) are examined. Additionally, we break the class of hyperthermophiles into two subsets, the Ancient hyperthermophiles that have been hyperthermophiles for their entire evolutionary history [160], and Recent hyperthermophiles like *Thermotoga maritima* that only recently became hyperthermophiles [149; 161]. This follows the lead of previous work where a similar split in the class of hyperthermophiles was used [150].

Materials and Methods

Data Collection

We use the recent data set of Greaves and Warwicker [51] that contains pairs consisting of a thermostable protein and a homologous mesostable protein. We use primarily the '67' set of data since these structures have pairs with chain lengths differing by 30 residues or less. Only such similarly sized structures are appropriate for most shape analysis. We do, however, use the larger '291' set for some additional analyses that do not depend on overall protein size. We examine both the moderate thermophiles and hyperthermophiles, and in addition we examine two subsets of the hyperthermophiles, the Ancient and Recent. The only organism in the dataset known to have recently adapted to extreme high temperatures is *Thermotoga Maritima* [149; 150; 161]. To be counted as Recent in our analysis, the protein must be from *T. maritima*, additionally it must not be from an Archaeal lateral gene transfer [161]. Each protein from *T. maritima* in the 67 set [51] has closest relatives from other bacteria and is therefore presumably not from lateral gene transfer from an already hyperthermophilic archaea. There are 12 pairs in the Recent-mesophile set, and 18 pairs in the Ancient-mesophile set, for a total of 30 pairs in the Combined hyperthermophile-mesophile set. There are 37 pairs in the moderate thermophile-mesophile set.

Files of single domains as specified [51] were downloaded from the Protein Data Bank [16]. In the case of multiple NMR structures, the structure closest to the average structure was used as representative of the set. All waters were removed from crystal structures. Hydrogens are assigned a radius of zero in our van der Waals radii set [3], effectively ignoring them for all analyses. A 1.8Å probe radius was used for all packing, surface construction, Travel Depth, Burial Depth analyses, as this is consistent with a water sphere. For the Burial Depth and Travel Depth analyses cavities are removed after surfaces were constructed.

Packing

Protein packing can be calculated by using the Voronoi construction [9; 162; 163; 164; 165; 166]. A Voronoi cell is defined as the volume that is closer to the given atom (or point in the more general sense) than any other atom [26; 45]. Defining the packing as the percentage of the volume of the Voronoi Cell filled by the van der Waals volume of the atom, packing is well-defined for completely buried or interior atoms. However, surface atoms have infinite Voronoi cells which must be restricted if a meaningful measure of their packing is to be computed. Methods of 'capping' the Voronoi cells of surface atoms include using the molecular surface as the bounding volume [167] or using crystallographic waters [9; 163; 166; 168]. Availability of sufficient crystallographic waters to cap depends on the resolution of the structure and how it was refined. Moreover, waters are entirely absent from NMR determined structures. For these reasons we decided to analyse surface and buried atoms separately. For the former we used the solvent accessible surface of each atom to generate the Voronoi cell capping. The solvent accessible surface is generated from van der Waals radius plus probe radius, so it lies a constant distance from the atom regardless of protein shape. However this method of capping is somewhat arbitrary, as are other methods used to determine the packing of surface atoms. For this reason, in detailed comparisons of packing, we believe it is more reliable to use just the interior atoms, which are bounded on all sides by identically defined and generated surfaces.

Travel Depth

Previously, Travel Depth was established as a useful measure of depth of the molecular surface for examining pockets and ligand binding sites as well as the overall depth of the surface [35]. Travel Depth is defined as the minimum distance from any surface point to the convex hull avoiding the protein interior, and is calculated using the Multiple Source Shortest Paths (MSSP) algorithm [23]. The original implementation has been improved for speed, flexibility and additional features [43].

Burial Depth

Atom burial depth has been used several times previously to analyze protein structure, although somewhat varying definitions exist in the literature, depending upon the exact implementation and desired use [19; 20; 21 17; 22; 169], see the review of Pintar et al [18]. However no measure of burial depth has previously been applied to analyzing differences in thermostable and mesostable structures. A closely related method uses burial and counting nearby hydrophobic residues to discriminate native folds from decoys [170]. Another related concept is that of centrality or closeness of a graph connecting nearby atoms, used in various applications [171 77; 172 173; 174]. Atom Burial Depth is defined here as the distance of the atom to the nearest point on the molecular surface. It is most efficiently calculated by starting from the molecular surface and labelling sequentially deeper points into the protein interior, using the same MSSP algorithm as for Travel Depth [35; 43].

Interatomic Distances, Wadell Sphericity, Convex Hull Volume

We also examined several other shape metrics to see if they could discriminate between thermostable and mesostable proteins. The first was the mean interatomic distance. In principle this could be sensitive to how spherical and well-packed the folded structure is, with the advantage that it is exceptionally simple to compute, requiring only atomic coordinates, not the protein surface. The second metric was the convex hull volume, which would be larger if the protein structure is more spread out, or less ‘compact’ at the larger, molecular scale. We note that compact is a pervasive yet ambiguous term in the literature on thermostable proteins. It has been used to refer to the efficiency of packing as measured by a Voronoi or similar analysis. It has also been used to refer to the number of contacts of a certain type, for instance hydrogen bonds, van der Waals contacts, etc. Finally it could refer to the extent of a protein, how ‘splayed out’ it is. Here we use compactness only in this sense, as defined by the convex hull volume. All other usages can be replaced by better terms.

The third metric, used previously to evaluate roundness of rocks and crystals, is Wadell Sphericity [175], a dimensionless ratio of volume and surface area designed to have an upper bound of 1 (perfectly spherical), and decreasing to 0 the further from perfectly spherical the shape is. The formula for this ratio is given by Equation 4-1 and it was calculated exactly from our triangulated molecular surfaces.

$$\psi = \frac{\pi^{1/3}(6V)^{2/3}}{A} \tag{4-1}$$

Statistical Tests

Statistical significance is evaluated by permutation testing, by randomly switching (or not switching) the labels on each thermostable and mesostable pair and recomputing the difference in means of each metric across each category, and evaluating if the original difference is extremal to each permuted difference [176]. We used two individual one-tailed tests, meaning the permuted means found are checked to see if they are less than or greater than the original. In all cases the lower p-value is reported. Each statistical test reported was done using 1,000,000 permutations in the case of overall tests and 10,000 permutations in the case of residue-specific tests. Importantly, in residue specific tests, the overall difference in means was used as a correction factor to the difference in means when analyzing which residues become more buried or unburied. A standard threshold of 0.05 was used as a cutoff for significance.

Results

Packing, Mean Distance, Convex Hull Volume and Wadell Sphericity

The packing analysis was performed on each atom in each structure, the results were separately accumulated over either all buried or all surface atoms in each

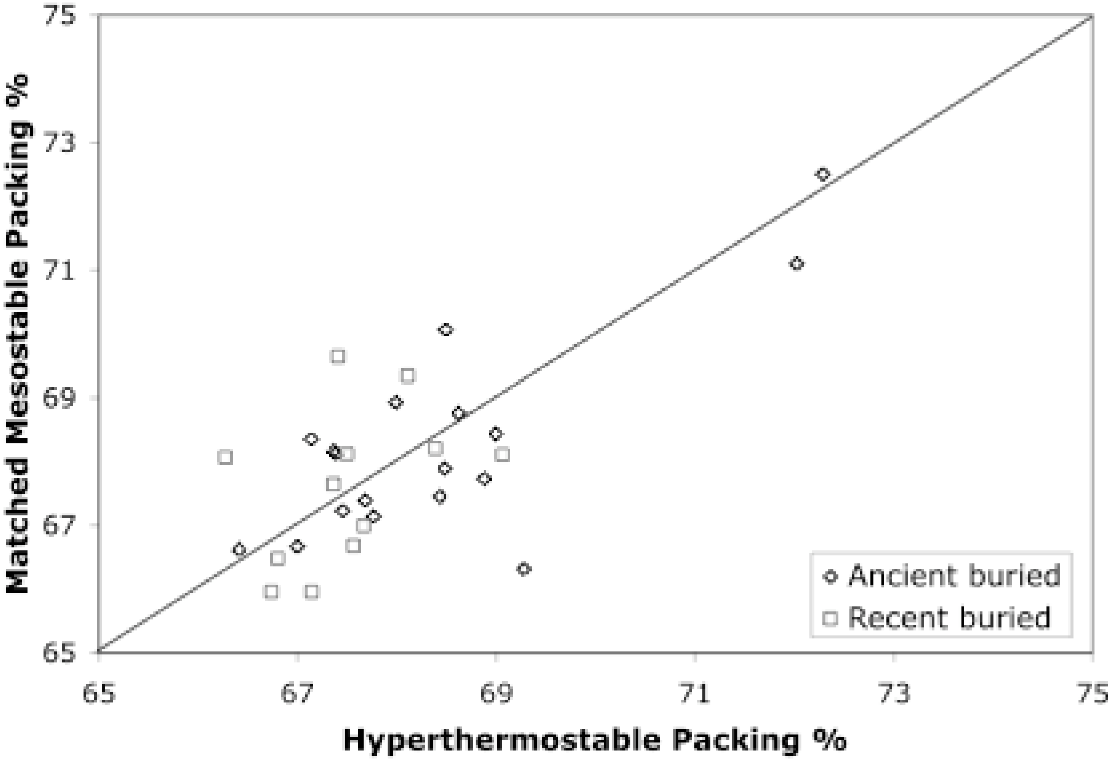


Figure 4-1 Packing in Hyperthermostable Proteins

Packing percentage in the buried category comparison between Recent or Ancient hyperthermostable vs. matched mesostable proteins for completely buried atoms.

protein and the mean of these numbers and the combined mean over all atoms was used to evaluate significance. Packing of recent and ancient hyperthermostable proteins is compared with their mesostable counterparts in Figures 4-1a and 4-1b for the buried and surface atoms, respectively. The results are summarized in Table 4-1 for the various thermostable categories compared to their mesostable counterparts. No significant differences were found in packing of interior atoms for any thermostable set. Surface atoms, however, are significantly more tightly packed in both the ancient and combined hyperthermostable proteins.

The mean of the interatomic distance was computed across all heavy atoms, the results between the various thermostable proteins and their matched mesostable proteins are summarized in Table 4-1 and the results for the hyperthermostable categories are shown in Figure 4-2. No significant differences were found for any thermostable set. The convex hull volume, a metric for the overall extent of the protein also showed no significant difference in any category, again shown in Table 4-1.

The Wadell Sphericity [175] of each protein surface was computed by calculating the area and volume of the triangulated surfaces, and the dimensionless ratio given by Eq. 1 computed. The results are shown in Figure 4-3 and summarized in Table 4-1. Hyperthermostable proteins are significantly more spherical than their mesostable counterparts. No difference is found for moderate thermostable proteins.

Wadell Sphericity is size independent so the analysis was also conducted on the ‘291’ set [51]. The difference in mean Wadell Sphericity between the 144 pairs of

Table 4-1 Summary of Differences in Mean Values of Geometric Measures

Geometric Measure[a]	Moderate Thermostable	Recent Hyper-thermostable	Ancient Hyper-thermostable	All Hyper-thermostable
Number of Pairs	37	12	18	30
Packing of Buried Atoms (%)	0.44 (0.06)	-0.10 (0.39)	0.16 (0.28)	0.06 (0.39)
Packing of Surface Atoms (%)	-0.01 (0.24)	0.28 (0.16)	**0.59** (<0.01)	**0.47** (<0.01)
Mean Interatomic Distance (Å)	0.08 (0.38)	-0.40 (0.27)	-0.50 (0.12)	-0.40 (0.08)
Convex Hull Volume (Å^3)	5 (0.5)	-1739 (0.3)	-3321 (0.08)	-2688 (0.06)
Volume (Å^3)	-430	1310	289	697

	(0.12)	(0.07)	(0.33)	(0.08)
Surface Area (Å^2)	-143	-227	**-533**	**-411**
	(0.23)	(0.27)	(<0.01)	(0.02)
Wadelll Sphericity	0.002	**0.027**	**0.039**	**0.034**
	(0.42)	(0.05)	(<0.01)	(<0.01)
Mean Travel	-0.06	**-0.40**	**-0.50**	**-0.46**
Depth (Å)	(0.34)	(0.05)	(<0.01)	(<0.01)
Mean Travel	-0.001	**-0.044**	**-0.047**	**-0.065**
Depth/H$^{1/3}$ (Å) [b]				
	(0.41)	(0.03)	(<0.01)	(<0.01)
Mean Burial	0.12	**0.13**	**0.12**	**0.13**
Depth (Å)	(0.10)	(0.01)	(<0.01)	(<0.01)

[a] Data is shown as the mean of the thermostable category minus the mean of the mesostable category. Statistical p-values for the lower of the two one-tailed tests follow in parentheses. Values with a p-value below the significance threshold of 0.05 are shown in bold.

[b] Mean Travel Depth divided by the cube root of the number of heavy atoms in the molecule, H.

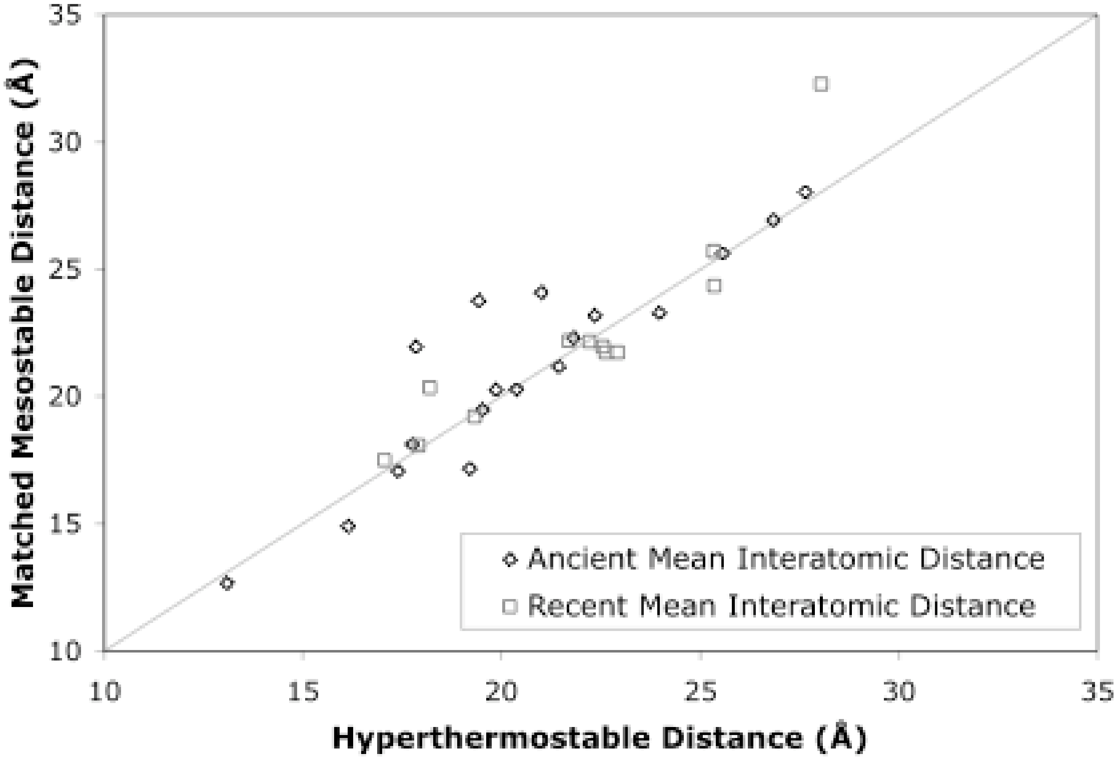


Figure 4-2 Interatomic Distances in Hyperthermostable Proteins

The mean interatomic distance compared across Recent or Ancient hyperthermostable vs. matched mesostable proteins.

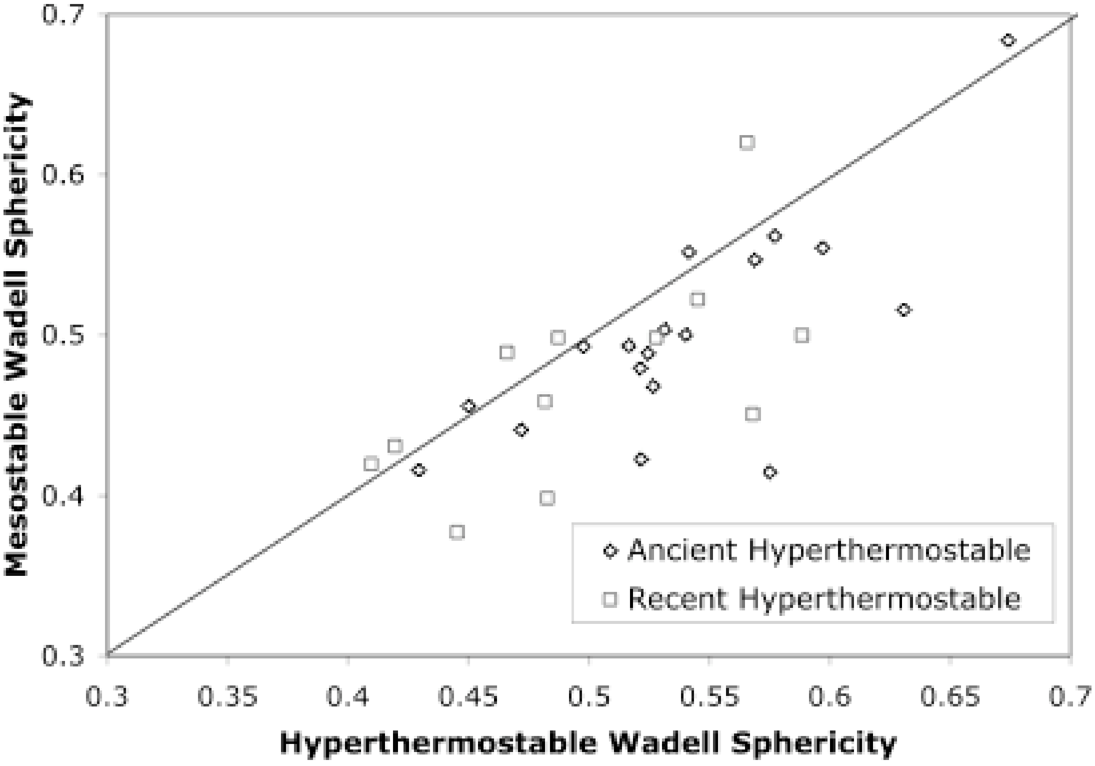


Figure 4-3 Wadell Sphericity in Hyperthermostable Proteins

Wadell Sphericity compared between Recent or Ancient hyperthermostable vs. matched mesostable proteins.

hyperthermostable-mesostable structures was 0.035, with $p<<0.001$. These results indicate that in the larger set of hyperthermostable proteins, they too have become more spherical.

Travel Depth and Burial Depth

The Travel Depth analysis was conducted on each protein structure. The mean Travel Depth, T_{av}, was computed by averaging over all surface points as described previously [35]. This mean was used to compare the various thermostable categories with their mesostable homologues, the results are shown in Figure 4-4 and summarized in Table 4-1. Hyperthermostable proteins have significantly smaller values of T_{av}, indicating a less convoluted surface. Since the maximum depth of a pocket is limited by the linear dimensions of the molecule, variation in size of proteins potentially complicates the interpretation of average travel depth. The volume of the protein is closely proportional to the number of heavy atoms, H, so $H^{1/3}$ provides a convenient measure of the average linear dimension of the molecule. Indeed Figure 4-5 shows that on average, travel depth increases linearly with the linear extent of the protein, for mesostable, thermostable, and hyperthermostable proteins. The scaled average travel depth, $T_{av}/H^{1/3}$ thus provides a good measure of the *relative* roughness of the molecule, in a fractal sense, as shown before on a larger class of small molecule binding proteins [35]. Differences in scaled average travel depth,$T_{av}/H^{1/3}$, are summarized in Table 4-1, and are also significantly smaller for both hyperthermostable categories.

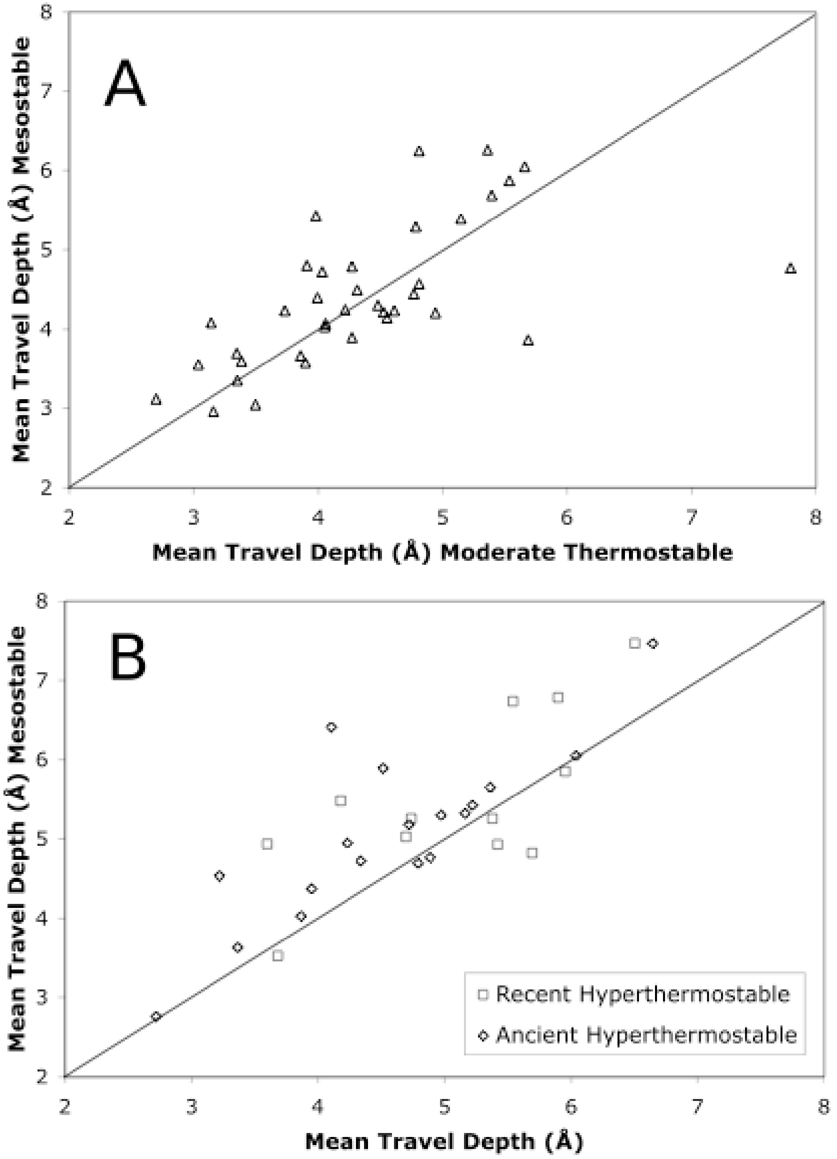


Figure 4-4 Travel Depth in Thermostable Proteins

Mean Travel Depth compared between (a) Moderate thermostable (b) Hyperthermostable and the respective matched mesostable proteins.

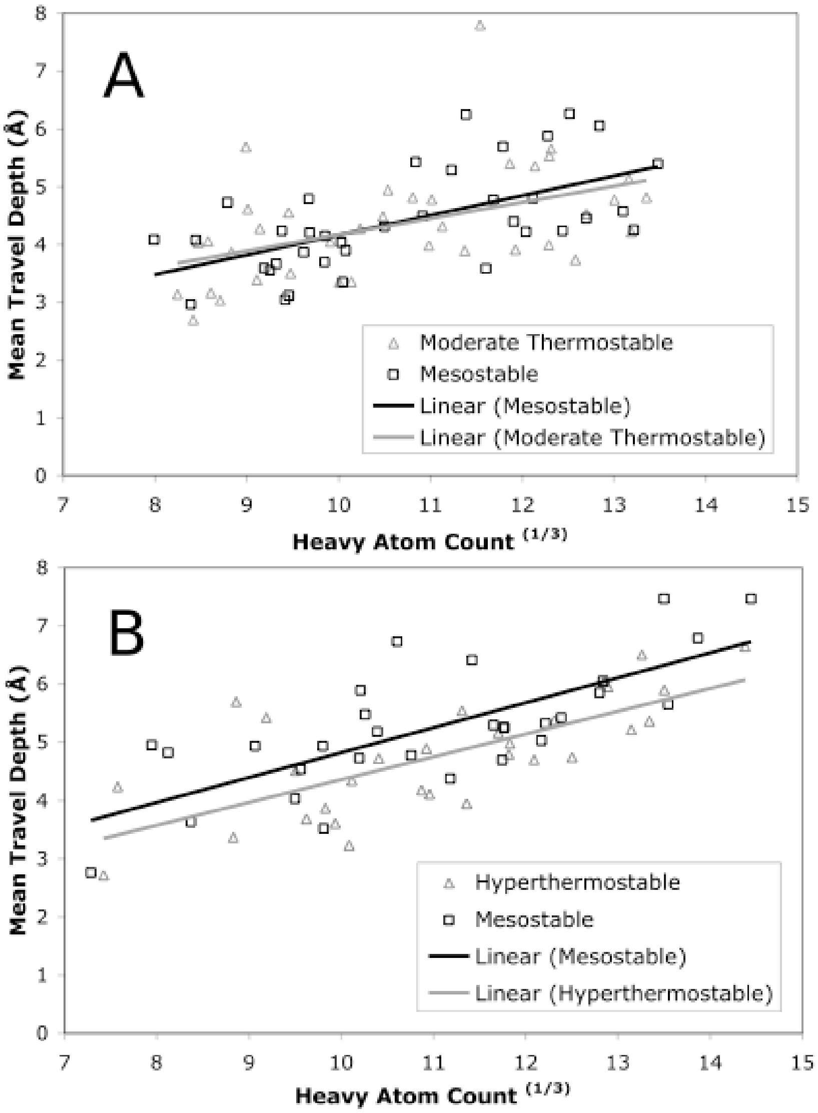


Figure 4-5 Size-Scaled Travel Depth in Thermostable Proteins

Mean Travel Depth is plotted vs. the cube root of the number of heavy atoms. (a) Moderate thermostable and the matched mesostable proteins. (b) Hyperthermostable and the matched mesostable proteins. Trendlines for each set are shown on the figure.

The Mean Burial Depth of all atoms in each protein structure was computed and the results are summarized in Table 4-1. Atoms are significantly more deeply buried in hyperthermostable proteins, by slightly more than a tenth of an Ångstrom. This is a very small difference, but it is an average over a large number of atoms, and it is statistically significant. The consistently deeper burial of atoms in hyperthermostable proteins is made more evident in complete histograms of Burial Depth accumulated over all atoms types, shown in Figure 4-6. The histogram for both hyperthermostable classes is consistently shifted to the right, indicating greater burial depth. This rightward shift is even clearer in the cumulative difference histogram. If just the C_β atom of each residue (C_a for glycine) is used for the burial depth analysis, very similar histograms, means, and p-values result. So for this kind of analysis the single atom burial is a good proxy for that of the entire residue.

Both smaller mean travel depths, and greater mean burial depths in hyperthermostable proteins indicate a more spherical shape in hyperthermostable proteins as compared to mesostable proteins. This is illustrated graphically for an ancient hyperthermostable-mesostable matched pair of protein structures of Phosphoserine Phosphatase in Figure 4-7. The molecular surface on the left is colored by Travel Depth, while the right hand bond representation is colored by Burial Depth. The hyperthermostable protein (upper panels) clearly has more red colored (shallow) surface, and more red colored (deeply buried) atoms.

Since the size scale Travel Depth metric, $T_{av}/H^{1/3}$, largely removes the effect of protein size, one can compare mesostable and thermostable proteins that differ substantially in size, for which there are more proteins to compare. This metric was

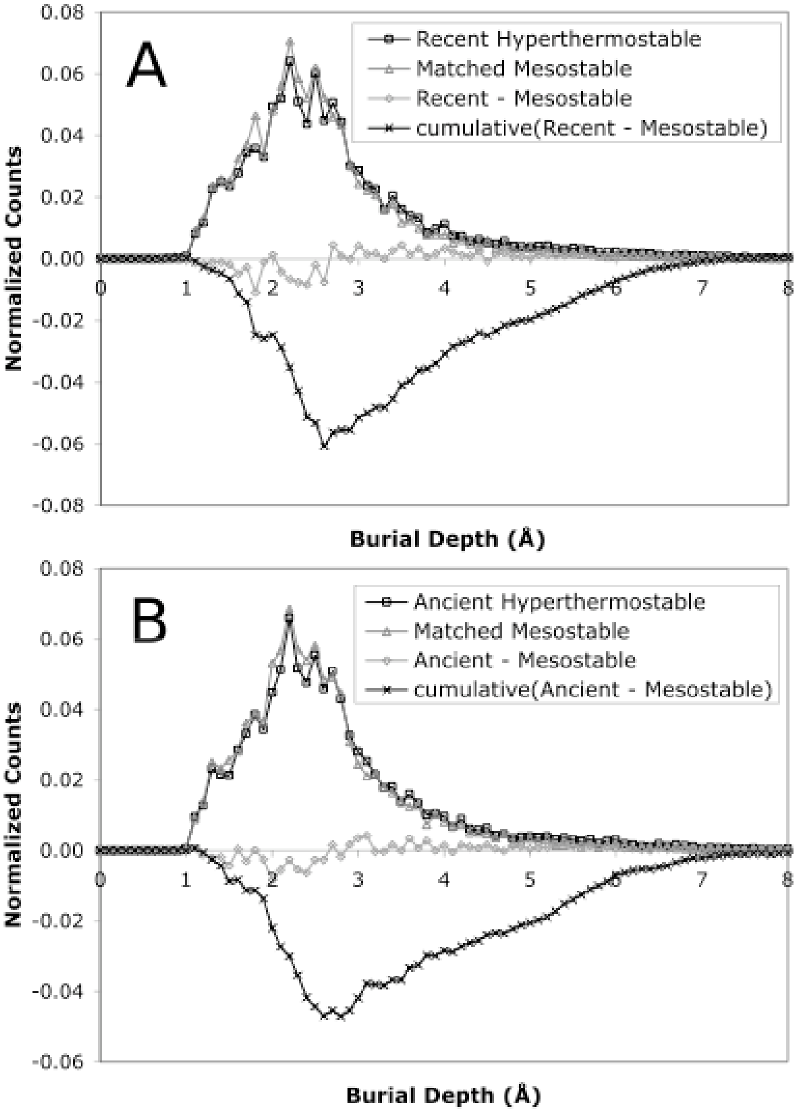


Figure 4-6 Burial Depth in Hyperthermostable Proteins

The figures show normalized counts of all atoms vs. burial depth for the thermostable and mesostable proteins, the difference in frequency distributions (Thermostable-Mesostable), and the cumulative frequency difference distribution. (a) Ancient hyperthermostable vs. matched mesostable proteins. (b) Recent hyperthermostable vs. matched mesostable proteins.

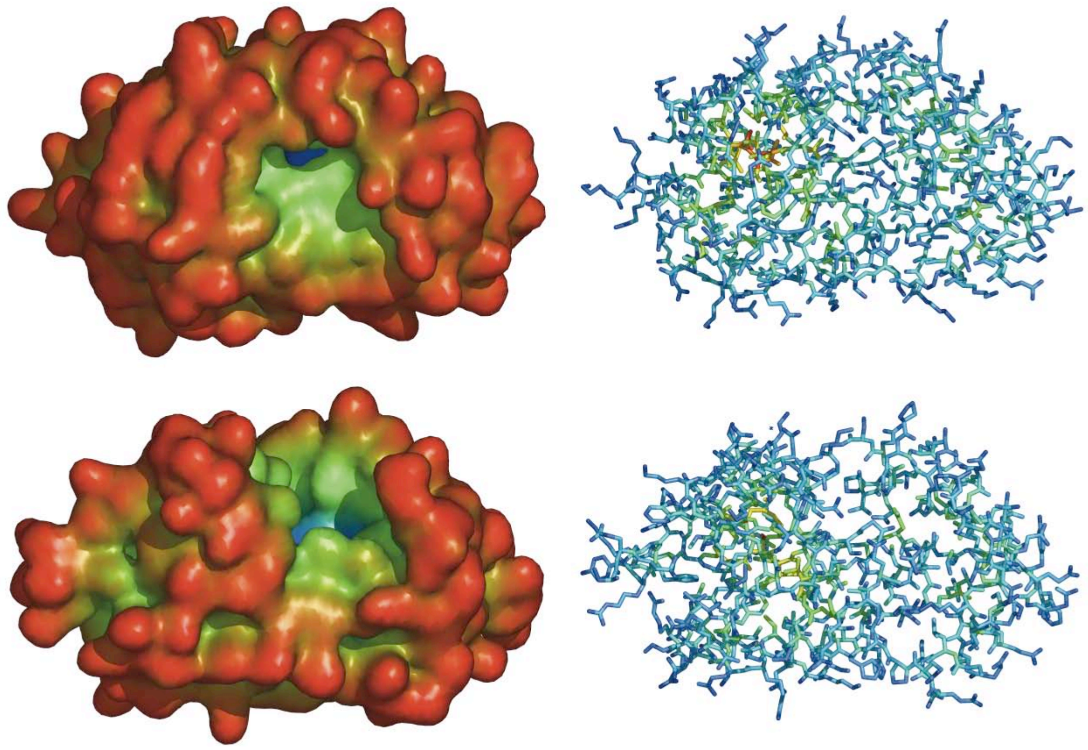


Figure 4-7 Example Structure Pair Colored by Travel Depth and Burial Depth

An example matched pair of protein structures: hyperthermostable Phosphoserine Phosphatase (top, PDB code 1L7M[177]) and mesostable Phosphoserine Phosphatase (bottom, PDB code 1NNL[178]). At left, the molecular surface is colored by increasing Travel Depth from red to green to blue. At right the wireframe representation is colored by increasing Burial Depth from blue to green to red. Images were generated using a customized PyMOL [64].

applied to the substantially larger '291' set of 144 hyperthermostable-mesostable pairs [51], and the results are shown in Figure 4-8. While there is considerable scatter, the data again show that hyperthermostable proteins have significantly shallower surfaces, as can be seen in the linear trendlines. By analyzing the means of the ratios $T_{av}/H^{1/3}$, a statistical analysis was performed, resulting in a p-value of 5.6×10^{-5} indicating that the differences are significant.

In order to check the sensitivity of the Burial Depth and Travel Depth analysis to slightly different structures, we ran the analyses on a complete set of NMR structures forming one hyperthermostable-mesostable pair. The structures chosen were PDB codes 1JDQ and 1JE3 [179]. Each had 20 models. The mean Burial Depth had standard deviations of only 0.017Å for both the hyperthermostable and mesostable protein. Since the mean difference in burial depth from Table 4-1 is nearly ten fold greater, this indicates that the Burial Depth analysis is not very sensitive to changes in which NMR structure was used. The mean Travel Depth had standard deviations of 0.24 Å and 0.13 Å for these two molecules, , which is smaller than the difference in means (0.5Å), also showing that Travel Depth is also not very sensitive to which NMR structure is used. Use of a single NMR structure out of the complete set of models seems reasonable.

The differences in burial depth of each specific residue type were also examined by comparing burials of the C_β atom (C_a for Glycine). Results are shown in Figure 4-9 for each of the 4 thermostable-mesostable classes. P-values for these individual residue burial differences were calculated by permutation, as described in the methods section. In computing the P-values, computed mean burial differences we first

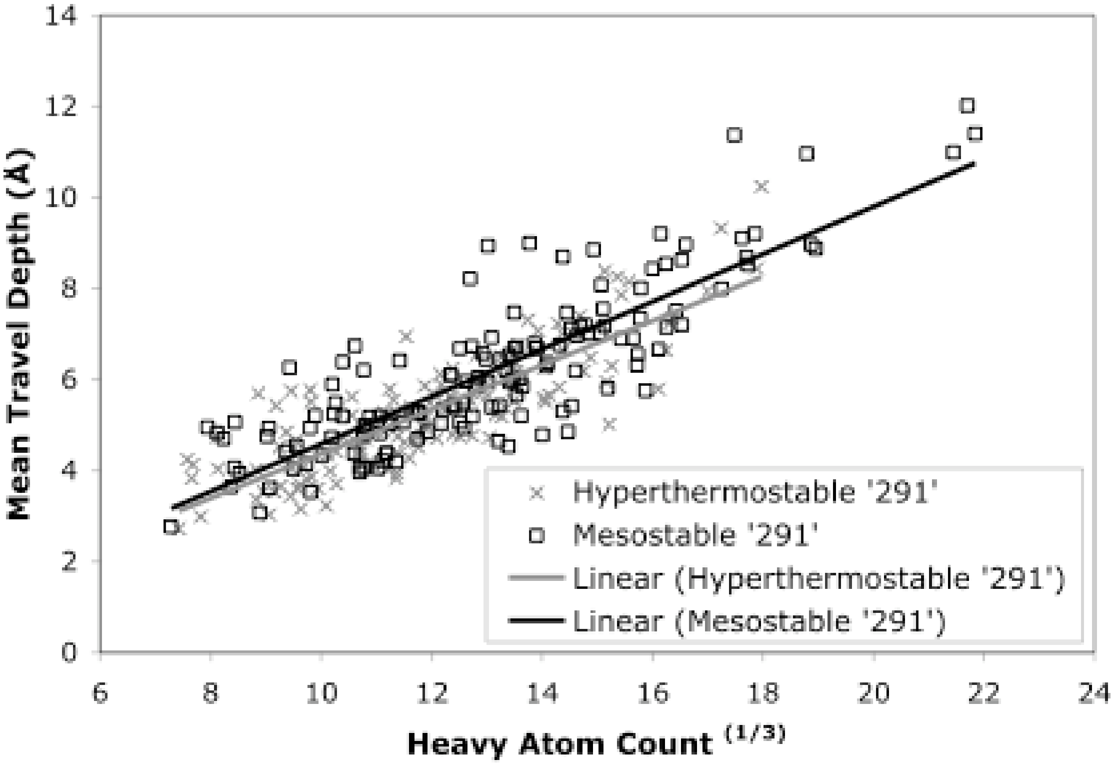


Figure 4-8 Size-Scaled Travel Depth for the Larger Hyperthermostable Set

Along the x-axis is the cube root of the heavy atom count, along the y-axis is the mean Travel Depth. Data for the '291' set of 144 hyperthermostable proteins and matched mesostable proteins is shown along with trendlines and p-values. The mean of the ratios for the hyperthermostable proteins is 0.445 and for mesostable proteins the mean is 0.472, the p-value of this difference is 5.6 x 10^{-5}.

corrected by subtracting the difference in mean Burial Depth of C_β atoms (C_a for Glycine) over all residue types as described above. These corrections were 0.028, 0.13, 0.135, and 0.126 for the moderate thermostable, hyperthermostable, recent and ancient classes respectively. Residues with significant differences ($p<0.05$) are indicated by their p-values on the figure.

The hyperthermostable-mesostable dataset has the largest amount of significant differences as shown in Figure 4-9b. Alanine shows the largest change and is significantly more buried in hyperthermostable proteins as shown in detail in Figure 4-10. Cysteine, Tryptophan and Valine show large trends to being more buried, but these changes are not statistically significant according to the analysis, after correcting for overall depth differences. Six residues are less buried in the hyperthermostable proteins. Given that the correction factors are all positive (*all* residues on average are more buried in hyperthermostable proteins) it would be more correct to say that these six residues stay unburied in hyperthermostable proteins while all other residues get slightly more buried and alanine is much more buried. These 6 residues are the 4 charged residues (Aspartic Acid, Glutamic Acid, Lysine and Arginine) as well as Histidine and Asparagine. Note that Histidine can also likely to be charged as the pKa is near physiological conditions. Since Asparagine is chemically labile at high temperatures [52] and may spontaneously deaminate to Aspartate, probably all the residues we find less buried in hyperthermostable proteins are charged. Only one that may be considered charged at high temperatures (Glutamine, chemically labile at high temperatures forming Glutamic Acid) is not less buried. This result is consistent with previous studies indicating

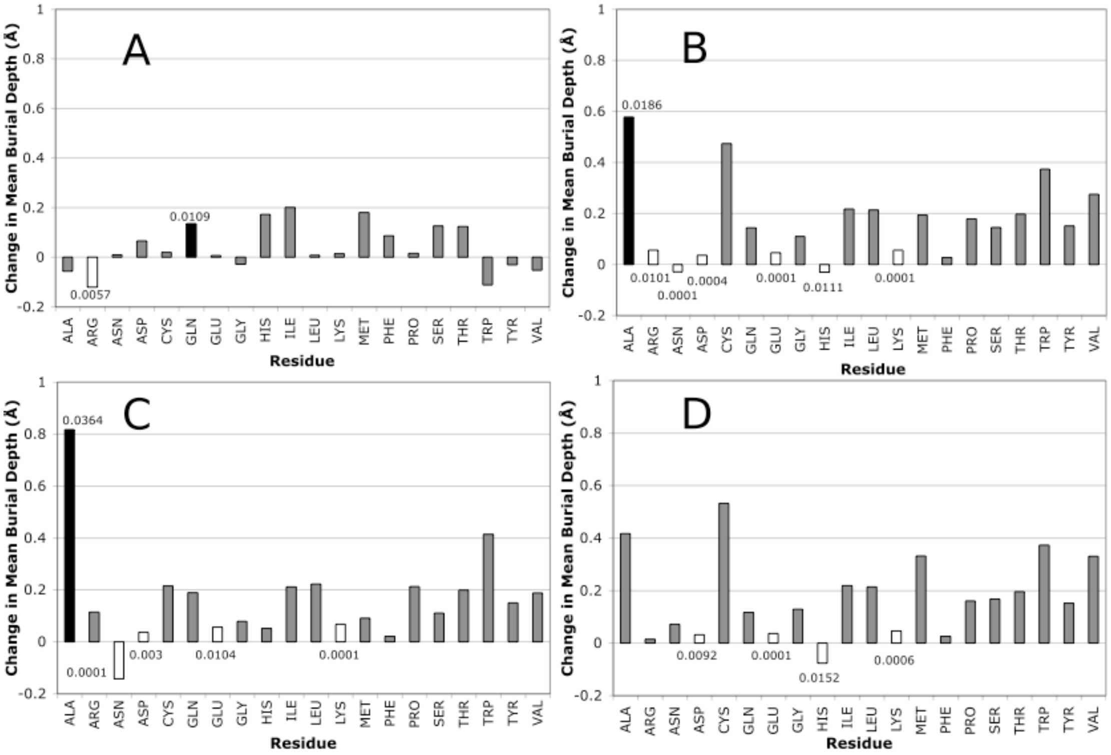


Figure 4-9 Residue Specific Burial Depth in Thermostable Proteins

The difference in Burial Depth for the C_β atom of each residue type (C_a for Glycine), expressed as Thermostable minus the matched Mesostable proteins. Significantly more buried residues are shown in black, significantly less buried residues are shown in white, p-values for significant differences are shown above or below each bar. These p-values were corrected for the overall Burial Depth differences seen between each thermostable-mesostable set. (a) Moderate Thermostable. (b) All Hyperthermostable. (c) Recent Hyperthermostable. (d) Ancient Hyperthermostable.

increased ion pairs and ion pair networks in thermostable proteins [48; 51; 52; 152; 153; 154; 155].

Recent and Ancient hyperthermophiles, as subsets of the entire hyperthermophile set, yield similar results as the latter (Figures 4-9c and 4-9d respectively). Interestingly, though Alanine is more buried in the Ancient class, it is not significant according to the statistical analysis, with a p-value of 0.15. However, four of the above six 'charged' amino acids are also significantly less buried in the Ancient class. In the Recent class, Alanine is significantly more buried, the difference being even more pronounced than in the combined data. Again, four of the six 'charged' amino acids are significantly less buried, although it is a different four from the Ancient category. We emphasize that these are significant differences in residue burial that do not show up with just a surface/interior binary data analysis [51; 53].

The distribution of burial depth of individual residue types that are significantly more or less buried in hyperthermostable proteins was examined in more detail by comparing the complete probability distribution histogram of burial depths. Results are shown in Figure 4-10 for just for one especially interesting case, alanine. Alanine was found to be significantly depleted overall in hyperthermostable proteins, while at the same time less exposed [51]. It has been suggested that this relative enrichment of buried alanine is due to the zero side chain entropy cost [51]. Our results also show that alanine is depleted near the surface of hyperthermostable proteins compared to mesostable proteins, and moreover that it is enriched right into the protein core (i.e. at Burial depths down to 6Å).

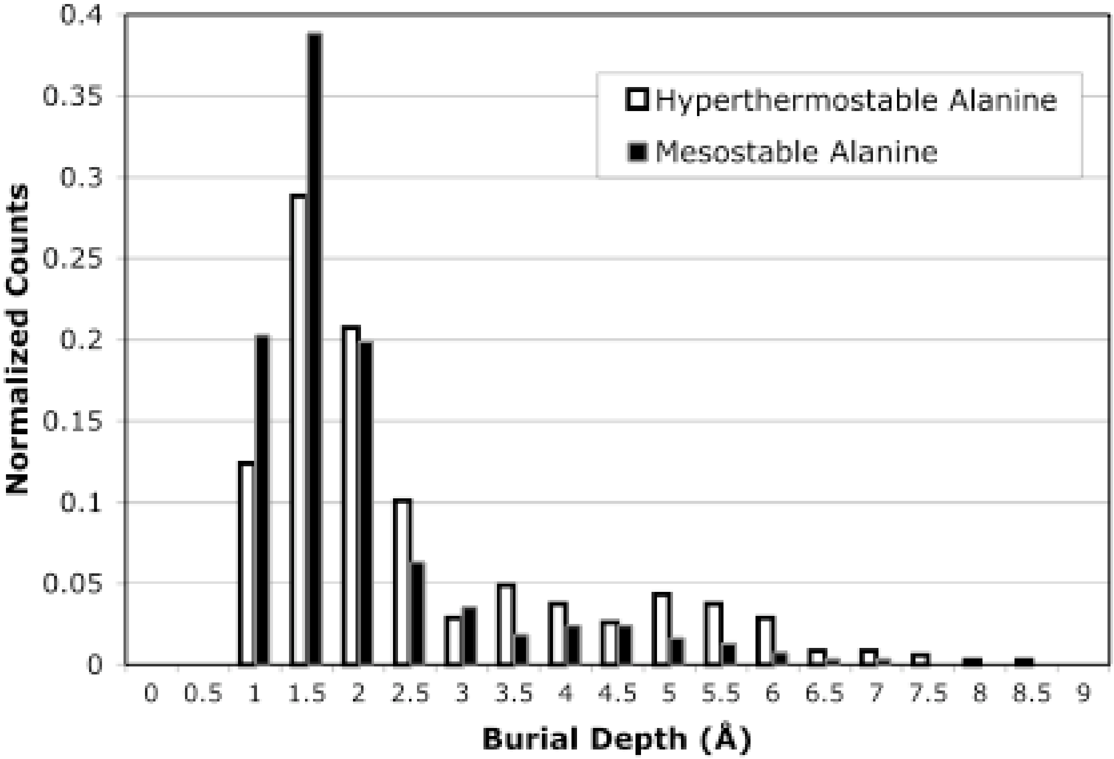


Figure 4-10 Histogram of Burial Depth of Alanine

Normalized frequency histograms of Burial Depth of the C_β atom for Alanine for mesostable and all hyperthermostable proteins.

Discussion

We examined ten different metrics of protein ‘shape’, using three sets of thermophile-mesophile pairs: Moderate thermophiles, ancient hyperthermophiles and recent hyperthermophiles, 67 pairs of proteins in all. Each thermostable protein was matched to the mesostable homologue of similar size, in order to increase the statistical resolution of the comparisons. For the moderate thermophile–mesophile set none of the metrics were significantly different. For the hyperthermophiles, significant differences in several metrics were found, including packing of surface atoms, surface area, Wadell Sphericity, travel depth and burial depth. Of these, only Wadell Sphericity, travel depth and burial depth were significantly different in both recent and ancient hyperthermostable proteins.

Hyperthermostable proteins on average have a higher Wadell sphericity, have fewer and or less deep pockets on their surface, and their residues are on average more deeply buried than in their mesostable counterparts. Taken together, these three metrics provide the first quantitative evidence that hyperthermostable proteins are more spherical. The fact that moderate thermostable proteins show no significant differences in overall shape while hyperthermostable proteins do is in line with previous reports moderate thermophiles and hyperthermophiles have achieved their necessary thermostability by different mechanisms [51]. We cannot rule out the possibility that some other shape metric would reveal differences between moderate thermostable proteins and their mesostable counterparts. However, given the fact that several of the shape metrics do reveal differences for hyperthermostable proteins, we conclude that adaptation to moderately elevated temperatures requires changes at the individual residue level that need produce little change in gross physical aspects of the proteins to achieve the necessary moderate increase in stability.

We consider now in more detail what the individual metrics reveal. Regarding packing efficiency, the most reliable metric is that for buried atoms, and this shows no significant difference between hyperthermostable and mesostable proteins. The similarity in interior packing is consistent with other analyses [159]. It is interesting to note that the observed increase in hydrogen bonding at higher temperatures does not correlate with increased packing in the interior [48; 150; 151; 152]. The packing of surface atoms is greater on average in one class, ancient hyperthermostable proteins and the superclass of all hyperthermostable proteins. This is consistent with some evidence that surface residues have more contacts in thermostable proteins than mesostable proteins [53]. However, the conclusion that in one class surface atoms are better packed must be qualified. The definition of packing of surface atoms is not agreed on, and another definition may lead to different results. This ambiguity is illustrated by considering the absolute values of the packing efficiency. Moreover, surface atom packing is confounded by curvature effects in some definitions. In our method, for example, exposed atoms near convex surfaces will have lower packing than exposed atoms near flat or concave surfaces.

Hyperthermostable proteins in all categories (Ancient, Recent and Combined) have significantly smaller mean Travel Depth and mean size-scaled Travel Depth than their mesostable counterparts, indicating fewer and shallower surface pockets. Using the size-scaled travel depth metric, we also find that hyperthermostable proteins have shallower and fewer surface pockets in a larger set of 144 protein pairs. This is one piece of evidence that hyperthermostable proteins are more spherical. The second piece of evidence for increased sphericity is that all hyperthermostable categories have higher mean Burial Depth than mesostable proteins, indicating that they bury more atoms overall. This agrees with increased hydrogen bonding [48; 150; 151; 152], increased apolar contact area [157] and increased van der Waals contacts [150; 152; 159]. This increased contact area that manifests across several types of interactions (hydrogen bonding, apolar, van der Waals) is reflected in the overall increase in burial depth.

The third metric related to overall sphericity of the protein is the Wadell Sphericity measure [175], which is simply a dimensionless ratio of volume to area, scaled so that its upper bound value of 1 indicates a perfect sphere. This measure is significantly increased for all three hyperthermophile categories: recent, ancient, and combined. We note that this ratio is considerably more sensitive than changes in volume or surface area alone. There is no significant difference in volume for any thermostable category, while a significant difference in area is only seen in ancient and combined hyperthermostable proteins.

The significant differences in these three metrics lead to the conclusion that hyperthermostable proteins are more spherical than their mesostable homologues. This difference is consistent across both Ancient and Recent hyperthermostable proteins, despite the different evolutionary paths those organisms have used to achieve thermostability [150].

While mean travel depth, mean burial depth and Wadell Sphericity all indicate increased sphericity in hyperthermostable proteins, they are by no means synonymous since they each reveal different, complementary, aspects of protein shape. Each is useful in analysing shapes as complex as those adopted by proteins since each can distinguish some feature that the other cannot. This is illustrated schematically in Figure 4-11, which depicts two idealised structures with identical volume and surface area (and hence Wadell sphericity), but with different mean travel depth and burial depth. The structure with one large pocket has less deeply buried atoms, and greater mean travel depth than the structure with four smaller pockets. In this case depth measures are more discriminating than Wadell Sphericity. On the other hand, the mean Travel Depth of any convex shape is zero, while the Wadell Sphericity (and mean burial depth) vary depending upon the shape, so the latter two would be more discriminating. Burial depth has the additional bonus of being able to examine changes in specific residues, whereas Wadell Sphericity only measures total changes in volume and surface area. Combining information from these complementary metrics can reveal other aspects of shape. Returning to Figure 4-11, we see that at constant Wadell sphericity the structure with a smaller mean travel depth has a more convoluted, one might say, 'rougher' surface. A straightforward measure of the roughness of the protein surface is not possible, however, when both Wadell sphericity and Travel Depth are different, as in the hyperthermostable-mesostable comparison. Here, as Figure 4-7 illustrates, the hyperthermostable protein has a smaller mean travel depth, and increased Wadell sphericity, and visually at least, has a less 'rough' surface.

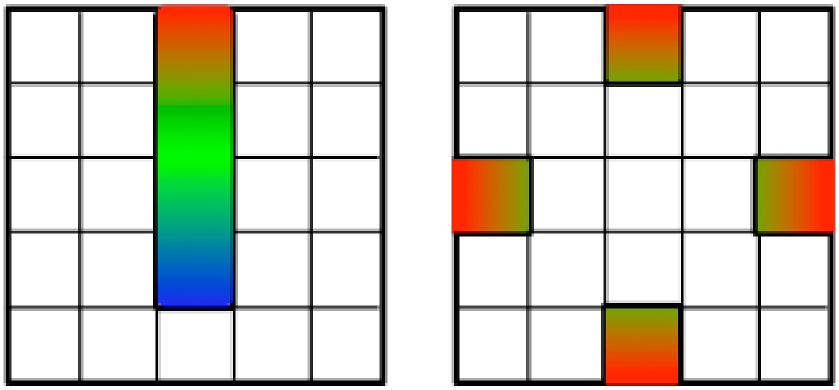


Figure 4-11 Equal Wadell Sphericity, Different Travel Depth and Burial Depth Example

2D Schematic indicating two shapes with equal volume and surface area, and hence equal Wadell Sphericity, but different mean Travel Depths and different mean Burial Depths. The ‘U’ shaped volume on the left has a greater mean and maximum Travel Depth than the ‘X’ shaped volume on the right. The ‘X’ has a higher mean Burial Depth, evident simply by observing the center square is not adjacent to the surface whereas all squares in the ‘U’ are adjacent.

The Burial Depth and Travel Depth analyses while fast, do require computing and working with a molecular surface and therefore take some significant calculation time. We attempted to come up with a faster measure that would also capture the different in ‘spherical property’ between mesostable and hyperthermostable proteins. To this end we calculated the interatomic distances between all pairs of heavy atoms in the proteins and examined the differences in the means. We found no significant differences as shown in Figure 4-2 and Table 4-1. In retrospect, the failure of the mean interatomic distance to detect differences makes sense as it reflects to a great extent the packing, which is not significantly different. In summary, the failure of interior packing and interatomic distances to differentiate thermostable from mesostable proteins shows that the Travel Depth and Burial Depth analyses are necessary to measure the spherical property.

The shapes proteins adopt have profound energetic effects on both charge-solvent and charge-charge interactions, and both travel depth and burial depth report on this. Greater burial depth indicates that more atoms are buried further from solvent. Although many of the deeply buried residues will have apolar sidechains, the backbone of each residue is still polar. Burying the backbone partial charges further from solvent lessens their favorable long range electrostatic with the higher dielectric solvent, increasing the desolvation penalty- these charges are less stable. Similarly a charged group at the bottom of a deep pocket, as measured by Travel Depth, will on average has less high dielectric solvent near it, and more low dielectric protein than a charge at the bottom of a shallow pocket, even though their solvent accessible surface areas are the same. Charges at the bottom of a deep pocket have a greater desolvation penalty and therefore are less stable. This is a manifestation of electrostatic focusing [180]. Considering now a favourable charge-charge interaction, increased burial depth or increased travel depth will strengthen it, thereby having the opposite effect on the protein's stability. The reason is the same. There is less effective solvent interaction with the deeper charges, so less dielectric screening. Fine tuning and balancing these competing effects is one possible way for stability to be controlled in the transition between mesophile and thermophile, and a reason why travel depth and burial depth show significant differences.

Examination of the burial depth of specific residues (Figure 4-9) adds another dimension to the previous surface area change analyses [48; 49; 50; 51; 52], which measured changes in solvent accessible surface area. Surface accessibility analysis on the same dataset used here led to the conclusion that Alanine and Proline have less surface non-polar area exposed in both moderate thermostable and hyperthermostable proteins, and that phenylalanine, methionine, tyrosine and tryptophan have more exposed non-polar surface area in the higher temperature classes [51]. In contrast, we find here that Alanine is buried significantly more deeply in the hyperthermostable proteins, while there is no appreciable change in the moderate thermostable proteins. Proline is more buried in hyperthermostable proteins, but not significantly more buried after correcting for overall burial, and again, no appreciable change is seen in the moderate thermostable proteins. Our disagreement with the result of the four residues having more exposed non-polar surface area could be due to two factors: The previous analysis was only of the nonpolar surface area and changes in surface area may not be directly comparable to counting residues at each burial depth. For instance a residue on the surface in our analysis will have the same burial depth no matter the local environment, however a concave region could result in less surface area where a convex region could result in more area.

Our examination of charged versus polar surface area show that there are indeed differences between the thermostable and mesostable proteins, as first indicated by Cambillau et al [49]. Charged residues are less buried in hyperthermostable proteins and half the polar residues (Serine and Threonine) are more buried, consistent with observed changes in surface area. However, the other polar residues (Asparagine and Glutamine) are significantly less buried, which disagrees with the surface area results. Since surface area changes are analyzed as percentages, difference overall surface areas between hyperthermostable and mesostable proteins does not account for this. These unburied residues, even though they are higher in number, may in fact expose less surface area. Regardless of the disagreement on polar residues, charged residues show large changes consistent across both surface area [49] [51] and burial depth analyses.

In the moderate thermostable proteins, shown in Figure 4-9a, the changes are the least pronounced of any of the four categories examined. However, two residues still show up as significantly different: Glutamine is more buried and arginine is less buried, something that could not be detected in previous analysis of nonpolar surface area.

Our residue-specific results can most easily be compared to results that used an interior/exterior definition based on surface area and then counted residues in these classes to see how they differed between thermophiles and mesophiles [53]. In that study, the interior counts showed very little changes, while the exterior showed many changes, though the significances were not analyzed statistically. Lysine, arginine, and glutaminic acid had increased exterior numbers in thermostable proteins, which our analysis would agree with as those residues remain unburied or become less buried in hyperthermostable proteins. Alanine, asparagine, aspartic acid, glutamine, threonine, serine and histidine all have less exterior residues in thermostable proteins. We agree on alanine's decreased exterior presence (and increased burial), but disagree on the other residues found less frequently on the exterior. This could be due to the differing data sets, use of a full spectrum of burial depths *vs.* a binary cutoff, or statistical variation.

A possible application for the burial depth preferences in hyperthermostable proteins is thermostable protein design. The amino acid burial preferences could be incorporated into models and design strategies. Additionally, using mesostable protein structures as a starting point, a design pipeline could incorporate the Travel Depth and Burial Depth analyses of the spherical property to find structures that bury more atoms and have fewer/smaller pockets. Obviously, a good protein design strategy is required as a starting point as it is not just the spherical property that ensures a protein is highly thermostable. A suggested pipeline would be to find many backbones [181], repack the native mesostable sidechains and mutations chosen from residue-specific Burial Depth changes [182], then evaluate the many possibilities to find proteins that are spherical but preserve the active site or desired function. Then traditional protein design tools, such as [183], could be used to find further mutations that enhance the stability of the new backbone, and could be modified to include residue burial depth preferences of hyperthermostable proteins. This method is obviously an addition to already existing approaches for thermostable protein design [184; 185; 186; 187].

When interpreting differences between thermostable and mesostable proteins, organism sources should be considered, as temperature is not the only difference between these organisms. In the hyperthermophilic set of 30 structures, 9 organisms are represented. At least one is a piezophile, but many are not, so it is doubtful that we are seeing results from changes due to adaptation to high pressure. However, all these hyperthermophiles are unicellular and many of the mesophile homologues come from multicellular organisms. At this point there is no evidence that this systematic difference is reflected at the level of protein structure, but it is a caveat nevertheless.

Another important caveat of our analysis (indeed of any type of analysis of the PDB database) is the experimental temperature at which the structures were determined. Protein crystal structures are now almost always solved at extremely low temperatures (ca. 130K). Even older structures or typical NMR structures are solved at room temperature, far from the environmental temperatures of hyperthermophiles. This could have several effects. Obviously at higher temperatures, the configurational entropy of the side chains will be higher and will explore more states. This could have some influence on hyperthermostability, as demonstrated by simulations showing that the charged residues in a hyperthermostable protein are able to interact cooperatively during the conformational fluctuations [188]. Our results that hyperthermostable proteins are more spherical could indicate a preference for finding conformations that have reduced flexibility, since the higher number of interactions will limit the number of states available, consistent with reduced rotameric states in hyperthermostable proteins [51]. These results are also consistent with results showing that mutations that support hyperthermostability are distributed throughout the protein and cause subtle changes in dynamics and distributed changes in stability [189; 190]. However, without structures solved at the ambient temperatures for mesophiles and hyperthermophiles, it is difficult to say how our results would be affected. The effect of experimental conditions on structures is a caveat of any research based on PDB structures [16].

Finally, in any discussion of thermostable proteins, it is important to note that the language used throughout almost all the literature (and in this work!) implies that previously mesophilic organisms have adapted to higher temperatures resulting in the hyperthermostable proteins. This is probably not the case[147; 148] except in specific cases like *T. maritima* [149; 161]. If the 'hot origin' of life theory is correct, then a common ancestor organism for these proteins was a hyperthermophile, though some organisms (and therefore their proteins) adapted to mesophilic conditions and then re-adapted to hyperthermophilic conditions [149; 161]. While this does not affect the observed differences and their statistical significance, this 'meso-centric' view does shade their evolutionary interpretation. Most mesostable proteins whose temperature dependence of stability has been examined in detail show a maximum in stability not too far from their working temperature with a substantial decrease in stability with increasing temperature above 45^{o}C. This tends frame the question in terms of how this stability profile is changed in thermostable proteins to ensure stability at high temperatures. This could be achieved by a) shifting the maximum of stability to a higher temperature, b) being more stable at all temperatures, c) reducing the rate at which stability decreases with temperature (or some combination of these effects). Viewed, however, from the perspective of the thermophile as the precursor, cases b) and c) present no problem in adaptation to mesophilic temperatures, since at these lower temperatures the thermostable protein is *already* stable. In this scenario there would be no selective pressure, and one would not expect to see pervasive stability related structure changes of the type observed here. In case a) however, presumably the stability of a thermostable protein at mesophilic temperatures would be low enough so that there would be selective pressure to adapt to lower temperatures, leading to significant stability correlated structure changes. Of course proteins need enough flexibility and dynamics to function, and cases b) and c) may result in too much stability at mesophilic temperature for optimal function, in which case again there would be selective pressure. In a recent review of available experimental evidence, hyperthermostable proteins used case b) most often, often combined with case a), whereas moderate thermostable proteins used case b) often combined with case c), however there is still not a lot of data available [191]. Considerably more data on the temperature stability profiles of matched mesostable-thermostable pairs is needed to distinguish these cases.

Conclusion

Protein structures from homologous mesophiles and hyperthermophiles have diverged due to evolution. Regardless of the age of the adaptation to hyperthermophilic conditions, the proteins adopt a more spherical structure, namely they have greater Burial Depth, lesser Travel Depth, and higher Wadell sphericity than their mesostable counterparts. The interiors of these hyperthermostable proteins are not more tightly packed, probably because mesostable proteins are already packed to near crystalline tightness. Rather these proteins have residue side-chain replacements and structural rearrangements that produce more spherical proteins. These changes are not detectable by other properties like mean interatomic distance or convex hull volume. The new metrics of Travel Depth and Burial Depth analyses are necessary to quantify the spherical property and complement Wadell Sphericity. All three metrics are applied here to proteins for the first time. In contrast to hyperthermostable proteins, moderate thermostable proteins do not show any significant differences in sphericity metric from their mesostable homologues. Moderate thermostable proteins adaptations to stability clearly do not drive them to more spherical structures. In this way, our results support the hypothesis that moderate thermophiles and hyperthermophiles achieve the enhanced stability of their proteins by different mechanisms.

Additionally, by adding a new dimension to specific residue analysis, distance of burial instead of the binary buried/exposed metric, key observations about hyperthermostable proteins can be made, specifically that charged residues stay unburied, alanine is considerably more buried and the rest of the amino acids become slightly more buried.

Chapter 5

This chapter will be published in the future [59].

Summary

The shape of the protein surface dictates what interactions are possible with other macromolecules, but defining discrete pockets or possible interaction sites remains difficult. First, there is the problem of defining the extent of the pocket. Second, one has to characterize the shape of each pocket. Third, one needs to make quantitative comparisons between pockets on different proteins. An elegant solution to these problems is to sort all surface and solvent points by Travel Depth, and then collect a hierarchical tree of pockets. The connectivity of the tree is determined via the deepest saddle points between each pair of neighboring pockets. The resulting pocket surfaces tessellate the entire protein surface, producing a complete inventory of pockets. This method of identifying pockets also allows one to easily compute important shape metrics, including the problematic pocket volume, surface area, and mouth size. Pockets are also annotated with their lining residue lists, polarity, and other residue based properties. Using this tree and the various shape metrics pockets can be merged, grouped, or filtered for further analysis. Since this method includes the entire surface it guarantees that any pocket of interest will be found among the output pockets, unlike previous methods of pocket identification. The resulting hierarchy of pockets is easy to visualize and aids users in higher level analysis. Comparison of pockets is done using the shape metrics, avoiding the shape

alignment problem. Example applications show that the method facilitates pocket comparison along mutational or time-dependent series. Pockets from families of proteins can be examined using multiple pocket tree alignments to see how ligand binding sites or other pockets have changed with evolution. Our method is called CLIPPERS, for Complete Liberal Inventory of Protein Pockets Elucidating and Reporting on Shapes.

Introduction

The shape and properties of the protein surface determine what interactions are possible with ligands and other macromolecules. Pockets are an important yet ambiguous feature of this surface. For example the first pass in screening for lead compounds and drug-like molecules is usually a filter based on the shape of the binding pocket [192], and shape plays a role in many computational pharmacological methods as reviewed by Kortagere et al [193]. A study of drug-binding pockets found that most features important to predicting drug-binding were related to size and shape of the binding pocket, with the chemical properties of secondary importance [90]. The surface shape is also important for interactions between protein and water. This depends, for instance, on how wide or narrow the pocket, or how deep or shallow the pocket as reviewed by Levitt and Park [14]. However, defining discrete pockets or possible interaction sites remains difficult despite many studies, for example see the review of Campbell et al. [58]. Compounding the problem is that the shape and location of nearby pockets can affect promiscuity and binding site diversity [194]. The primary difficulty is in defining the border of a pocket, as most pockets are open to solvent. Those closed to solvent we refer to as buried cavities.

Buried cavities are more straightforward to locate as they have a well defined extent, area and volume. In contrast, the border of an open pocket defines its mouth and it provides the cut-off for determination of the surface area and volume. The border definition problem for open pockets has been discussed before as a 'can-of-worms problem' [103]. Even defining the pocket as a set of residues does not define the volume or the mouth of the pocket.

Several very different solutions, and therefore pocket definitions, have been proposed. These include fattening the atoms to close off pockets [103], defining pockets as clustered sets of spheres [71; 86; 195; 196; 197; 198; 199; 200], by using discrete flow analysis on alpha-shapes[72], and by using a larger probe radius to construct a surface or alpha-shape that acts as the pocket mouth [70; 90; 201], by examining clusters of lines through solvent [202; 203], by defining pockets of interest to only fall in a narrow range of surface areas and shapes and then generating multiple overlapping pockets covering the protein surface for evaluation [204]. Other methods focus only indirectly on shape, for instance by examining pockets predicted by evolution [89] or by protein motion changes upon binding [205]. Various combinations of these methods are also employed [76; 206], including methods that find regions where certain combinations of features are clustered or combined within a statistical framework [207; 208].

A common problem with any specific definition of a pocket or any method for finding a small number of non-overlapping pockets on a protein is that they may miss the actual pocket of biological interest. For example defining pockets to be bottlenecks (a narrowed region of the pocket that defines the mouth) as several methods do will miss non-bottleneck pockets, such as clefts, entirely. Other methods and definitions can also miss certain types of pockets or need parameter adjustment to capture relevant pockets.

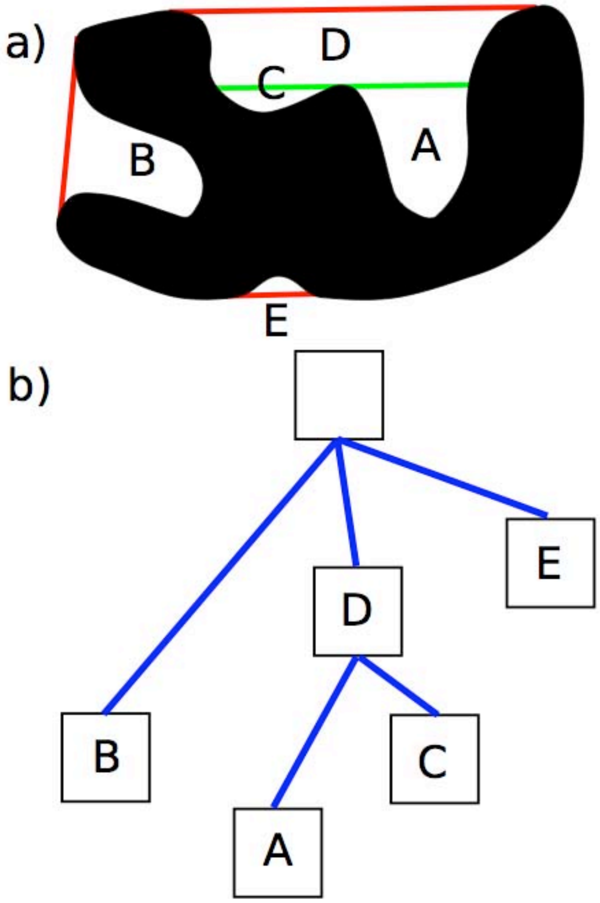


Figure 5-1 Pockets Example

a) Schematic protein, molecular volume shown in black, the convex hull shown as red lines. Pockets are labeled and the split line where two sub-pockets are joined is shown in green. b) Corresponding pocket tree.

We present here an alternative definition of pockets, one general enough to create what we call a complete inventory of pockets: In this inventory the entire surface is tessellated into protein pocket regions, each pocket being organized into a hierarchical tree of sub-pockets. The basic idea is illustrated in Figure 5-1, and it described in detail in the methods section. If a protein had a molecular surface which was convex everywhere, this surface would be identical to what is known as its convex hull [25]. Clearly such a protein would have no pockets, however relaxed the definition. However a real protein's molecular surface is not identical to its convex hull; it lies within the latter surface at many points (Figure 5-1a). Thus in seeking pockets our attention is directed to *both* the molecular surface that lies within the convex hull, *and* the solvent accessible volume that lies between the two surfaces (the intermediate volume). It is in this combined surface/volume region that every protein pocket must lie. The foundation for inventorying the pockets is Travel Depth [35]. Travel Depth is an efficient way to determine the shortest distance, traveling only through solvent, from any point on the molecular surface point or in the intermediate volume to the convex hull, this distance provides the basis for the inventorying step.

In addition to presenting a new definition of pockets, a new way of comparing pockets is described. Most algorithms for comparing two binding sites assume the binding site is known or locate it solely based on proximity to a ligand in the co-crystal structure. After that most algorithms that use spatial information to come up with a motif of various chemical properties and their arrangements in space, and rely on some alignment or geometric hashing technique to compare binding sites based on these structural motifs [209 206; 210; 211; 212; 213; 214; 215; 216; 217; 218; 219; 220; 221]. Motif definitions can involve hydrogen bond donors or acceptors, residues, or atom types based on residues or can involve the complete set of docked substrates [222]. Here we present a new method of comparison of pockets based solely on the shape features.

We first describe the use of Travel Depth to create a complete inventory of protein pockets, including construction of the complete tree of protein pockets, then we describe the computation of various pocket metrics and a way to quantitatively compare pockets. We then show various applications of the methods, including display of pockets and visualization of pocket properties, analysis of pockets along mutational and time series of structures, and the clustering of pockets from different members of evolutionarily related protein families.

Methods

Computation of Travel Depth

This work builds on the concept of Travel Depth, first used to analyze surfaces and ligand binding sites [35], with subsequent speed and algorithm improvements [43]. The Travel Depth algorithm computes the shortest molecule interior-avoiding paths from all surface points to the convex hull of a given macromolecule. The algorithm also computes the Travel Depth of points in the intermediate volume between the molecular surface and the convex hull. Additionally the algorithm puts the surface points and volume grid points in a graph structure with the distances between each point as the edge lengths between adjacent nodes, which aids in later steps. The outline of the algorithm is as follows:

Starting with the atomic coordinates, the molecular surface [7] is generated using a standard 1.2Å solvent probe radius. The convex hull of this surface is generated using the Qhull algorithm [25]. These surfaces are mapped onto an appropriately scaled cubic grid, and all grid points are assigned to either to the interior of the molecular surface, outside the convex hull, or between the two surfaces. The Travel Depth of all molecular surface points and intermediate volume grid points is computed as described previously using the multiple source shortest paths algorithm [23], avoiding the interior points.

We extend the original Travel Depth algorithm here to include a definition of Travel Depth for buried cavities. Previously these cavities were removed completely, which made analyzing ligands inside them impossible. The extension to buried cavities is done by adding one 'virtual' edge per cavity to connect it to the exterior molecular surface. This edge connects the closest cavity and exterior surface points. The length of this edge defines the Burial Depth of that cavity [57]. After adding a virtual connecting edge to each buried cavity the Travel Depth algorithm is applied as described above. Due to these connecting edges, Travel Depth values are now propagated to all buried cavity surface points and their enclosed volume grid points.

The rationale for defining the burial depth of a cavity by the shortest distance to the main surface is that this route would require the least amount of protein motion to open the cavity to bulk solvent. Of course the protein may open by a different route, and if experimental or simulation data were available, a more accurate burial depth estimate could be made. Nevertheless, the closest distance connection is a useful device to seamlessly include cavities in the analysis of pockets.

Pocket Inventory

The goal of this step of the algorithm is to enumerate all pockets by analyzing all regions of the molecular surface that lie below the convex hull. By enumerating all pockets over the entire protein surface we produce an unbiased collection, rather than focusing *a priori* on a subset of possible pockets.

The inventory algorithm has two phases. In the first phase, all surface and intermediate volume grid points with a defined Travel Depth are put into a list and that list is sorted so the deepest points are first. Ties are broken randomly, but the sorted order is kept fixed throughout the algorithm. To keep track of pockets, a union-find data structure P, is initialized, [223] [28]. P is essentially a list of lists, each sub-list containing the surface and volume points belonging to a single pocket, Pj. Also a tree data structure T, whose nodes will be pockets, is initialized.

In the second phase of the algorithm, each point in the sorted list is examined in turn, starting with the point with the greatest Travel Depth. For each point, *i*, there are three possible cases:

i) The point *i* has no neighbors already in P. In this case, a new pocket P_j is added to P, the point *i* is added to P_j's list of points, and a new leaf node P_j added to the tree T. The depth of point *i* will be the maximum depth of the new pocket.

ii) The point *i* has neighbor(s) in only one pocket of P, P_k. The point is added to P_k's list of points.

iii) The point *i* has neighbors in two or more pockets in P, say pockets P_j….P_k. The point *i* and the point lists of all sub-pockets P_j….P_k are added into the point list of a new pocket P_l. The pocket P_l. is added as a new node in T, and the existing sub-pockets nodes P_j….P_k are indexed as descendents of P_l. The depth of point *i* will be simultaneously the minimum depth of all the pockets P_j….P_k and the height of the deepest saddle point connecting these sub-pockets.

In summary, in this phase of the algorithm there are three possible operations: i) finding a new pocket, ii) adding to an existing pocket, iii) merging pockets.

Once all points have been examined, the points in all the top level pockets of T are unioned into a final mother of all pockets which forms the root of T. This pocket contains all parts of the molecular surface that lie within the convex hull, and the entire intermediate volume.

The result of the algorithm is therefore a complete tree of pockets, T. Each node of T is a pocket, and each pocket contains all the volume and surface points of each of its descendent pockets, plus points specific to itself, i.e. the smaller pockets are nested inside the larger pockets. Every molecular surface point and intermediate volume point has been assigned to a pocket and hence to all antecedents of that pocket. Each saddle point has been assigned to two or more pockets, and the resulting merged pocket. Each leaf node of this tree represents a pocket containing a single local maximum in Travel Depth, i.e. a simple pocket. As we ascend the tree, the pockets become increasingly larger and more complex, with multiple local maxima in depth (sub-pockets), i.e. they are compound pockets. The mouth or mouths of a given pocket are defined as the union of surface and volume points belonging to that pocket which are on its boundary. i.e. that have at least one neighbor that is *not* in that pocket. Each pocket has other associated shape, physical and protein related properties as described in the next section.

Pocket Collation

To facilitate collation, filtering, comparison and clustering of pockets, various features or metrics of each pocket are computed.

First are the global geometric features: volume, surface area, and principal axis dimensions. Second are the mouth geometric features: number of mouths, mouth area(s), and largest mouth linear dimension(s). Third are residue based properties: Lists of residues lining the entire pocket andlining the mouth. Fourth are physico-chemical properties: including surface area of positively charged, negatively charged, or neutral (apolar) atoms. Fifth are secondary surface properties: mean curvature and mean absolute curvature (roughness). The sixth set of properties, unique to this work, are Travel Depth related: height (maximum Travel Depth – minimum Travel Depth), mean height (mean Travel Depth – minimum Travel Depth), absolute maximum Travel Depth.

Curvatures are computed by analyzing the angle between adjacent triangles of the surface, and these are mapped from edges to points by weighting according to the length of the edge. This gives local curvatures, not regional curvatures as computed by other methods [87]. The mouth linear dimension and pocket dimensions are computed by finding the principal components [224] of the mouth or pocket points and then measuring the distance along each dimension. The pocket principal dimensions could be considered similar to finding the global fit of a sphere through all pocket surface points to judge how open the pocket is [87] [225]. Partial charges are assigned using the PARSE parameter set [226], using a cutoff of -0.45 and 0.45 to determine polarity of lining atoms.

These pocket properties are principally designed for quantitative comparison of pockets, as described in the next section. We note that these features could also be used to automate the qualitative classification into pocket types, i.e. bottlenecks, clefts, tunnels, etc based on ratios of appropriate metrics, although we don't pursue that application here.

Another use for these metrics is to identify biological activity associated with various pockets. This would include assessing the likelihood the pocket is an active site, or if the pocket is druggable. This application will be pursued in future work.

Pocket Comparison

To compare the shape of two pockets using either the actual surfaces or lining residue positions requires first, that the surface points or residue atoms of the two pockets be put into a 1-1 correspondence (aligned). The two objects are then overlaid using rigid body superposition, to yield the minimum root mean square deviation (rmsd) for that set of pair alignments. Since it may not be *a priori* evident which parts of each pocket correspond with the other, especially in pure shape matching, many alternate alignments may have to be considered until the global minimum rmsd is found. An alternative is to examine motifs of lining atoms or residues, which may generate thousands of descriptors which have to be matched. Thus pocket shape comparison using positional alignment or indirect lining residue information is fraught with difficulty. In this work each pocket is described by a modest number of shape descriptors, and our goal is to use these descriptors to quantitatively compare pockets avoiding the aforementioned alignment problem.

Since the numerical range and units of each descriptor differ widely, we first express them in dimensionless, normalized units using the information contained in the pocket tree(s), as follows. For the protein or set of proteins of interest, and their resulting pocket trees we first select all the relevant shape descriptors for the particular application. The mean and standard deviation of each descriptor is calculated over all these trees. Each descriptor for the two pockets to be compared is turned into a Z-score by subtracting the mean (for that descriptor) and dividing by the standard deviation (again for that descriptor). Each pocket now has an n-dimensional vector of Z-scores where n is the number of descriptors. The rectilinear, or 'Manhattan' distance in shape space between two pockets P_i and P_j is defined as

$$D_{ij} = \sum_{m=1}^{n} \left| Z_i^m - Z_{ji}^m \right| \qquad (5\text{-}1)$$

where Z_i^m is the Z-score of the m'th descriptor of the pocket *i*.

The default set of descriptors used for shape comparison in this work are: volume, surface area, height, mean height, mean curvature, principal dimensions, number of mouths, mean mouth area and mouth longest dimension.

Use of Z-scores removes differences in numerical range and units for each descriptor and gives each descriptor equal weight in the final analysis. So for example a difference in surface area equal to one standard deviation over the set of all pockets is the same as a difference in one standard deviation in volume. This method of pocket shape comparison requires no alignment, and hence is extremely rapid. It does however use the descriptors as a proxy for full shape comparison. False negative type errors are demonstrably small: If two pockets are significantly different in a single descriptor, say volume or height, then they really must be different. Conversely, if two pockets are similar in all descriptors, and the descriptors are well chosen to represent non-redundant aspects of shape, it is highly likely that they truly are similar in shape and size. However, it does not preclude the possibility that the pockets differ in some aspect of shape that is not measured by the descriptors, so false positive type errors are possible. Using visual examination of many dozens of pairs of matched pockets we found no egregious examples of this error, so we judge it uncommon enough to consider this method of shape comparison robust.

To estimate the descriptor means and standard deviations to compute Z-scores we use the population of pockets for the protein or protein trees under comparison. An alternative approach to this internal standard would be means and standard deviations calculated from a suitable 'standard set' of protein structures. This choice of reference will likely have little effect as the means and standard deviations of the many shape descriptors across several of our data sets were found to be very similar.

Selecting Unique Pockets

For various applications, it is useful to have a measure of pocket uniqueness. This was calculated by comparing each pocket in a given tree to all other pockets in that tree that did not have any lining atoms in common. The distances between the pocket of interest P_j and all m non-overlapping pockets P_i are computed, and the uniqueness score of P_j is defined as

$$R_j = {}^{1}\!/_{m} \sum_{i=1}^{m} {}^{1}\!/_{D_{ij}} \qquad (5\text{-}2)$$

the mean of the reciprocal distances. Unique pockets will have a low value of redundancy, R, since there will be no pockets close in shape space. Conversely, pocket types seen frequently (like small dimples occurring between two or three neighboring non-bonded atoms) will have a high value of R. The uniqueness score allows one to filter out ‘uninteresting pockets’ to focus on ones that have a unique shape and that are therefore more likely to support specific ligand binding.

The uniqueness score is most useful for pockets lower on the pocket tree, where there are many non-overlapping pockets to compare. Pockets very high up on the pocket tree contain large amounts of surface, and there will be few, perhaps no pockets without any atom overlap. These would correctly get low uniqueness scores, but only because the sample size is small. For this reason, in most applications one would only use a uniqueness score combined with some suitable upper volume bound.

The uniqueness filter step in our algorithm takes the place of filtering strategies or parameter variation employed by other methods to generate only the most interesting pockets or those likely to be active sites. The difference is that here all pockets of interest are already contained in the complete pocket tree, so if a particular filtering step does not pick out the required pockets, one can re-examine the complete list.

Clustering and Ordering Pockets

With a well defined pocket-pocket distance in shape space it is straightforward to cluster trees of pockets using standard clustering algorithms. To get useful clustering, however, we add the uniqueness score R as a penalty into the distance formula. This penalizes common uninteresting pockets such as dimples, which would otherwise dominate the clustering. The term in the penalty function used for clustering, due to a pocket pair A-B is

$$\frac{1}{D_{AB}} - \alpha(R_A + R_B) \qquad (5\text{-}3)$$

where D_{AB} is the rectilinear distance in shape space between pockets A and B, and R_A and R_B are the two pockets' uniqueness scores. α is a parameter that can be adjusted to emphasize different sets of pockets. A low value of α favors redundant pockets, a high α value de-emphasizes the redundant pockets. However, α in no way affects the number of pockets found or their position in the tree, only the order in which they are ranked and clustered together. In clustering whole trees we also exclude pockets with volume less than 25Å^3 or more than 2000 Å^3. The former are too small to be of any relevance, while the latter are large compound pockets that consist of multiple sub-pockets which are already included individually in the clustering operation.

For applications involving transitions along a single dimension (like a transition pathway or molecular dynamics run), we found it useful to create minimum spanning "lines". These are similar to minimum spanning trees [28; 227] except the maximum degree of any node is 2 so when the minimum spanning line is fully constructed it gives a connected series from one end to another, each end being defined as having degree one. This is an approximation to the Traveling Salesman Problem [28], where the best solution is one that minimizes the total pocket-pocket distance while visiting each pocket exactly once.

Output files are created that can can be used to visualize these clusters or minimum spanning trees in the graph drawing software packages GraphViz [228; 229] and aiSee[230]. The aiSee version is annotated with snippets of code that can be used to quickly display the pockets of interest in PyMOL [64], a common operation. Nodes can be colored according to which tree they belong to, or by the amount of residue overlap (ignoring ordering) of each pocket to all adjoining pockets.

Additionally, heatmaps of the pocket-pocket distance matrix can be created, which are useful for looking at the variation between sets of pockets of interest or among pockets from a single tree.

Pocket Selection

Once an entire tree of pockets has been collated, a common task will be to examine a pocket of interest. This can be done interactively with PyMOL [64] using our customized scripts. The tree can be followed up or down the branches to look at progressively larger or smaller pockets.

Another common task is to select a pocket or pockets based on a set of residues of interest. This is done most simply by computing a Tanimoto type overlap score: the size of the intersection of the list of residues of interest with the list of pocket lining residues, divided by the size of the union of the same two lists. Perfect overlap gives a score of one, no overlap gives zero. The pocket that maximizes the Tanimoto overlap score, T, is then picked. This part of the procedure is automated. The user can then use this pocket as a good starting point for an interactive search of related pockets up and down the tree using PyMol to refine the pocket selection for a specific application.

A more advanced pocket selection routine for a series of closely related pocket trees involves the following procedure. One initial pocket is selected from each tree based on residue overlap using the Tanimoto type score. All pocket-pocket distances for this pocket set are computed. The pocket with the greatest mean distance to all other pockets is removed, and all other pockets from the same tree with at least 0.5 in overlap to the removed pocket are examined to see which has the lowest mean distance to the other pockets remaining in the set. The one with the lowest mean distance is added to the set so there remains one pocket from each tree. This swapping operation is done iteratively until the pockets remain the same even after examining all pockets in descending order of mean pocket distance. The swapping optimization potentially involves a large number of steps so the procedure is terminated if a large cutoff number of swaps is reached though this cutoff was not reached in our experiments. The swapping optimization leads to a consistent set of pockets along a transition pathway or a mutational series so the differences can be analyzed with minimal bias from the initial residue overlap selection step.

Results & Discussion

We now present various application of the CLIPPERS program for finding and analyzing pockets. As part of this we include several important objective tests of CLIPPERS. First, we claim to generate pockets for every portion of the surface and therefore at least one pocket for any given bound ligand should exist. This is tested on a diverse set of structures with bound ligands, where the resulting pocket trees are searched for pockets that have a high Tanimoto score between the residues lining the pocket and the residues near each ligand.

Second, given a series of structural snapshots of a protein undergoing a transition between two very different conformations, one should be able to follow an evolving pocket through this transition pathway. More specifically, if the pocket shape distance measure is robust, distances between pockets in structures that are neighbors should be smaller than between non-neighbors. In other words, a complete reconstruction of the pocket ordering through the transition pathway should be possible from just the pocket-pocket distance matrix. This is tested in the section of the paper on adenylate kinase.

Finally, the ability to distinguish between pockets associated with less dramatic conformational change, such as those in protein tyrosine phosphatase 1b (ptp1b), can be tested by comparing the pocket-pocket distances between and within evolutionarily related groups, as demonstrated on the protein tyrosine phosphatome.

Comparison of binding site location in SURFNET, CAST and CLIPPERS

As a comparison to two other widely used approaches to finding pockets, we analyze a data set of 67 monomeric proteins with diverse enzymatic activity, originally compiled and analyzed using SURFNET[86]. SURFNET identifies all active sites at least partially, but we note that the algorithm has several parameters that were adjusted to get this recognition. This same data set was also used to test against CAST, though only 51 of the structures were used [72]. 14 structures were excluded since CAST could not analyze the known binding site since the discrete flow method could not find the pocket. Two other structures were eliminated in the original CAST work since they had been superseded in the PDB. We use the newer versions of these two structures here.

These 67 monomers were downloaded from the PDB [16]. Waters were removed and ligands were separated for later analysis. Some complexes contained multiple ligands bound in spatially separated sites. These were split by clustering using a 5Å cutoff, resulting in 92 individual binding sites in these 67 structures. Special attention was paid to including non-standard residues with the protein and to identify peptide ligands correctly. Radii were assigned to the atoms using the radius set of Bondi [3], which is a standard set in the area of macromolecular analysis. SURFNET does not use an explicit probe sphere to construct the surface it uses. However CAST does use a solvent probe sphere, of radius 1.4Å. For CLIPPERS, we used a probe radius of 1.2Å as previously described [43]. Since the three methods have different methods of surface generation, and different radii sets, the surfaces will differ somewhat leading to minor differences in volumes and surface areas. This may contribute to differences in results, although the major effect is the method of pocket finding. In collecting pockets, a lower bound volume cutoff of 25 $Å^3$ was used in CLIPPERS since this represents the volume of a typical heavy atom. This is the smallest pocket that could be considered relevant to molecular recognition, as one ion, water, or other heavy atom could fit into a dimple of that size. Since some structures had ligands in buried cavities, we included these cavities while computing the pockets, as described in the methods section. We note that several of these 67 structures have ligands binding in the non-physiological active site, and some of the active site ligands are much smaller than the actual substrate, as in PDB code 1PII [231] which contains phosphates and not the entire substrate and PDB code 1ONC [232] which contains a sulfate in the active site of an RNase, so while these are valid ligands for the test, they do not reflect accurately the physiological ligand.

Considering first the success rate in finding ligand binding pockets the mean number of pockets per protein generated by CLIPPERS for this dataset of 67 proteins is 431±161. Thus a large number of possible pockets are found covering the entire surface. To score these pockets, the set of residues within a cutoff of 5Å from any ligand atom is generated, and then the Tanimoto overlap score of this residue set with the lining residues of all the CLIPPERS pockets is computed. For all 92 ligands, at least one pocket is generated with a significant Tanimoto overlap, indicating 100% success in generating the binding pocket. Selecting the most overlapped pocket for each ligand, the mean Tanimoto overlap score over the 92 sites was 0.5±0.2, even though the set contained very exposed sites or sites that bound very small ligands like sulfate or phosphate. In other words using CLIPPERS there are enough pocket candidates generated that one finds on average a pocket that overlaps at least 50%, as identified by proximity to the ligand. This is in contrast to CAST, which fails in 14 cases to define a ligand binding pocket, since the discrete flow method cannot find pockets without bottleneck mouths. In examining all 92 pockets found for these ligands, we note that most cases of a low Tanimoto overlap are with ligands that are bound to a very shallow pocket near the convex hull of the protein. The pockets near such ligands tend to be less 'pocket' like. The Tanimoto overlap score can be less than 1 if either the pocket is too small or too large. One example is shown in Figure 5-2. The middle panel on the bottom row has a pocket far larger than one would expect, with a Tanimoto score of 0.25. Despite this poor overlap, CLIPPERS outperforms CAST which cannot find this ligand at all. Also CAST fails on the ligand in the upper right panel of Figure 5-2, which CLIPPERS finds easily. Interesting cases where T<<1 because the pockets are too large are shown in the upper middle and the lower right panels of Figure 5-2. Low T scores for this reason are not necessarily bad: These pockets contain additional volume that could guide the design by medicinal chemists of more specific or higher affinity ligands by indicating areas where functional groups can be added. More generally, once having identified a ligand binding pocket, nearby pockets may be a good target for fragment based drug design [233; 234; 235; 236; 237; 238; 239] or interaction sites for added groups. Since CLIPPERS inventories all the pockets and places them in a tree, it facilitates such an approach. For example one may easily search for 'siblings' pockets in the tree: Ones which are joined by the lowest barriers forming natural routes across which the fragments would be joined. While CAST and SURFNET can sometimes identify these nearby pockets, only CLIPPERS identifies all such pockets and the saddle points joining them.

Comparing now the number, shape and size of pockets generated by the different methods, CAST typically generate tens of pockets per protein, SURFNET generates more, typically a hundred or so. CLIPPERS generates considerably more candidate pockets, usually several hundred per protein, and due to the hierarchical and inclusive way they are generated, smaller pockets are nested inside larger pockets, all the way down to the smallest dimple. Neither CAST nor SURFNET generates overlapping or nested pockets. Both methods also prune the number of possible pockets to focus on ones that hopefully include the site of interest. In SURFNET this is done by adjusting the parameters use in the sphere clustering method. In CAST this is done using the discrete flow technique to join the tetrahedra and decide

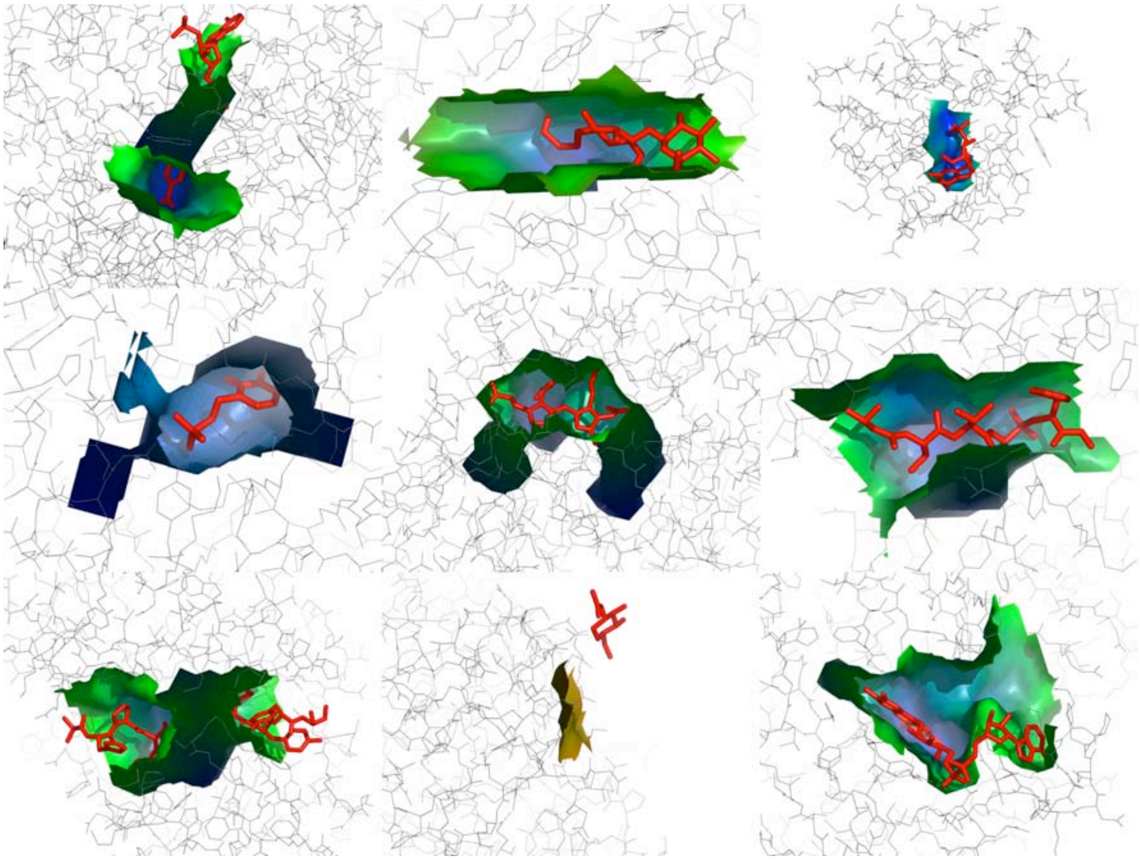


Figure 5-2 Pocket Finding Montage

9 example pockets found using CLIPPERS having the greatest Tanimoto score to ligand-neighboring residues. From left to right, top to bottom, the structures are PDB codes 1ADS, 1BYH, 1FUT, 1GPB, 1PDA, 1PPL, 1SMR, 1THG, 2CND. The protein is shown as grey lines, the ligand is shown in red sticks, the pocket is colored according to Travel Depth, figure created using PyMOL[64].

where pocket mouths lie. However, in each method the number of pockets is well correlated with the protein volume, shown for CLIPPERS in Figure 5-3a. In the SURFNET study, only the volume of the biggest and second biggest clefts were compared the protein volume. As another comparison to CAST, we show that the pocket areas and volumes correlate linearly with total protein area and volume, respectively as shown in Figures 5-3b and 5-3c.

Analyzing the 92 ligand binding pockets further, we find, as with CAST, that there is no correlation of protein size with binding site pocket size, as measured either by volume or surface area (Figures 5-4a and 5-4b). The mean of various statistics of these 92 pockets is as follows: Volume: 530 Å^3, Surface Area: 319 Å^2, mean Travel Depth: 12.8 Å, maximum Travel Depth: 17.2 Å, height: 7.2 Å, mean height: 2.8 Å, mean curvature: 5 degrees, principal dimensions: 16.8Å, 11.6Å, 7.1Å, fraction apolar surface area, 0.31: fraction negative surface area: 0.25, fraction positive surface area: 0.44.

Analyzing the mouth statistics in CLIPPERS, there is only one cavity in the set of 92, 83 pockets have single mouths, 5 have 2 mouths, 1 has 3 and 2 have 4. The mean mouth area is 147.5 Å^2 and the mean mouth longest dimension is 14.5 Å. The relationship between mouth number and pocket volume is shown in Figure 5-4c, as in CAST there is a slight correlation with mouth number and volume. The relationship between mouth diameter and mouth area is shown in Figure 5-4d, a line representing a perfect circle is shown for reference. Most mouths show some

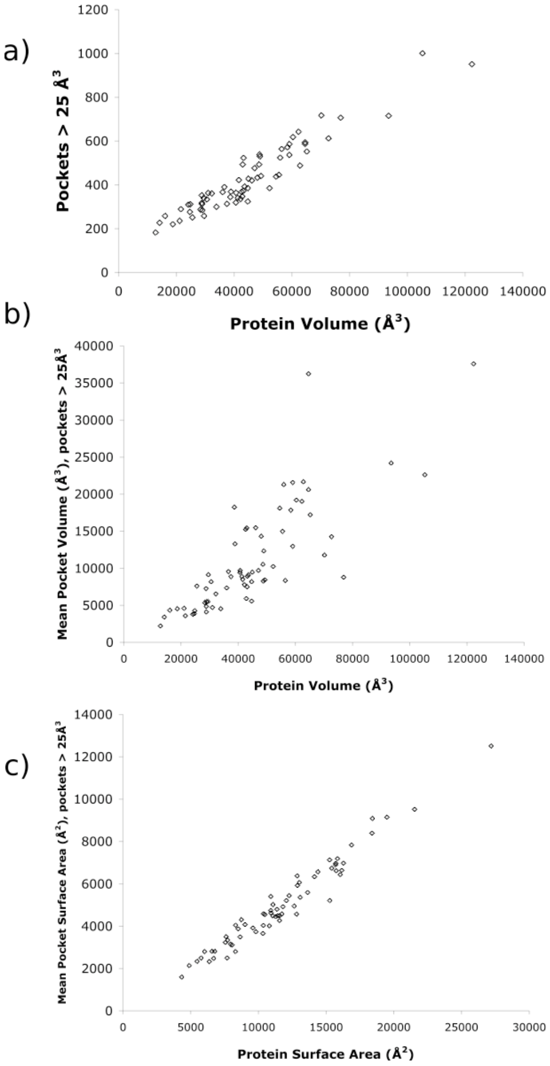


Figure 5-3 Pocket Finding Comparison

67 protein structures [71] analyzed using CLIPPERS. a) The protein volume compared with the total number of pockets with volume greater than 25 Å^3. b) Protein volume compared to mean pocket volume for pockets with volume greater than 25 Å^3. c) Protein surface area compared to mean pocket surface area for pockets with volume greater than 25 Å^3.

deviation from this, many mouths tend to be longer in one dimension than would be expected of perfectly circular mouths, since our mouths are not constrained to be bottlenecks as in CAST. This makes sense as mouths of grooves or clefts would by nature be very elongated. The mouth diameter is measured from point to point and not necessarily along the Travel Depth isosurface representing the mouth, this explains the few mouths with diameters smaller than possible for two dimensional circles.

A major feature of the CLIPPERS program is improved visualization of pockets with PyMOL[64] using customized python scripts. Once pockets have been inventoried and the resulting pocket data file loaded, each pocket surface can be displayed and colored individually. The default coloring is by Travel Depth, but other coloring schemes include pocket size, curvature, electrostatic potential and polarity. Another feature of CLIPPERS is that the lining atoms can be easily displayed. Several examples are shown in a montage in Figure 5-2.

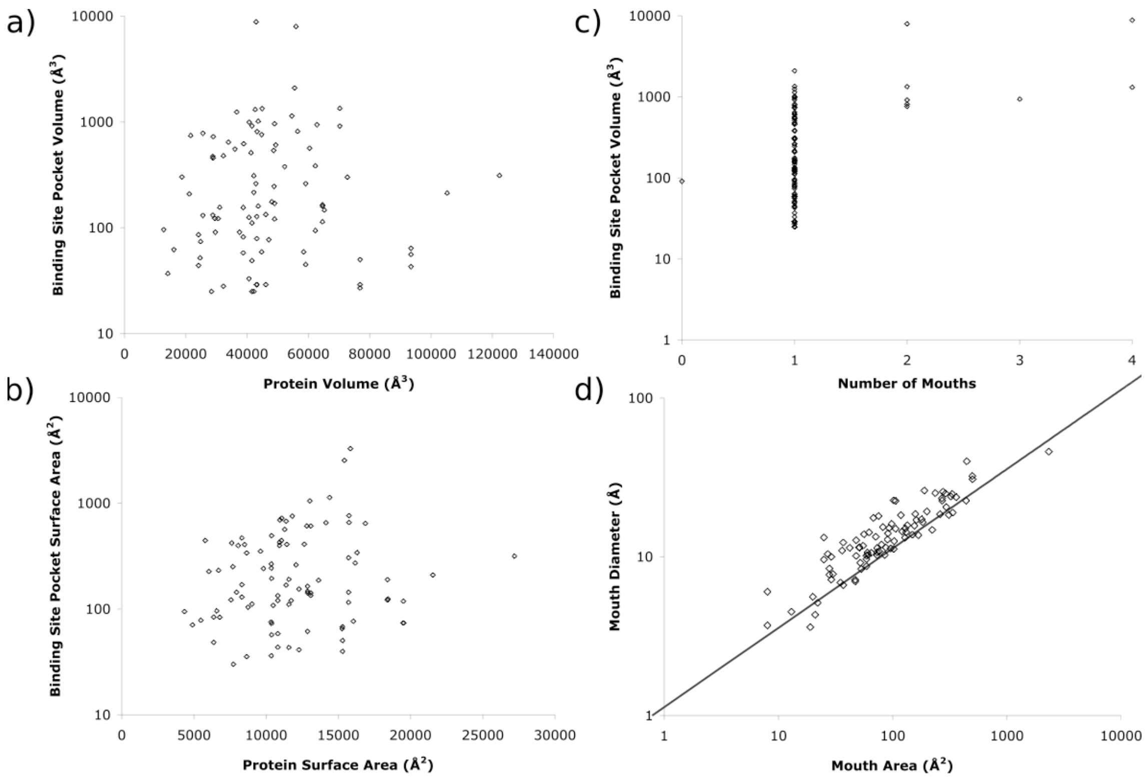


Figure 5-4 Pocket Finding Comparison – Binding Sites and Mouths

92 binding sites in 67 protein structures [71] analyzed using CLIPPERS. a) Protein volume compared to binding site volume (on log scale). b) Protein surface area compared to binding site surface area (on log scale). c) Number of mouths compared to the binding site volume (on log scale). d) Mouth area compared to mouth diameter (on log-log scale). The line corresponds to perfect circles.

Adenylate Kinase Transition Pathway

Adenylate kinase undergoes a significant conformational transition between the open inactive form PDB code 4AKE [240] and the closed active form Pdb code 1AKE [241]. This transition has been modeled by examining various crystal structures at the endpoints and in the middle of the transition pathway[242; 243], or by more extensive experiments [189; 244]. We generated a full transition from the closed to open forms using Climber, a morphing method that takes into account the energy of each structure when determining the step size to the next structure [245]. 82 were generated along the pathway and analyzed. The purpose of generating this transition pathway was twofold. First, to show how CLIPPERS can be used to track and examine pocket shape changes due to conformational changes. Second, to test the objectivity of the pocket-pocket distance function. Adjacent pockets in the pathway should have smaller separations in shape space than pockets further apart in the transition pathway.

To select the initial CLIPPER pocket series a set of 41 lining residues around the active site pocket was chosen, and the iterative Tanimoto overlap/swapping procedure described in Methods was used to pick a single pocket from each of the 82 structures. The resulting pairwise distance matrix was computed for this set of 82 pockets. We then used just this distance matrix, without reference to the known conformational sequence, to construct the minimum spanning lines of these pockets, i.e. the pocket sequence that minimized the total neighbor neighbor distance. We

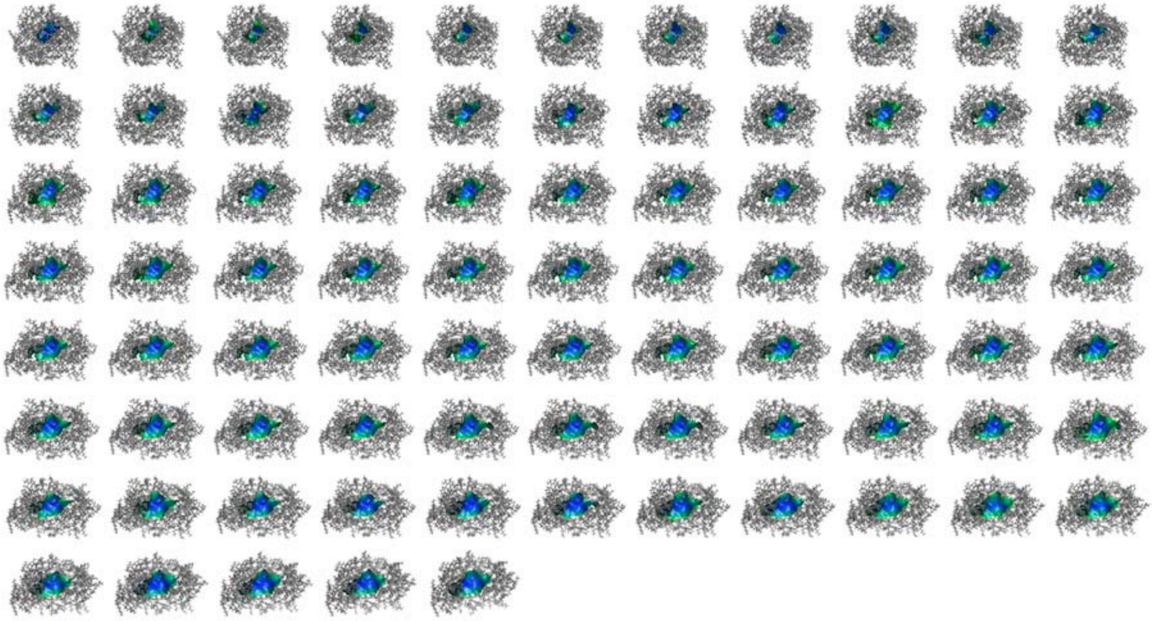


Figure 5-5 Adenylate Kinase Transition Visualization

The steps of the transition pathway shown from left to right, top to bottom adenylate kinase changes conformation from closed to open. The pockets found with CLIPPERS are visualized with Travel Depth.

then compared this reconstructed sequence with the actual sequence through the transition pathway.

We varied the descriptors included in the distance function and the distance metric (Manhattan or Euclidean) to determine which gave us the best reconstructed pathway. The Manhattan metric provided the best results, along with the following 11 descriptors: 1) surface area 2)volume 3) height 4) mean curvature 5) mouths 6) longest dimension 7) middle dimension 8) short dimension 9) area of biggest mouth 10) diameter of biggest mouth 11) mean height. The reconstructed ordering of the minimum spanning line had a Spearman Rank Correlation Coefficient of 0.999 with the actual ordering, indicating almost perfect ordering. This is excellent considering the degree of similarity of many pockets to each other in the open form. These 11 descriptors and the Manhattan metric were used for all further pocket-pocket distance comparisons.

Using the advanced pocket selection criteria that involve iterative swapping of pockets that have good residue overlap led to a transition pathway of pockets that was visually smooth and plausible, as shown in Figure 5-5. With this sequence many useful pocket properties can now be tracked smoothly throughout the entire transition pathway, as shown in Figure 5-6.

Finally the heatmap of the matrix of pocket-pocket distances across the entire transition pathway was computed (Figure 5-7). This representation confirms that adjacent pockets (near the diagonal) have low distances and pockets far away in the pathway have high distances. One interesting observation is that the open pockets

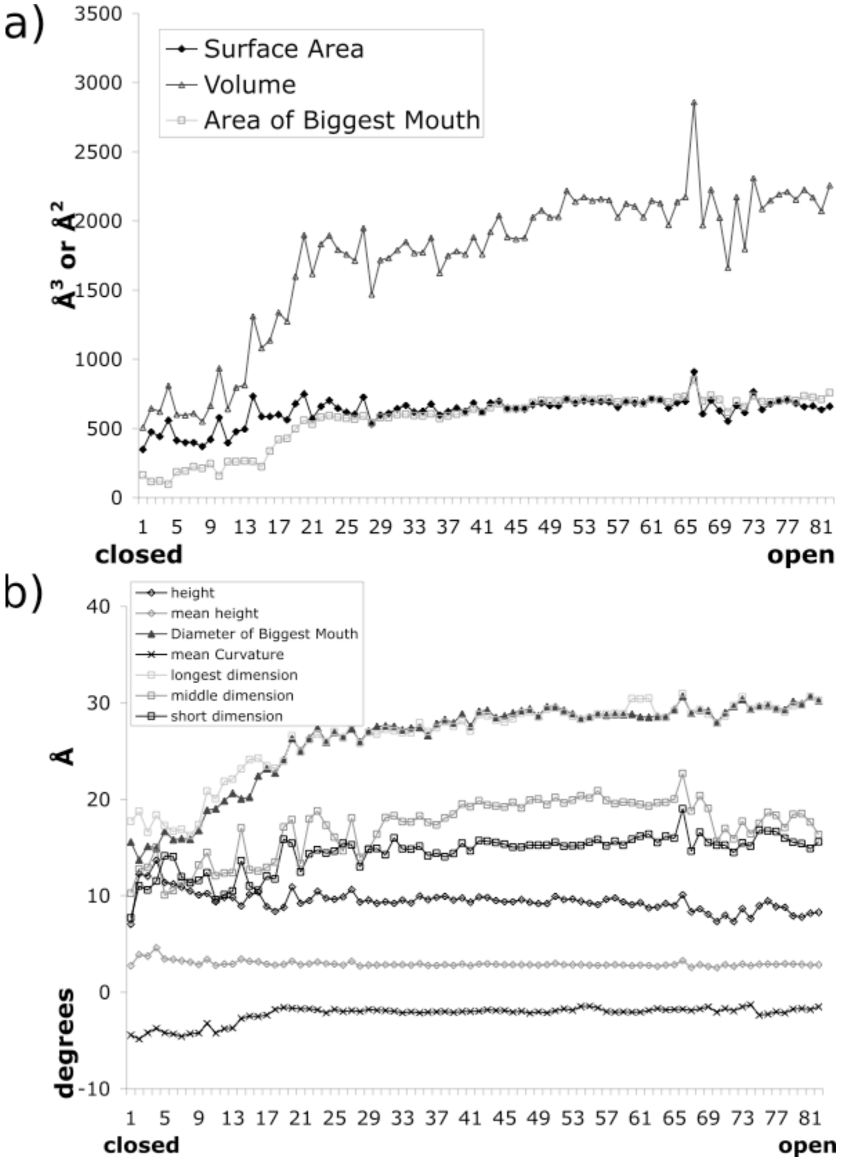


Figure 5-6 Adenylate Kinase Transition Properties

Properties of the binding pocket tracked over the transition pathway between the closed and open adenylate kinase structures. a) Volume, Surface Area, and Area of Biggest Mouth. b) Height, Mean height, Diameter of Biggest Mouth, principal dimensions, all in Å, and mean Curvature in degrees.

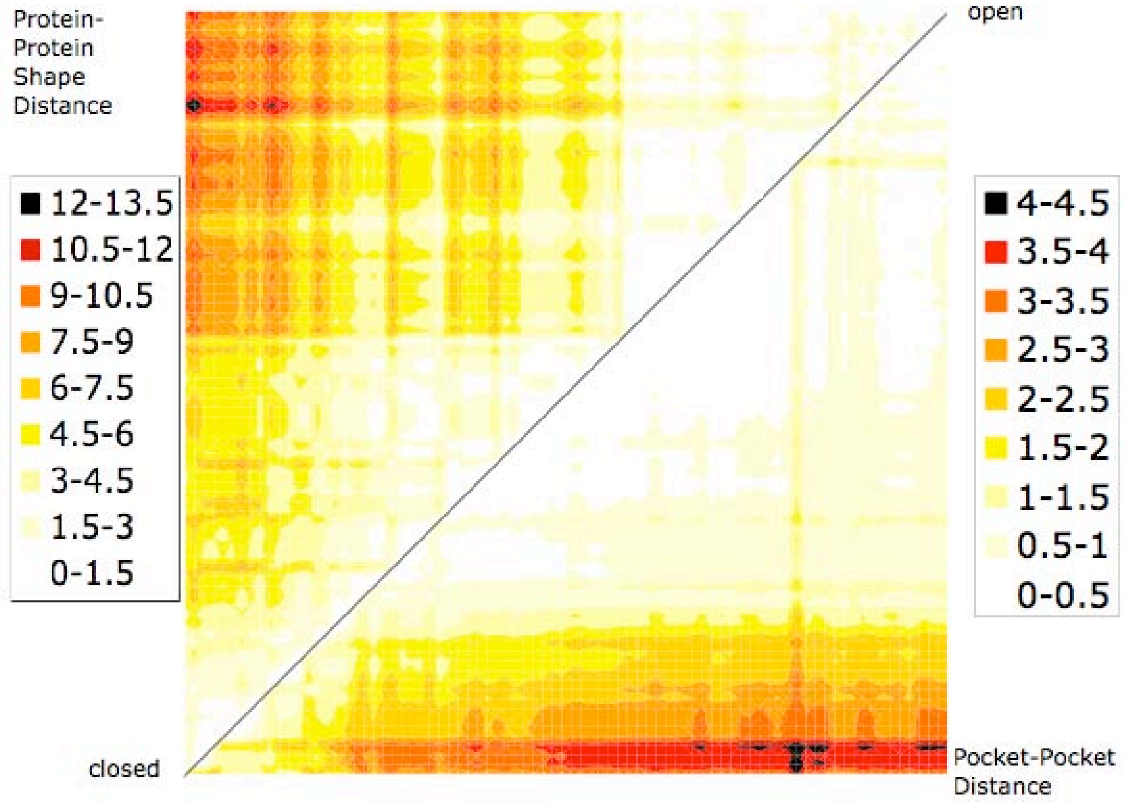


Figure 5-7 Adenylate Kinase Heatmap

The differences between all 82 structures along the transition pathway. Upper left half: root pocket-pocket distance. Lower right half: binding site pocket-pocket distance. Note that the scale for the two comparisons is different.

are more similar to each other than the closed and intermediate pockets, indicating that for a given overall structural change, the pocket shape change is more rapid as the protein approaches the closed form.

β-lactamase

β-lactamase is the enzyme responsible for bacterial resistance to penicillin and newer classes of antibiotics like cephalosporins. As such, the protein is under selective pressure due to the many new antibiotics used and it is of clinical and medicinal interest. Many mutants have been isolated from patients with resistant bacterial infections, and the structure of many of these determined. Thus β-lactamase is a good example of an enzyme whose active site has been well studied, and where high resolution structures of many active site variants are known. These include the wild-type [246], structures that have mutations conferring activity against cephalosporins [247], structures that have stabilizing mutations [247; 248], a structure with active site mutations that should destroy activity but do not due to sidechain and water rearrangements [249], structures bound to different inhibitors [250], and structures with inhibitor resistant mutations [251; 252]. Not all the mutations are in the active site, for instance many of the stabilizing mutations are far from the active site.

Table 5-1 β-lactamase Structure Comparison

PDB Code	Mutation(s)	Ligand[a]	Reference	Pocket Distance to WT Pocket	Volume (Å^3)	Surface Area (Å^2)
1XPB	WT		[246]		561	1013
1ESU	S235A		[246]	0.6	519	971
1JWP	M182T		[247]	1.0	475	926
1JWV	G238A	Yes	[247]	1.3	394	760
1JWZ	E104K/R164S/ M182T	Yes	[247]	2.3	404	695
1NYY	M182T	Yes	[250]	0.9	444	810
1NYM	M182T	Yes	[250]	1.7	366	592
1NY0	M182T	Yes	[250]	1.2	407	718

1NXY	M182T	Yes	[250]	1.6	398	695
3CMZ	L201P		[248]	0.9	468	904
1YT4	S130G		[249]	1.0	432	854
1LHY	R244S		[251]	1.4	445	857
1LI0	M69I/M182T		[251]	1.4	386	764
1LI9	M69V		[251]	1.0	436	848
1CK3	N276D		[252]	1.0	423	786

[a] Sulfate, Phosphate, Potassium and Bicarbonate Ions not included

15 structures were analyzed as summarized in Table 5-1. Each structure was downloaded, water and ligands removed, and a surface was created with a probe radius of 1.2Å. 44 residues were chosen to represent the active site, based on proximity to one of the ligands. These residues, numbers 69, 70, 72, 73, 103-105, 107, 127-130, 132, 166-171, 214-220, 234-238, 240-244, 268-273, 275, 276 were used to find the active site pocket with the Tanimoto overlap method. Pocket descriptor means and standard deviations were calculated from this data set.

With the refined list of active site pockets, all pocket-pocket distances can be examined. The pocket distance to the wild-type pocket is shown in Table 5-1. Ten of the structures have a mutual pocket-pocket distance of less than 1.1 (which is very low), these are typically mutations not in the active site or mutations or bound ligands that do not affect the overall shape of the active site. However, very few (six) of these pocket-pocket distances are less than 0.5. To put this figure of 0.5 in perspective, if the choice of pockets is refined simply by choosing the lowest distance regardless of whether the pocket is a ligand binding pocket, the resulting distance is 0.27, giving a rough lower bound to pocket-pocket distances. So a value of 0.5 indicates that each of the structures examined is somewhat different in a small but significant way.

The five structures that show extreme variation fromm these ten (and sometimes with each other) are PDB codes 1JWZ, 1ESU, 1XPB, 1LI9 and 1NYM. 1JWZ is a triple mutant with a bound ligand, one mutation is stabilizing the other mutations increase activity against cephalosporins, this pocket was observed by hand to be bigger in the original report [247]. Interestingly this is the only pocket that has 2 mouths and a much smaller mouth diameter of 17Å. 1XPB and 1ESU are the wildtype protein and a protein with a mutation that slows down activity against cephalosporins. If the refinement method is used on the original set of pockets found from residue overlap, these structures join the group with a low pocket-pocket distance to each other, indicating there are similar pockets in the tree with slightly rearranged lining residues. 1NYM is a structure that has the common M182T stabilizing mutation but is bound to a different inhibitor which mimics a transition states and is slightly smaller than the other pockets found. Two structures 1NXY and 1NYY have a pocket-pocket distance of 1.06, without this distance the highest pocket-pocket distance in the mutually similar set of ten is 0.93. 1LI9 has a mutation that is inhibitor resistant that makes the pocket slightly larger. Again these join the large group of mutually similar pockets if the refinement method is used.

Also of interest is 1YT4, an inhibitor resistant mutant that has a lot of rearrangements in the active site resulting in a differently placed but similarly shaped pocket. It is likely a technique based on residue motifs would not identify this pocket as similar due to the extensive rearrangements, despite having the same shape and function. The volume and surface area for all the active site pockets is shown in Table 5-1. As many of the structures have the stabilizing M182T mutation, the structure of just that mutation, 1JWV, along with the structure 1YT4 with the rearranged but similarly shaped active site, and the very different active site, 1JWZ, are all shown in Figure 5-8, along with their pocket-pocket distances. Note the 'failure' of pure pocket shape distance to discriminate the changes that take place in the 1YT4 active site, however the similarity is actually a success, even though the

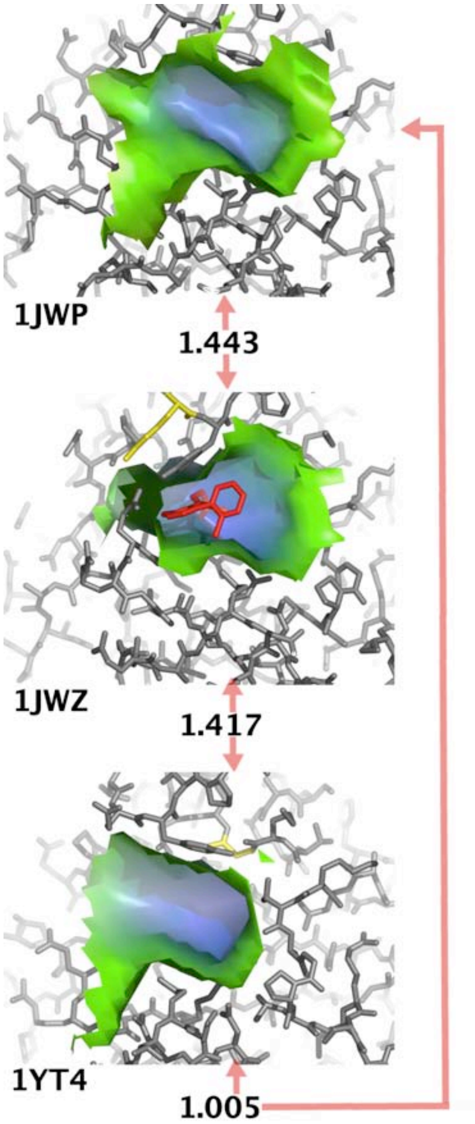


Figure 5-8 β-lactamase comparisons

Shown are three β-lactamase pockets colored by Travel Depth. The protein is colored grey, mutations from wild-type are colored yellow, and the ligand is colored red. The pocket-pocket distance between each pair of pockets is shown.

residues are very different the pocket shape and enzymatic activity is still very similar and is identified as such.

Enzyme Pocket Shape

To test the ability of our shape comparison to discern differences between active site pockets of proteins we used a benchmark dataset used to train a geometric hashing comparison algorithm based on atoms in the active site [214]. This data set contains 79 proteins from 13 diverse protein families. Although the activity and enzyme classification of these proteins within a family are identical, it is not necessarily true that binding pocket shapes within one class will be similar. In the original study describing this dataset, it was possible to cluster these binding sites into the correct classifications, but knowledge of the binding location was used. In contrast, in the test here of CLIPPERS, no prior information about the binding site was used in the clustering. Instead, each family was examined in turn to see if pocket shapes cluster together. Then these clusters were examined to see if they corresponded to the ligand binding sites.

First protein structures were downloaded, waters and ligands removed, and nonstandard amino acids preserved. Then CLIPPERS was run on each protein to inventory the pockets. To compute shape descriptor Z scores the means and standard deviations of the descriptors taken from the 67 proteins in the SURFNET/CAST dataset. For clustering, the penalty score given by Eqn. 3 with α=1 was used. Again no sequence or structural alignment of pockets was necessary. Clustering of the complete pocket trees of two or more proteins in a family proceeds by comparing pair wise pocket distances and connecting pocket pairs with an ‘edge’ if their similarity rises above some threshold. As each edge is added to the clustering, the residue overlap (ignoring residue order) of each distinct cluster was computed, along with how many structures have a pocket in that cluster. We search this output for clusters that have pockets representing all structures and that have the highest residue overlap score otherwise. Each pocket in the cluster is then examined to see if it corresponds to the ligand binding site. The results are summarized in Table 5-2. The total number of connections used up to the point where that cluster is created is also reported.

Eight of the thirteen families are complete successes, the cluster with at least one pocket from each structure with the highest residue overlap contains pockets representing each individual binding site. Though the residue overlap scores may not seem high, considering that mutations and size variation among pockets will affect this score, they are reasonable in the successful cases. In these cases we presume the shape and enzymatic activity are linked and note that the relatively simple scoring system of finding the cluster with at least one pocket from each structure with the highest residue overlap is sufficient to identify the binding sites for all such structures.

The cases where this simple scoring scheme fails to identify a cluster of active site pockets were examined further. In the set of ten serine/threonine kinases, no cluster with a nonzero overlap score containing pockets from all ten structures existed, and the highest scoring cluster with nine structures represented was not the binding site. The highest scoring cluster with four structures represented does indeed contain the

Table 5-2 Enzyme Shape Clustering

Enzyme Name	Total Connections	Cluster Size	Overlap	Structures in Cluster	Found Ligands	Total Structures
Aldose reductase	1823	130	0.368	8	8	8
Isocitrate dehydrogenase	3441	16	0.5	7	7	7
p-Hydroxybenzoate hydroxylase	640	15	0.73	7	7	7
Kinases (serine/threonine)	2859	51	0.019	9	0	10
Kinases (tyrosine)	166	2	0.5	2	1	2
Thymidylate kinase	6572	114	0.13	11	11	11
Subtilisin	243	2	0.763	2	2	2
Acid protease	694	26	0.725	7	7	7
Carbonic anhydrase	3140	16	0.444	6	6	6

Methionine gamma-lyase	6	2	1	2	0	2
D-Xylose isomerase	3801	32	0.115	8	0	8
Phosphoglycerate mutase	4990	31	0.507	4	4	4
D-Glutamate ligase MurD	166	11	0.917	5	0	5

binding sites from those structures, and has a residue overlap of 0.487. Further examination of the structures shows that the ones found represent an open conformation of the binding site, leading to a wide-mouthed but similarly shaped pocket in those four structures. Other structures in the set of ten for this enzyme class have a more closed conformation of the binding site, leading to a very different pocket with a much smaller mouth. So these would not show up in a clustering scheme based on shape, even though they do cluster together when residue type, position and known binding site location are the basis of the clustering [214].

The class of tyrosine kinases is represented by only two structures in this dataset. The highest overlap cluster found only contains one binding site, clustered with a similarly shaped cleft in the other structure. As there are only two structures in this cluster, these small clusters of just two pockets are very common and drown out possible clusters where many similar pockets all cluster together that have lower residue overlap due to size differences. Also, the binding site shapes of these two structures are somewhat different, while both bind in a cleft with an open mouth, one structure has two very deep lobes that extend beyond the volume taken up by the ligand, the other structure has much less volume below the ligand and no lobes. So, while CLIPPERS and the simple scoring scheme fail to identify the binding site here, this is not unreasonable as the binding site is not similarly shaped.

Methionine gamma-lyase has only two structures in this dataset, a cluster is found containing perfect residue overlap between 8 residues, however this is not the binding site cluster. D-xylose isomerase has many more structures but again suffers a similar result, the highest residue overlap cluster is not the binding site. Both these classes share some features: they have multiple binding sites per structure, their ligands are small, their binding sites are very deep and limited to just the volume near the ligand. In the methionine gamma-lyase structures, the probe radius of 1.2Å used appears to be a bit small and some of the ligand is inside the surface, this is possible where the surrounding protein is packed very closely to many parts of the ligand. Having multiple binding sites per structure, and having those binding sites be very small bottleneck pockets means they will be penalized highly by our redundancy clustering scheme.

The class of D-Glutamate ligase MurD contains five structures and the cluster with the most overlap is a conserved shallow dimple on the surface. The cluster ranked 3rd by residue overlap contains the five correct binding site pockets in a cluster of size 54 but with a low residue overlap of 0.211. Since there are 54 pockets in this cluster, and they are of variable size, the union of their residue counts is from the largest pocket while the intersection of the residue counts is from the smallest pocket, accounting for the very low overlap score.

Overall, while other methods can completely cluster these classes correctly [214], they have prior knowledge of the binding site location. Without binding site location, CLIPPERS correctly clusters the shapes of about two third of the classes. The other classes present a challenge for any shape based comparison. While binding and functional site location is not the major motivation for developing CLIPPERS, we note that the successes here show promise that additional methods or a better clustering and scoring system could prove useful. Regardless, similarly shaped binding sites can be identified using the clustering and scoring system, a useful test of the pocket similarity and redundancy formulas.

Protein Tyrosine Phosphatome

Protein tyrosine phosphatases are involved in many functions by specifically controlling dephosphorylation of their peptide or protein substrates. They share a conserved fold but have diverse shapes surface properties even across just the human forms. Recent structural genomics work in addition to previous crystal structures have presented an excellent opportunity to analyze the entire diverse class [253] [254]. Here we add to the previous analyses using CLIPPERS. We refer to the main site of dephosphorylation activity as the main active site. The photyrosine binding site sometimes found near the active site is referred to as the secondary site. When both sites are present in a single super-pocket found by CLIPPERS we will refer to this as the joint site. The conserved site putatively assumed to be involved with protein-protein interaction and regulation will be referred to as the distal site [254]. These sites are challenging since the interaction is between two proteins, so the binding face may have few deep or well-defined pockets.

32 crystal structures were downloaded that span the human phosphatome as detailed by Barr et. al. [253]. 12 of these were soluble proteins. In the other 20, the phosphatase domain was in a cytosolic region of a membrane protein. Waters were removed from each structure, non-standard residues were kept and then ligands and peptides were removed. Note that no structural alignment is necessary, nor is a sequence alignment used later to analyze residue conservation of pockets. This is

done purely on the residue types and counts lining each pocket. To compute shape descriptor Z scores the means and standard deviations of the descriptors taken from the 67 proteins in the SURFNET/CAST dataset.

Initially, the soluble and transmembrane classes were analyzed separately. The pockets from the 12 soluble structures were clustered using the redundancy scores to (hopefully) give preference to the more interesting and unique active site pockets. Several α values were used, from 0 (using only the pocket-pocket distances) to 10 (very highly weighting against redundant pockets). When using α=0, almost no active site pockets were identified in the output, when using α=10, pockets far from the active site that were unique to each structure were so heavily weighted that most other pockets were not present in the clustering output. At α=4, there were several active site or nearby active site pockets as well as many other pockets in the clustering output, so the α=4 clustering result was examined further. An arbitrary number of connections of 5000 was used to create the output graph shown in Figure 5-9, even at this scale most of the clusters have joined together but the structure can still be seen in the output graph. Among the output were many main active site pockets and many joint site pockets that also included the nearby PTP1B-like pocket, many distal site pockets, some small deep pockets, and many medium sized pockets varying from very flat to somewhat medium in depth. The medium size and flat to shallow bowl shaped pockets were clustered together in a large 'smear' which is also connected at this clustering threshold to the main site and joint site pockets.

The small deep pockets had almost very little sequence conservation (all much less than 0.5) and were located at different places in the structure. 5 of the non-

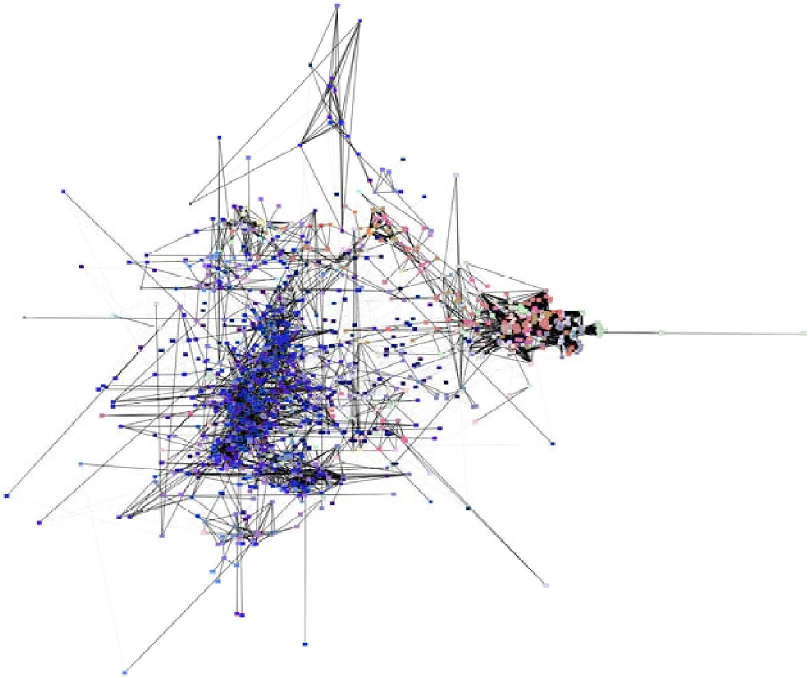


Figure 5-9 Protein Tyrosine Phosphatome Non-transmembrane Domain Comparison

12 structures and the resulting pockets found using CLIPPERS are shown. Only pockets within 25 Å^3 and 2000 Å^3 of volume are compared using α=4 to weight the uniqueness score and 5000 connections are shown in dark black lines. Thin grey lines connect the trees of pockets together. Each node has a border color unique to the structure. Each node is colored from blue to white to red according to the mean residue overlap over all connected nodes. This layout was created by aiSee [230].

transmembrane structures had these pockets, all were in different areas and only PDB code 1WCH [255] had such a pocket in the main active site. None of these pockets were very large, the largest in PDB code 2SHP [256] had a volume and surface area of about 200 Å^3 or Å^2 respectively, and a maximum and minimum travel depth around 20 and 11 Å. 2SHP also contained an additional similar pocket separate from this large one, the larger of these pockets are formed by the interface with the N-terminal SH2 domains, the smaller is in the PTP domain itself. These pockets, except for the one that is the closed active site found in 1WCH could potentially be sites of specific allosteric control, though the small size and depth would probably not favor natural binding partners or designed inhibitors. These pockets are clustered together at the top of Figure 5-9.

The elongated cluster running diagonally from just below the top to the right of Figure 5-9 contains main active site pockets, joint pockets and distal pockets. The upperleft most are the small main active sites alone. In the middle are many joint pockets. The big cluster at right contains both joint pockets and distal pockets. The joint pockets usually have a higher residue overlap and are therefore more red than blue. No pockets representing just the secondary site are in this output, as this is a shallow pocket and not always present, we assume it does not score well enough to show up at this arbitrary threshold. The distal pocket clusters near large sometimes oversized joint pockets (oversized meaning they contain more volume than just the main active site and secondary site), as they are both shallow and pockets with the same size can be found amongst the output for some structures. Again this distal pocket has been found before using computational techniques [254] though it appears to have not been confirmed experimentally.

Immediately to the right of the large cluster representing joint and distal pockets in Figure 5-9 is a small cluster of pockets with medium (around 0.5) residue conservation. These pockets are not near the distal pocket or the main active site, but appear to be a collection of pockets that are very long shallow trenches, the longest principal dimension of one pocket is almost 48Å.

The large unconserved, blue cluster dominating the left of Figure 5-9 is a collection of either small medium depth bowl-shaped pockets or small flat pockets, few of which show any residue conservation or spatial proximity. Again, by choosing higher α values this cluster will be smaller in the resulting graphs but at the expense of the interesting main active site and joint pockets.

The 20 receptor structures of protein tyrosine phosphatases [253] were examined and clustered in a similar manner to the non-transmembrane structures. Again, α=4 highlighted the most interesting set of output clusters. 10000 connections were used to create the output in Figure 5-10. Again we examine the more interesting clusters, though the large smear contains small pockets of either shallow dimples or close to flat shapes, and they do not have much residue overlap with their connections.

The cluster at the top left of Figure 5-10 is an interesting case. One of the two structures involved has pockets representing the main active site, 2A8B. The other structure, 2FH7, has pockets formed between two domains, it is striking that these pockets are both very similar in shape. The output cluster is colored white, indicating

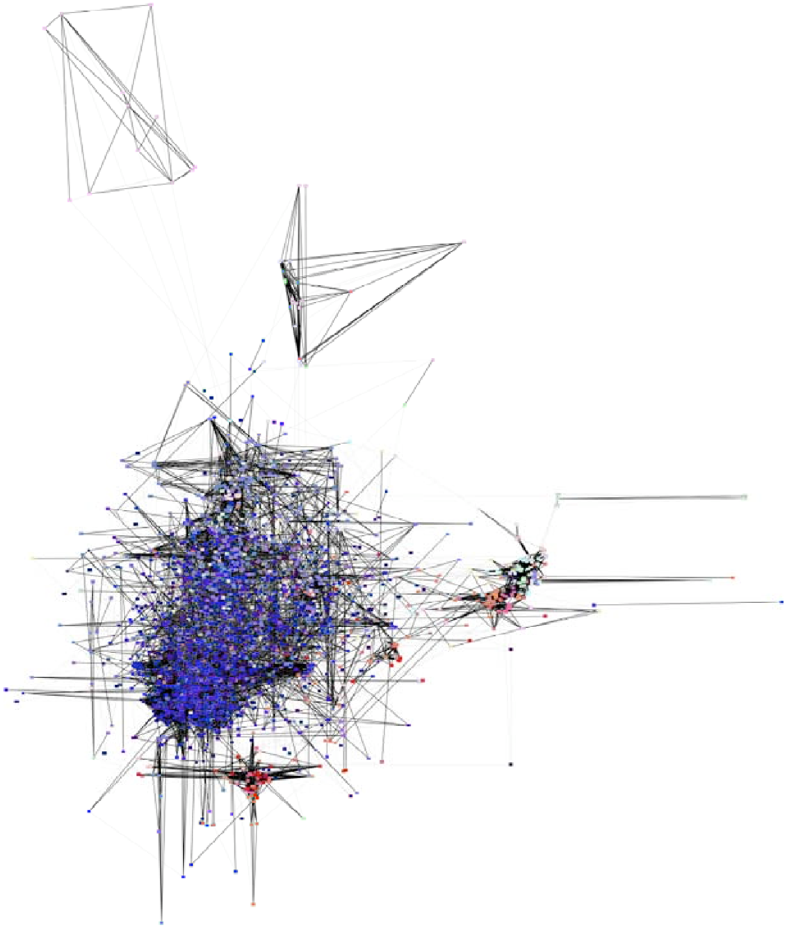


Figure 5-10 Protein Tyrosine Phosphatome Receptor Comparison

20 structures and the resulting pockets found using CLIPPERS are shown. Only pockets within 25 Å^3 and 2000 Å^3 of volume are compared using α=4 to weight the uniqueness score and 10000 connections are shown in dark black lines. Thin grey lines connect the trees of pockets together. Each node has a border color unique to the structure. Each node is colored from blue to white to red according to the mean residue overlap over all connected nodes. This layout was created by aiSee [230].

about 50% residue overlap. The next cluster examined is the one in the center but above the main cluster. This cluster is comprised solely of pockets formed at the domain boundaries of just three structures. There is again a fair but more variable of residue overlap in this cluster.

In the right of Figure 5-10 is a cluster with residue overlap from 0.5 and higher. Roughly half the cluster seems to be distal sites, the other half is main active sites with some extra nearby pockets, several structures are represented. The rough half containing main active sites have slightly higher residue overlaps overall. It may seem strange that these shapes cluster together as they sometimes did for the non-transmembrane structures as well, but it makes sense when the shapes are examined, as both contain one deep pocket and several connected shallow grooves.

At the bottom of Figure 5-10 is a cluster representing all distal pockets or pockets far from the main active site. These pockets are characterized by many shallow grooves and at most one deep depression. This cluster shows a lot of residue overlap and 6 structures are represented.

Clustering all thirty-two structures of protein tyrosine phosphatases at best led to a result where the ‘boring’ small flat dimpled pockets were clustered in one big cluster and some of the main active site or joint active site pockets were clustered together in another smaller cluster, further examination of this large clustering was not very interesting. Note that the thirty-two structures produced well over ten thousand pockets for over ten million possible connections.

As a closer examination of the ability of CLIPPERS to discern differences among related conformations and bound states, four structures of PTP-1b were examined. These 4 structures all have bound inhibitors that exploit both the main site and the nearby secondary pocket, though the structures are in different conformations. Two structures have a closed active site: PDB codes 1PTY [257] and 1Q1M [258]. Two structures have an open active site: 1NNY [259] and 1ONZ [260]. 31 residues were chosen that lie near any of the ligands in the structures and used to choose pockets for comparison. When the pockets chosen are compared, the open and closed states are discriminated according to their pocket-pocket distances. The differences between closed pockets is 0.41, between open pockets is 0.61. The difference between the two sets range from 1.05 to 1.24. If the refinement method is used to pick pockets that are as close to each other as possible, these numbers change in magnitude, dropping to 0.28 and 0.36 within the classes and ranging from 0.68 to 0.86 between the classes. When the refinement method is used, slightly larger pockets are chosen for all four structures, indicating that the smaller pockets are less alike than the larger pockets that contain them. Note that these differences between the sets are quite small but these pockets are very similar in shape, with the open sites having higher volumes, surface areas, bigger mouths and longer first principal dimensions than the closed sites, the other shape descriptors do not vary much. The fraction of apolar surface area of these pockets is between 0.3 and 0.4, confirming analysis that the PTP-1b site is not very druggable [199] and that the site is hard to search for using a formula based on finding hydrophobic concave regions [204]. The binding site pockets are shown in Figure 5-11.

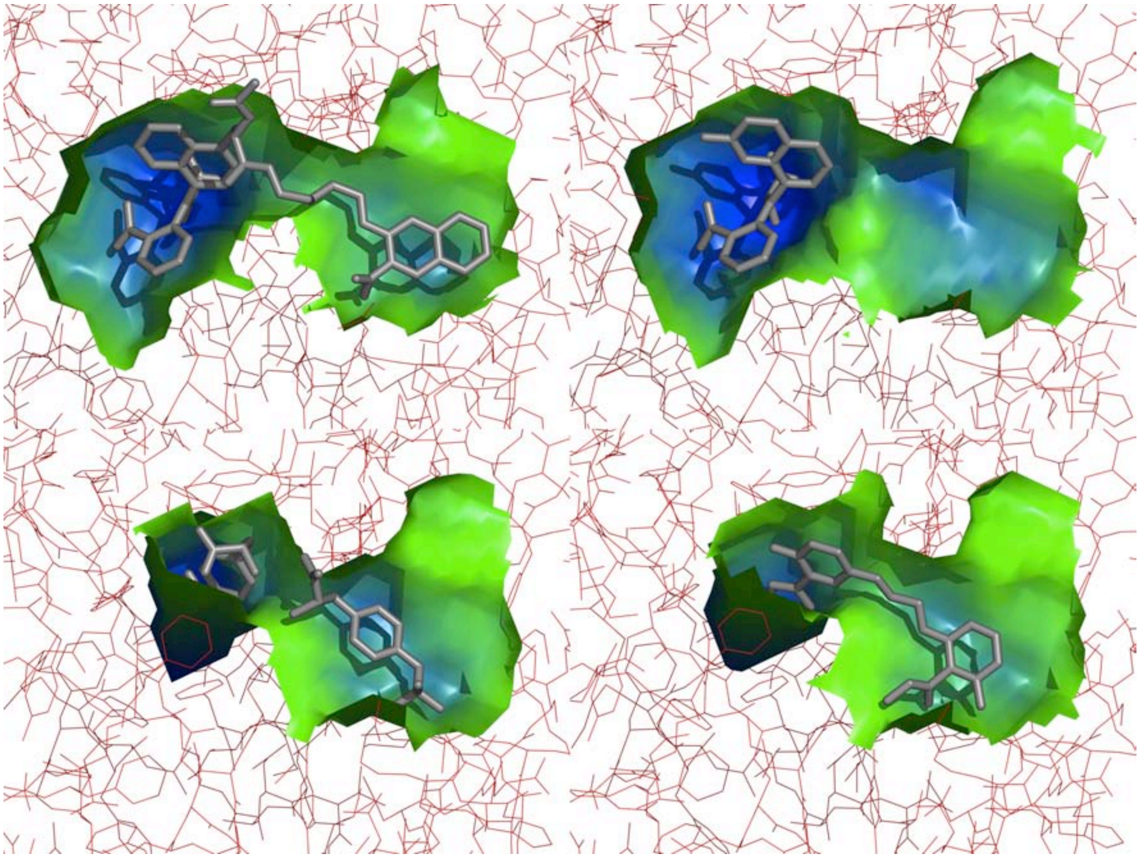


Figure 5-11 PTP-1b Binding Site Pockets

Pockets from the PTP-1b set of structures colored by Travel Depth, with the ligands shown in gray and the proteins shown in red. Top two panels: Open form. Bottom two panels: Closed form.

Future Work

We present here some of the many potential applications now possible with CLIPPERS, identifying all pockets, calculating shape properties, and comparing them. The shape framework and pocket hierarchy could be adapted to many others needs and applications, for instance to aid in functional site location and predciction [58], finding druggable binding sites [199; 204; 225] or especially druggable binding spots in protein-protein interfaces [261; 262; 263; 264; 265], finding sites amenable to fragment based drug design [233; 234; 235; 236; 237; 238; 239] or identifying transient pockets as proteins undergo motions [266].

The influence of pocket shape on chemical shape space and ligand shape is obviously important as well, and perhaps a complete classification of pocket shape will assist or provide guidance in these areas[267; 268; 269; 270; 271; 272; 273]. Also, allosteric site discovery [274; 275 204] is a very important application of finding potential binding sites. Additionally, cataloging protrusions of the protein surface could provide the positive shape to the negative shape provided here to search for protein-protein binding sites and partners. This could perhaps be done by using distance from the convex hull inwards and into the protein surface to catalog each protrusion in the same way CLIPPERS analyzes pockets.

Conclusion

CLIPPERS is a new computational technique capable of cataloging all the potential pockets on a protein surface, and this cataloging is done without any tunable parameters or user intervention. CLIPPERS passes three objective tests, first, it always finds a pocket with a reasonable and sometimes high residue Tanimoto score to bound ligands in a diverse test set of proteins,with a mean score of T=0.5. Second, it can reconstruct the ordering of pockets formed along a transition pathway purely from their pocket-pocket distances, as shown in the adenylate kinase transition pathway. Finally it gives lower pocket-pocket distances within groups of similar conformations than between them, as is shown with PTP1B.

Many applications need a list of pockets as a starting point for later analysis, some are presented here including tracking pockets through dynamic changes, comparing pockets across protein families or across different bound ligands. CLIPPERS provides excellent visualization and characterization of pocket shape through customized PyMOL scripts [64] and output of many shape features, including the difficult volume and mouth descriptors. CLIPPERS computes pocket-pocket distances without doing full pocket alignments of any kind and clusters pockets according to shape and uniqueness to visualize the many possible interacting pockets on a set of protein surfaces. The framework and approach is adaptable. For example it could be integrated with tools like multiple sequence alignment, so residue overlaps could be scored based on alignment profiles rather than residue identity scores with a single sequence. This would be expected to improve the pocket classification of multi-protein families.

Chapter 6

Conclusions and Future Work

There are many analyses possible now with the use of the shortest paths algorithm in the context of surfaces in structural biology, including many applications to protein-ligand binding, ion channel and pore examination, and the shape of proteins and their pockets, including changes in shape due to increased thermostability. These results have generated new hypotheses to be tested experimentally and are the building blocks of further computational techniques.

Summary of Results

CHUNNEL successfully identifies tunnels in many proteins, with interesting results. In the porin family, all tunnels are successfully found, the size of the choke points is correlated with the size of the molecules that can be transported, and the analysis of the residues involved indicates the as-expected polar residues line the pore, with arginine, tyrosine, glutamic acid, and proline significantly enriched near the choke point.

In analyzing the entire set of known transmembrane proteins, significant new facts were discovered. Many structures contain no putative physiological holes, those that span both membrane layers, but several membrane spanning segments do contain these tunnels. Many other kinds of tunnels exist, for instance those that exit within the bilayer or that transverse from two points interior to the bilayer. Of particular note was the discovery of a new class of tunnels which older algorithms like HOLE [40] could never discover; namely branched tunnels. These branches split off a putative physiological tunnel that spans both membrane bilayers and the branch exits with the bilayer. As no previous work could discover and classify bifurcated tunnels, CHUNNEL is the first to identify them. From their size and lining residue makeup, it would appear these tunnels are not involved in ion transport. They may be water-filled as they are lined primarily with tryptophan, known to prefer the polar head groups and water in membrane bilayers [276]. These tunnels may be involved in ion desolvation and resolvation for transport, as many ion channels conduct ions with fewer waters bound to them than in bulk solvent.

Turning to analyzing hyperthermostable and mesostable homologous pairs of proteins, there are several interesting results. Most importantly, it was shown that hyperthermostable proteins have significantly fewer in number and shallower pockets using Travel Depth analysis. Also hyperthermostable proteins bury more atoms further from the surface as a result of the Burial Depth analysis. This combined with a lack of significant change in buried atom packing, interatomic distances or convex hull volume leads to the conclusion that hyperthermostable proteins are not better packed, but instead that they are more spherical. Analysis using Wadell Sphericity [175], applied for the first time to proteins, supports this conclusion.

After correcting for overall differences in burial of atoms, it was shown by a residue-specific Burial Depth analysis that the charged residues of hyperthermostable proteins stay unburied significantly. The other residues are more buried, but not by any significant amount once the correction factor is employed, except for alanine which is much more buried in hyperthermostable proteins, consistent with previous analyses using surface area analyses, for instance that of Greaves and Warwicker [51].

Travel Depth confirms and quantifies many observations about macromolecular structure. First the relative depths of grooves of the canonical forms of DNA are shown to match the initial observations of the crystallographers solving the structures. Using a large database of protein-ligand co-crystal structures [33; 34], a large portion of binding sites (839 of 887) were shown to be quantitatively and significantly deeper than would be expected from a random binding site. The depths of the entire protein, a measure of number and depth of pockets, were shown to correlate strongly with protein size, also the Travel Depth of the binding sites correlated with the overall protein size, though to a lesser degree. This is interesting as the volume of binding sites does not correlate well with protein size using very different analysis [72] or CLIPPERS. The binding affinity of these protein-ligand complexes does not generally correlate with depth, as expected since binding affinity is more affected by other less global descriptors of the binding site.

CLIPPERS inventories and analyzes a nested set of pockets that completely cover a protein surface. The volume and surface area of these pockets correlates with the volume and surface area of the entire protein, confirming previous analyses [72] [86]. Binding site pocket size however, does not correlate with protein size according to CLIPPERS and previous work. In contrast however, where previous methods can fail to find many binding sites [72], CLIPPERS succeeds in finding a set of pockets that cover the entire surface. For any set of binding site residues, a pocket exists in the output with a good Tanimoto overlap of lining residues. CLIPPERS also provides the best visualization of pocket surfaces, and the first visualization of pocket depth, going hand in hand with the excellent protein surface visualization provided by Travel Depth.

On an objective test of a set of nearby pocket shapes constructed by a transition pathway of adenylate kinase, the pocket ordering can be reconstructed from shape similarity alone, providing an objective test of the pocket-pocket shape similarity distance, done with residue knowledge or alignments. Additionally, it is shown that the open state pockets of adenylate kinase are all very similar to each other and that while the structure must continue to transition to the completely open form, the pocket has already opened up to a likely inactive shape.

Analyzing a set of β-lactamase structures, or a set of protein tyrosine phosphatase structures, also led to good results in identifying conformational changes that have functional consequences. For instance a β-lactamase structure with many rearrangements in the pocket due to mutations still has a very similar shape as judged by CLIPPERS, this structure has retained activity despite these mutations and rearrangements. Structures of PTP1B can be discriminated between their open and closed conformations. Additionally, when sets of enzymes structures are examined, the active sites of functionally similar enzymes cluster together and can be picked out using the simple qualification of residue identity and count overlap. In eight of the thirteen classes examined this simple scoring scheme picks out the active site pockets for all the enzyme structures in the dataset [214], the other five cases, while similar in enzyme function, are not always similar in pocket shape.

Experimental Future Work

Many discoveries from CHUNNEL are bioinformatic in nature, for instance the enrichment of putatively functional amino acids at choke points in tunnels that span the membrane bilayer. Many structural and functional techniques have confirmed how some of these amino acids aid in ion or small molecule transport, but there are some amino acids whose exact contribution is still unknown. This is part of the already very active pursuit of understanding the structural and molecular nature of ion channels and pores.

The discovery of an entirely new class of tryptophan lined channels exiting in headgroup region that branch from membrane spanning channels opens the door to many experiments. Are these channels filled with water? Do these channels play a role in ion desolvation or solvation as most ions seem to be partially or completely desolvated when passing through the membrane? One system where these channels exist that is that of the inward rectifying potassium channel [277], and could present a good model system for experimental validation of this new theory.

In investigating the correlation of shape features and thermostability, many experiments are possible. One is to verify the effects of residue burial differences found, by making systematic mutations to design newly thermostable proteins. Also proteins could be adapted to a more spherical shape by searching for new backbone conformations and sidechain choices that stabilize the new backbone arrangement, while retaining the necessary active site shape and function.

The work here as well as much work with thermostable proteins suggests many additional experiments. The most prominent of these would be to obtain structures at more relevant temperatures, as many crystal structures are at cryonic temperatures. Some NMR experiments are conducted at 35 degrees C, still a far cry from the optimal conditions of some hyperthermophiles that can exist at temperatures above 80 degrees C. Having structures or ensembles of structures solved at these temperatures would aid greatly in understanding the nature of hyperthermostability.

Additionally, more experimental work correlating the structural and sequence features (including the shape features from this work) and the method of increased stability is suggested. Here method is meant be one of the following mechanisms: 1) a global increase in stability 2) a stability maximum that has been shifted up in temperature or 3) a higher heat capacity, in other words a wider range of temperatures at which the protein is stable. It is possible that the shape features correlate with only one or all of these methods, the available data now is not conclusive [191].

The work here on depth of pockets suggests several experiments. Are deeper pockets with similar shapes and conformations more or less hydrophobic? CLIPPERS could be adapted to find similar shape and residue lined pockets at various depths, but experimental techniques to determine water affinity at specific sites are still very difficult. Similar experiments could be done with small molecules, by finding small molecule binding sites that are similar in all respects except the absolute depth at which they bind. Again, comparisons between vastly different protein structures and experiments on them would necessitate many experiments to be positive of any findings.

Computational Future Work

The ability of CHUNNEL to find and analyze tunnels is a first step in any automated process to predict function from structure. A system that could predict the ion or other substrate would be very useful. This would again be just a first step of a procedure to find tunnels, predict what they transport, the kind of transport (transporters, pumps, channels), the conductance rate, and finally the effects of pH and voltage differential on these properties. Such a fully automated prediction scheme in likely many years away, especially as the number of membrane protein structures is still small [37].

Additionally CHUNNEL currently works well for single structures, but finding tunnels that are never completely open, i.e. transporters, is not possible with the current techniques. For this, finding tunnels in four dimensions, the fourth being time, will be necessary, and this severely complicates topological techniques that should augment any tunnel finding procedure.

Moving from a grid-based volume representation to a Voronoi-based one [45] for CHUNNEL may prove advantageous, as other methods have done[44; 46], though it is important to keep the topological features of the algorithm since these enable complete automation and reporting of all topologically distinct paths.

Finding more mesostable/hyperthermostable pairs of structures to analyze shapes (and other sequence and structural features) would lead to a better understanding. Also it would be good to rule out possibly mitigating factors like the multicellular nature of many of the mesophiles used in the current data set or the few hyperthermophilic and piezophilic organisms. This requires care in searching the available database of structures. Also, the shape features discussed here were not examined in psychrophiles, cold adapted organisms. They could be different in interesting ways, or the shapes may not change significatly, as found in the moderate thermophile against mesophile case.

Shape features like curvature [87] or roughness were not examined but could be significantly different between the various protein types. Also, the depth and shape of the active sites in the mesostable/hyperthermostable pairs could be examined with CLIPPERS. Finally, the flexibility of the structures could be examined, or the multiple conformations that may be present at high temperature, as suggested by some molecular dynamics simulations [188].

CLIPPERS has opened many new doors: identifying all pockets, calculating shape properties, and comparing them. The shape framework and pocket hierarchy could be adapted to many others needs and applications, for instance to aid in functional site location and predciction [58], finding drugable binding sites [199; 204; 225] or especially drugable binding spots in protein-protein interfaces [261; 262; 263; 264; 265], finding sites amenable to fragment based drug design [233; 234; 235; 236; 237; 238; 239] or identifying transient pockets as proteins undergo motions [266].

The influence of pocket shape on chemical shape space and ligand shape is obviously important as well, and perhaps a complete classification of pocket shape will assist or provide guidance in these areas[267; 268; 269; 270; 271; 272; 273]. Also, allosteric site discovery [274; 275 204] is a very important application of finding potential binding sites. Additionally, cataloging protrusions of the protein surface could provide the positive shape to the negative shape provided here to search for protein-protein binding sites and partners. This might be done by using distance from the convex hull inwards and into the protein surface to catalog each protrusion in the same way CLIPPERS analyzes pockets.

Conclusion

In conclusion, the new algorithms and analyses enabled by Dijkstra's shortest paths algorithm[23] lead to many new discoveries in structural biology which were not previously possible. Applications include: protein shape changes due to increased thermostability, finding and examining ion channels and pores, the depth of binding sites and how this affects binding affinity, and finally finding and comparing the shapes of pockets. While analyzing the distances from these various surfaces to each other or into the molecule or solvent has led to many new and different applications from the original surface analyses [4], there remain many avenues for inquiry into macromolecular shape and biological function.

Bibliography

1. van der Waals, J. D. (1873). Over de Continuïteit van den Gas - en Vloeistoftoestand, Leyden University.
2. van der Waals, J. D. (1910). The Equations of State for Gases and Liquids, pp. 254-265. Nobel Lecture.
3. Bondi, A. (1964). van der Waals Volumes and Radii. *J. Phys. Chem.* 68, 441-451.
4. Richards, F. M. (1977). Areas, volumes, packing, and protein structure. *Ann. Rev. Biophys. Bioeng.* 6, 151–76.
5. Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J Mol Biol* 105, 1-12.
6. Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature* 277, 491-493.
7. Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221, 709-713.
8. Connolly, M. L. (1983). Analytical molecular surface calculation. *J Appl Cryst* 16, 548-558.
9. Gerstein, M. & Lynden-Bell, R. M. (1993). What is the Natural Boundary of a Protein in Solution? *J Mol Biol* 230, 641-650.
10. Akkiraju, N. & Edelsbrunner, H. (1996). Triangulating the surface of a molecule. *Disc. Appl. Math.* 71, 5-22.
11. Edelsbrunner, H. & Mücke, E. P. (1994). Three-dimensional alpha-shapes. *ACM Trans. Graph.* 13, 43-72.

12. Bhat, S. & Purisima, E. O. (2005). Molecular surface generation using a variable-radius solvent probe. *Proteins: Struct. Funct. Bioinf.* 62, 244-261.

13. Lewis, M. & Rees, D. C. (1985). Fractal surfaces of proteins. *Science* 230, 1163-5.

14. Levitt, M. & Park, B. H. (1993). Water: now you see it, now you don't. *Structure* 1, 223-226.

15. Raschke, T. M. (2006). Water structure and interactions with proteins surfaces. *Curr Opin Struct Biol* 16, 152-9.

16. 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. *Nuc. Acid. Res.* 28, 235-242.

17. Chakravarty, S. & Varadarajan, R. (1999). Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure* 7, 723-732.

18. Pintar, A., Carugo, O. & Pongor, S. (2003). Atom depth in protein structure and function. *Trends Biochem Sci* 28, 593-7.

19. Pintar, A., Carugo, O. & Pongor, S. (2003). DPX: for the analysis of the protein core. *Bioinformatics* 19, 313-314.

20. Pintar, A., Carugo, O. & Pongor, S. (2003). Atom Depth as a Descriptor of the Protein Interior. *Biophys J* 84, 2553-2561.

21. Varrazzo, D., Bernini, A., Spiga, O., Ciutti, A., Chiellini, S., Venditti, V., Bracci, L. & Niccolai, N. (2005). Three-dimensional computation of atom depth in complex molecular structures. *Bioinformatics* 21, 2856-60.

22. Yuan, Z. & Wang, Z.-X. (2007). Quantifying the relationship of protein burying depth and sequence *Proteins: Struct. Funct. Bioinf.* 70, 509-16.

23. Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik* 1, 269-271.

24. Nicholls, A., Sharp, K. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* 4, 281-296.

25. Barber, C., Dobkin, D. & Huhdanpaa, H. (1993). *The Quickhull Algorithm for Convex Hull*, Geometry Center Technical Report GCG53, Univ. of Minnesota, MN.

26. de Berg, M., van Kreveld, M., Overmars, M. & Schwarskopf, O. (2000). *Computational Geometry: Algorithms and Applications*. 2nd edit, Springer, Berlin.

27. O'Rourke, J. (1998). *Computational Geometry in C*. 2nd edit, Cambridge University Press.

28. Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2001). *Introduction to Algorithms. Second Edition*. 2nd edit, McGraw-Hill Higher Education.

29. Canny, J. & Reif, J. (1987). *IEEE Symp. Found. Comp. Sci.*

30. Dickerson, R. E., Drew, H. R., Conner, B. N., Kopka, M. L. & Pjura, P. E. (1983). Helix geometry and hydration in A-DNA, B-DNA, and Z-DNA. *Cold Spring Harb. Symp. Quant. Biol.* 47 Pt 1, 13-24.

31. Dickerson, R. E., Drew, H. R., Conner, B. N., Wing, R. M., Fratini, A. V. & Kopka, M. L. (1982). The Anatomy of A-, B-, and Z-DNA. *Science* 216, 475-485.

32. Wang, A. H.-J., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. H., van der Marel, G. & Rich, A. (1979). Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* 282, 680-686.

33. Wang, R., Fang, X., Lu, Y. & Wang, S. (2004). The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* 47, 2977-2980.

34. Wang, R., Fang, X., Lu, Y., Yang, C.-Y. & Wang, S. (2005). The PDBbind Database: Methodologies and updates. *J. Med. Chem.* 48, 4111-4119.

35. Coleman, R. G. & Sharp, K. A. (2006). Travel Depth, a New Shape Descriptor for Macromolecules: Application to Ligand Binding. *J Mol Biol* 362, 441-458.

36. Petrek, M., Otyepka, M., Banás, P., Kosinová, P., Koca, J. & Damborsky, J. (2006). CAVER: A New Tool to Explore Routes from Protein Clefts, Pockets and Cavities. *BMC Bioinformatics* 7, 1-9.

37. White, S. H. (2004). The progress of membrane protein structure determination. *Prot. Sci.* 13, 1948-9.

38. White, S. H. (2007). Membrane Proteins of Known Structure.

39. Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T. & MacKinnon, R. (1998). The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science* 280, 69-77.

40. Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. P. (1996). HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* 14, 354-360.

41. Smart, O. S., Goodfellow, J. M. & Wallace, B. A. (1993). The Pore Dimensions of Gramicidin A. *Biophys. J.* 65, 2455-2460.

42. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. (2006). OPM: orientations of proteins in membranes database. *Bioinformatics* 22, 623-5.

43. Coleman, R. G. & Sharp, K. A. (2009). Finding and Characterizing Tunnels in Macromolecules with Application to Ion Channels and Pores. *Biophys J* 96, 632-645.

44. Petrek, M., Kosinová, P., Koca, J. & Otyepka, M. (2007). MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure* 15, 1357-1363.

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

46. Yaffe, E., Fishelovitch, D., Wolfson, H. J., Halperin, D. & Nussinov, R. (2008). MolAxis: Efficient and Accurate Identification of Channels in Macromolecules. *Proteins: Struct. Funct. Bioinf.* 73, 72-86.

47. Atomi, H. (2005). Recent progress towards the application of hyperthermophiles and their enzymes *Curr Opin Chem Biol* 9, 166-173.

48. Vogt, G., Woell, S. & Argos, P. (1997). Protein thermal stability, hydrogen bonds, and ion pairs. *J Mol Biol* 269, 631-43.

49. Cambillau, C. & Claverie, J.-M. (2000). Structural and Genomics Correlates of Hyperthermostability. *J Biol Chem* 275, 32383-86.

50. Fukuchi, S. & Nishikawa, K. (2001). Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *J Mol Biol* 309, 835-43.

51. Greaves, R. B. & Warwicker, J. (2007). Mechanisms for stabilisation and the maintenance of solubility in proteins from thermophiles. *BMC Struct Biol* 7, 1-23.

52. Yano, J. K. & Poulos, T. L. (2003). New understandings of thermostable and peizostable enzymes. *Curr Opin Biotech* 14, 360-365.

53. Glyakina, A. V., Garbuzynskiy, S. O., Lobanov, M. Y. & Galzitskaya, O. V. (2007). Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. *Bioinformatics* 23, 2231-2238.

54. Lin, Y.-S. (2008). Using a strategy based on the concept of convergent evolution to identify residue substitutions responsible for thermal adaptation. *Proteins: Struct. Funct. Bioinf.* 73, 53-62.

55. Robinson-Rechavia, M., Alibésb, A. & Godzik, A. (2006). Contribution of Electrostatic Interactions, Compactness and Quaternary Structure to Protein Thermostability: Lessons from Structural Genomics of Thermotoga maritima. *J Mol Biol* 356, 547-557.

56. Coleman, R. G. & Sharp, K. A. (2009). Thermophilic protein structure adaptation examined with Burial Depth and Travel Depth [abstract]. *Biophys J* 96, 584a.

57. Coleman, R. G. & Sharp, K. A. (2009). Shape and Evolution of Thermostable Protein Structure. *Proteins: Struct. Funct. Bioinf.* X, X.

58. Campbell, S. J., Gold, N. D., Jackson, R. M. & Westhead, D. R. (2003). Ligand binding: functional site location, similarity and docking. *Curr. Opin. Struct. Biol.* 13, 389-395.

59. Coleman, R. G. & Sharp, K. A. (2009). Protein Pockets: Inventory, Shape and Comparison. *J Mol Biol* submitted.

60. Wang, A. H.-J., Fujii, S., van Boom, J. H. & Rich, A. (1983). Right-handed and left-handed double-helical DNA: structural studies. *Cold Spring Harb. Symp. Quant. Biol.* 47 Pt 1, 33-44.

61. Cheng, H. L., Dey, T. K., Edelsbrunner, H. & Sullivan, J. (2001). Dynamic skin triangulation. *Disc. Comp. Geom.* 25, 525-568.

62. Sridharan, S., Nicholls, A. & Honig, B. (1992). A new vertex algorithm to calculate solvent accessible surface areas. *Biophys. J.* 61, abstract 995.

63. Eppstein, D. (2002). Priority Dictionary.

64. DeLano, W. L. (2002). The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, California.

65. Hill Jr, F. S. (2001). *Computer Graphics Using Open GL*. 2nd edit, Prentice Hall, Upper Saddle River, New Jersey.

66. Pearlman, D. A., Case, D. A., Caldwell, J. C., Seibel, G. L., Singh, U. C., Weiner, P. & Kollman, P. A. (1991). AMBER 4.0. University of Californa San Francisco, San Francisco, California.

67. Drew, H. R., Wing, R. M., Takano, T., Broka, C., Tanaka, S., Itakura, K. & Dickerson, R. E. (1981). Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. USA* 78, 2179-2183.

68. Crawford, J. L., Kolpak, F. J., Wang, A. H., Quigley, G. J., van Boom, J. H., van der Marel, G. & Rich, A. (1980). The tetramer d(CpGpCpG) crystallizes as a left-handed double helix. *Proc. Natl. Acad. Sci. USA* 77, 4016-4020.

69. Buchbinder, J. L., Stephenson, R. C., Scanlan, T. S. & Fletterick, R. J. (1998). A comparison of the crystallographic structures of two catalytic antibodies with esterase activity. *J. Mol. Biol.* 282, 1033-1041.

70. Peters, K. P., Fauck, J. & Frommel, C. (1996). The Automatic Search for Ligand Binding Sites in Proteins of Known Three-dimensional Structure Using only Geometric Criteria. *J. Mol. Biol.* 256, 201-213.

71. Laskowski, R. A. (1995). SURFNET: A program for visualizing molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph.* 13, 323-330.

72. Liang, J., Edelsbrunner, H. & Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Prot. Sci.* 7, 1884-1897.

73. (1999). Origin 6.0. Microcal Software, Inc., Northampton, MA.

74. Kuntz, I. D., Chen, K., Sharp, K. & Kollman, P. (1999). The maximal affinity of ligands. *Proc. Natl. Acad. Sci. USA* 96, 9997-10002.

75. Misra, V. K. & Draper, D. E. (2000). On the role of magnesium ions in RNA stability. *Biopolymers* 48, 113-135.

76. Glaser, F., Morris, R. J., Najmanovich, R. J., Laskowski, R. A. & Thornton, J. M. (2006). A Method for Localizing Ligand Binding Pockets in Protein Structures. *Proteins: Struct. Funct. Bioinf.* 62, 479-288.

77. Ben-Shimon, A. & Eisenstein, M. (2005). Looking at Enzymes from the Inside out: The Proximity of Catalytic Residues to the Molecular Centroid can be used for Detection of Active Sites and Enzyme–Ligand Interfaces. *J. Mol. Biol.* 351, 309-326.

78. Davis, M. E. (1995). Molecular Elevation GRASP Macro.

79. Agarwal, P. K., Edelsbrunner, H., Harer, J. & Wang, Y. (2004). *Symp. Comp. Geo.*

80. Wang, Y., Agarwal, P. K., Brown, P., Edelsbrunner, H. & Rudolph, J. (2005). *Pac. Symp. Biocomp.*

81. Choi, J., Sellen, J. & Yap, C.-K. (1994). *Symp. Comp. Geo.*

82. Lozano-Pérez, T. & Wesley, M. A. (1979). An Algorithm for Planning Collision-Free Paths Among Polyhedral Obstacles. *Comm. of the ACM* 22, 560-570.

83. Mitchell, S. A. & Vavasis, S. A. (1992). *Symp. Comp. Geo.*

84. Sethian, J. A. (1999). *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. 2nd edit, Cambridge University Press, Cambridge.

85. Hugg, J., Rafalin, E., Seyboth, K. & Souvaine, D. L. (2006). An Experimental Study of Old and New Depth Measures. *ALENEX06, Springer-Verlag Lect. Notes Comp. Sci.*, 51-64.

86. Laskowski, R. A., Luscombe, N. M., Swindells, M. B. & Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Prot. Sci.* 5, 2438-2452.

87. Coleman, R. G., Burr, M. A., Souvaine, D. L. & Cheng, A. C. (2005). An intuitive approach to measuring protein surface curvature. *Proteins: Struct. Funct. Bioinf.* 61, 1068-1074.

88. Sharp, K. & Honig, B. (1990). Electrostatic Interactions in Macromolecules: Theory and Applications. *Ann. Rev. Biophys. Chem.* 19, 301-332.

89. Lichtarge, O. & Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* 12, 21-27.

90. Nayal, M. & Honig, B. (2006). On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins: Struct. Funct. Bioinf.* 63, 892-906.

91. Elcock, A. H. & McCammon, J. A. (1996). Evidence for Electrostatic Channeling in a Fusion Protein of Malate Dehydrogenase and Citrate Synthase. *Biochemistry* 35, 12652-12658.

92. Elcock, A. H. (2004). Molecular Simulations of Diffusion and Association in Multimacromolecular Systems. *Meth. Enzymology* 383, 166-198.

93. Bass, R. B., Strop, P., Barclay, M. & Rees, D. C. (2002). Crystal structure of Escherichia coli MscS, a voltage-modulated and mechanosensitive channel. *Science* 298, 1582-7.

94. Moarefi, I., Jeruzalmi, D., Turner, J., O'Donnell, M. & Kurlyan, J. (2000). Crystal structure of the DNA polymerase processivity factor of T4 bacteriophage. *J Mol Biol* 296, 1215-1223.

95. Voss, N. R., Gerstein, M., Steitz, T. A. & Moore, P. B. (2006). The Geometry of the Ribosomal Polypeptide Exit Tunnel. *J Mol Biol* 360, 893-906.

96. Roll-Mecak, A. & Vale, R. D. (2008). Structural basis of microtubule severing by the hereditary spastic paraplegia protein spastin. *Nature* 451, 363-367.

97. Taylor, T. C. & Andersson, I. (1997). The structure of the complex between rubisco and its natural substrate ribulose 1,5-bisphosphate. *J Mol Biol* 265, 432-44.

98. Gilson, M. K., Straatsma, T. P., McCammon, J. A., Ripoli, D. R., Faerman, C. H., Axelsen, P. H., Silman, I. & Sussman, J. L. (1994). Open "back door" in a

molecular dynamics simulation of acetylcholinesterase. *Science* 263, 1276-1278.

99. Murray, J. W. & Barber, J. (2007). Structural Characteristics of channels and pathways in photosystem II including the identification of an oxygen channel. *J Struct Biol* 159, 228-237.

100. Hankamer, B., Glaeser, R. & Stahlberg, H. (2007). Electron Crystallography of membrane proteins. *J Struct Biol* 160, 263-264.

101. Walian, P., Cross, T. A. & Jap, B. K. (2004). Structural genomics of membrane proteins. *Genome Biology* 5, 215.

102. Bansal, A. & Sankararamakrishnan, R. (2007). Homology modeling of major intrinsic proteins in rice, maize and Arabidopsis: comparative analysis of transmembrane helix association and aromatic/arginine selectivity filters. *BMC Struct Biol* 7.

103. Kleywegt, G. J. & Jones, T. A. (1994). Detection, Delineation, Measurement and Display of Cavities in Macromolecular Structures. *Acta Cryst D* 50, 178-185.

104. Damborsky, J., Petrek, M., Banás, P. & Otyepka, M. (2007). Identification of Tunnels in Proteins, Nucleic Acids, Inorganic Materials and Molecular Ensembles. *Biotechnology Journal* 2007, 62-67.

105. Feldman, J. & Singh, M. (2006). Bayesian estimation of the shape skeleton. *PNAS* 103, 18014-9.

106. Larn, L., Lee, S.-W. & Suen, C. Y. (1992). Thinning Methodologies-A Comprehensive Survey. *IEEE Trans. on Patt. Anal. and Mach. Intel.* 14, 869-85.

107. Reinders, F., Jacobson, M. E. D. & Post, F. H. (2000). Skeleton Graph Generation for Feature Shape Description. *Eurographics-IEEE TCVG Symp. on Visualization*, 73-82.

108. Foskey, M., Lin, M. C. & Manocha, D. (2003). Efficient Computation of A Simplified Medial Axis. *J of Comp and Info Sci in Engi* 3, 274-284.

109. Coleman, R. G. (2004). Finding Knotted and Linked Vorticity Lines in 3D Vector Fields. Master's Thesis, Tufts University.

110. Mardia, K. V. (1975). Statistics of Directional Data. *J Roy Stat Soc B (Method)* 37, 349-393.

111. Lundstrom, K. (2006). Structural genomics for membrane proteins. *Cell and Mol Life Sci* 63, 2597-2607.

112. Nikaido, H. (2003). Molecular Basis of Bacterial Outer Membrane Permeability. *Micro and Mol Biol Reviews* 67, 593-656.

113. Koebnik, R., Locher, K. P. & Van Gelder, P. (2000). Structure and function of bacterial outer membrane proteins: barrels in a nutshell *Mol Microbiol* 37, 239-253.

114. Weiss, M. S. & Schulz, G. E. (1992). Structure of porin refined at 1.8 Å resolution. *J Mol Biol* 227, 493-509.

115. Kreusch, A., Neubüser, A., Schiltz, E., Weckesser, J. & Schulz, G. E. (1994). Structure of the membrane channel porin from Rhodopseudomonas at 2.0 Å resolution. *Prot. Sci.* 3, 58-63.

116. Dutzler, R., Rummel, G., Albertí, S., Hernández-Allés, S., Phale, P., Rosenbusch, J., Benedí, V. & Schirmer, T. (1999). Crystal structure and

functional characterization of OmpK36, the osmoporin of Klebsiella pneumoniae. *Structure* 7, 425-434.

117. Zeth, K., Diederichs, K., Welte, W. & Engelhardt, H. (2000). Crystal structure of Omp32, the anion-selective porin from Comamonas acidovorans, in complex with a periplasmic peptide at 2.1 Å resolution. *Structure* 8, 981-992.

118. Zachariae, U., Klühspies, T., De, S., Engelhardt, H. & Zeth, K. (2006). High resolution crystal structures and molecular dynamics studies reveal substrate binding in the porin Omp32. *J Biol Chem* 281, 7413-7420.

119. Cowan, S. W., Schirmer, T., Rummel, G., Steiert, M., Ghosh, R., Pauptit, R. A., Jansonius, J. N. & Rosenbusch, J. (1992). Crystal structures explain functional properties of two E. coli porins. *Nature* 358, 727-733.

120. Baslé, A., Rummel, G., Storic, P., Rosenbusch, J. & Schirmer, T. (2006). Crystal Structure of osmoporin OmpC from E. coli at 2.0 Å. *J Mol Biol* 362, 933-942.

121. Subbarao, G. V. & van der Berg, B. (2006). Crystal Structure of the monomeric porin OmpG. *J Mol Biol* 360, 750-759.

122. Yildiz, O., Vinothkumar, K. R., Goswami, P. & Kühlbrandt, W. (2006). Structure of the monomeric outer-membrane porin OmpG in the open and closed confirmation. *EMBO J* 25, 3702-3713.

123. Meyer, J. E., Hofnung, M. & Schulz, G. E. (1997). Structure of maltoporin from Salmonella typhimurium ligated with a nitrophenyl-maltotrioside. *J Mol Biol* 266, 761-75.

124. Schirmer, T., Keller, T. A., Wang, Y. F. & Rosenbusch, J. (1995). Structural basis for sugar translocation through maltoporin channels at 3.1 Å resolution. *Science* 267, 512-514.

125. Forst, D., Welte, W., Wacker, T. & Diederichs, K. (1998). Structure of the sucrose-specific porin ScrY from Salmonella typhimurium and its complex with sucrose. *Nat Struct Biol* 5, 37-46.

126. Moraes, T. F., Bains, M., Hancock, R. E. & Strydnadka, N. C. (2007). An arginine ladder in OprP mediates phosphate-specific transfer across the outer membrane. *Nat Struct Mol Biol* 14, 85-87.

127. Hedfalk, K., Törnroth-Horsefield, S., Nyblom, M., Johanson, U., Kjellbom, P. & Neutze, R. (2006). Aquaporin gating. *Curr Opin Struct Biol* 16, 447-456.

128. Sui, H., Han, B. G., Lee, J. K., Walian, P. & Jap, B. K. (2001). Structural basis of water-specific transport through the AQP1 water channel. *Nature* 414, 872-8.

129. Lomize, A. L., Pogozheva, I. D., Lomize, M. A. & Mosberg, H. I. (2006). Positioning of proteins in membranes: a computational approach. *Prot. Sci.* 15, 1318-33.

130. Senes, A., Chadi, D. C., Law, P. B., Walters, R. F. S., Nanda, V. & DeGrado, W. F. (2007). Ez, a Depth-dependent Potential for Assessing the Energies of Insertion of Amino Acid Side-chains into Membranes: Derivation and Applications to Determining the Orientation of Transmembrane and Interfacial Helices. *J Mol Biol* 366, 436-48.

131. Choe, S., Hecht, K. A. & Grabe, M. (2008). A Continuum Method for Determining Membrane Protein Insertion Energies and the Problem of Charged Residues. *J Gen Physiol* 131, 563-73.

132. Dunbrack Jr., R. L. & Wang, G. (2003). PISCES: a protein sequence culling server. *Bioinformatics* 19, 1589-1591.

133. Yau, W.-M., Wimley, W. C., Gawrisch, K. & White, S. H. (1998). The Preference of Tryptophan for Membrane Interfaces. *Biochem* 37, 14713-14718.

134. Wiener, M. C. & White, S. H. (1992). Structure of a fluid dioleoylphosphatidylcholine bilayer determined by joint refinement of x-ray and neutron diffraction data. III. Complete structure. *Biophys J* 61, 434-47.

135. Wallace, I. S. & Roberts, D. M. (2004). Homology Modeling of Representative Subfamilies of Arabidopsis Major Intrinsic Proteins. Classification Based on the Aromatic/Arginine Selectivity Filter. *Plant Physiol* 135, 1029-1068.

136. Hille, B. (1992). *Ionic channels of excitable membranes*, Sinauer Associates, Sunderland, Massachusetts.

137. Chen, H., Sesti, F. & Goldstein, S. A. N. (2003). Pore- and State-Dependent Cadmium Block of IKs Channels Formed with MinK-55C and Wild-Type KCNQ1 Subunits. *Biophys J* 84, 3679-3689.

138. Yellen, G., Sodickson, D., Chen, T. Y. & Jurman, M. E. (1994). An engineered cysteine in the external mouth of a K+ channel allows inactivation to be modulated by metal binding. *Biophys J* 66, 1068-1075.

139. Gross, A. & Hubbell, W. L. (2002). Identification of Protein Side Chains near the Membrane-Aqueous Interface: A Site-Directed Spin Labeling Study of KcsA. *Biochem* 41, 1123-1128.

140. Bronson, J., Lee, O.-S. & Saven, J. G. (2006). Molecular Dynamics Simulation of WSK-3, a Computationally Designed, Water-Soluble Variant of the Integral Membrane Protein KcsA *Biophys J* 90, 1156-1163.

141. Pathak, M., Kurtz, L., Tombola, F. & Isacoff, E. (2004). The Cooperative Voltage Sensor Motion that Gates a Potassium Channel *J Gen Physiol* 125, 57-69.

142. Akiba, T., Toyoshima, C., Matsunaga, T., Kawamoto, M., Kubota, T., Fukuyama, K., Namba, K. & Matsubara, H. (1996). Three-dimensional structure of bovine cytochrome bC1 complex by electron cryomicroscopy and helical image reconstruction. *Nat Struct Biol* 3, 553-561.

143. Beckstein, O. & Sansom, M. S. P. (2004). The influence of geometry, surface character and flexibility on the permeation of ions and waters through biological pores. *Phys Biol* 1, 42-52.

144. Nagle, J. F. & Morowitz, H. J. (1978). Molecular Mechanisms for Proton Transport in Membranes. *PNAS* 75, 298-302.

145. Kazhdan, M., Funkhouser, T. & Rusinkeiwicz, S. (2003). Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors. *Symp Geom Proc*, 167-175.

146. Rahi, S. J. & Sharp, K. A. (2007). Mapping Complicated Surfaces onto a Sphere. *Int. J. Comput. Geom. Appl.* 17, 305-329.

147. Stetter, K. O. (2006). Hyperthermophiles in the history of life. *Phil Trans Roy Soc B* 361, 1837-1843.

148. Nisbet, E. G. & Sleep, N. H. (2001). The habitat and nature of early life. *Nature* 409, 1083-1091.

149. Puigbòa, P., Pasamontesa, A. & Garcia-Vallve, S. (2008). Gaining and losing the thermophilic adaptation in prokaryotes. *Trends in Genetics* 24, 10-14.

150. Berezovsky, I. N. & Shakhnovich, E. I. (2005). Physics and evolution of thermophilic adaptation. *PNAS* 102, 12742-12747.

151. Vogt, G. & Argos, P. (1997). Protein thermal stability: hydogren bonds or internal packing? *Folding & Design* 2, S40-S46.

152. Szilágyi, A. & Závodszky, P. (2000). Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure* 8, 493-504.

153. Karshikoff, A. & Ladenstein, R. (2001). Ion pairs and the thermotolerance of proteins from hyperthermophiles: a "traffic rule" for hot roads. *Trends Biochem Sci* 26, 550-556.

154. Xiao, L. & Honig, B. (1999). Electrostatic contributions to the stability of hyperthermophilic proteins. *J Mol Biol* 289, 1435-44.

155. Alsop, E., Silver, M. & Livesay, D. R. (2003). Optimized electrostatic surfaces parallel increased thermostability. *Prot. Eng.* 16, 871-4.

156. Capistran-Licea, V. M., Millan-Pacheco, C. & Pastor, N. (2009). Thermal Adaptation Strategies used by TBP [abstract]. *Biophys J* 96, 331a.

157. Paiardini, A., Sali, R., Bossa, F. & Pascarella, S. (2008). "Hot cores" in proteins: Comparative analysis of the apolar contact area in structures from hyper/thermophilic and mesophilic organisms *BMC Struct Biol* 8, 14.

158. Rader, A. J. (2009). Thermostabilization Due to Rigidity: A Case Study of Rubredoxin [abstract]. *Biophys J* 96, 330a.

159. Karshikoff, A. & Ladenstein, R. (1998). Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. *Prot. Eng.* 11, 867-72.

160. Ogata, Y., Imai, E.-I., Honda, H., Hatori, K. & Matsuno, K. (2000). Hydrothermal Circulation of Seawater Through Hot Vents and Contribution of Interface Chemistry to Prebiotic Synthesis. *Origins Life Evol Biosphere* 30, 527-537.

161. Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., McDonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., White, O., Salzberg, S. L., Smith, H. O., Venter, J. C. & Fraser, C. M. (1999). Evidence for lateral gene transfer between archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* 399, 323-329.

162. Richards, F. M. (1974). The interpretation of protein structures: Total volume, group volume distributions and packing density. *J Mol Biol* 82, 1-14.

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

164. Poupon, A. (2004). Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Curr Opin Struct Biol* 14, 233-241.

165. Levitt, M., Gerstein, M., Huang, E. S., Subbiah, S. & Tsai, J. (1997). PROTEIN FOLDING: The Endgame. *Ann Rev Biochem* 66, 549-579.

166. Gerstein, M., Tsai, J. & Levitt, M. (1995). The Volume of Atoms on the Protein Surface: Calculated from Simulation, using Voronoi Polyhedra. *J Mol Biol* 249, 955-966.

167. Liang, J. & Dill, K. A. (2001). Are proteins well-packed? *Biophys J* 86, 751-766.

168. Gerstein, M. & Chothia, C. (1996). Packing at the protein-water interface. *PNAS* 93, 10167-10172.

169. Liu, S., Zhang, C., Liang, S. & Zhou, Y. (2007). Fold recognition by concurrent use of solvent accessibility and residue depth. *Proteins: Struct. Funct. Bioinf.* 68, 636-45.

170. Huang, E. S., Subbiah, S. & Levitt, M. (1995). Recognizing Native Folds by the Arrangement of Hydrophobic and Polar Residues. *J Mol Biol* 252, 709-720.

171. Pereira de Araújo, A. F., Gomes, A. L. C., Bursztyn, A. A. & Shakhnovich, E. I. (2007). Native atomic burials, supplemented by physically motivated hydrogen bond constraints, contain sufficient information to determine the tertiary structure of small globular proteins. *Proteins: Struct. Funct. Bioinf.* 70, 971-983.

172. del Sol, A., Fujihashi, H., Amoros, D. & Nussinov, R. (2006). Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families. *Prot. Sci.* 15, 2120-8.

173. Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I. & Pietrokovski, S. (2004). Network analysis of Protein Structure Identifies Functional Residues. *J Mol Biol* 344, 1135-46.

174. Brinda, K. V. & Vishveshwara, S. (2005). A Network Representation of Protein Structures: Implications for Protein Stability. *Biophys J* 89, 4159-70.

175. Wadell, H. (1935). Volume, Shape and Roundness of Quartz Particles. *J Geol* 43, 250-280.

176. Fisher, R. A. (1936). "The Coefficient of Racial Likeness" and the Future of Craniometry. *Journal of the Royal Anthropological Institute* 66, 57-63.

177. Wang, W., Cho, H. S., Kim, R., Jancarik, J., Yokota, H., Nguyen, H. H., Grigoriev, I. V., Wemmer, D. E. & Kim, S. H. (2002). Structural characterization of the reaction pathway in phosphoserine phosphatase: crystallographic "snapshots" of intermediate states. *J Mol Biol* 319, 421-431.

178. Peeraer, Y., Rabijns, A., Verboven, C., Collet, J. F., Van Schaftingen, E. & De Ranter, C. (2003). High-resolution structure of human phosphoserine phosphatase in open conformation. *Acta Crystallogr D Biol Crystallogr* 59, 971-977.

179. Yee, A., Chang, X., Pineda-Lucena, A., Wu, B., Semesi, A., Le, B., Ramelot, T., Lee, G. M., Bhattacharyya, S., Gutierrez, P., Denisov, A., Lee, C. H., Cort, J. R., Kozlov, G., Liao, J., Finak, G., Chen, L., Wishart, D., Lee, W., McIntosh, L. P., Gehring, K., Kennedy, M. A., Edwards, A. M. & Arrowsmith, C. H. (2002). An NMR approach to structural proteomics. *Proc Natl Acad Sci USA* 99, 1825-1830.

180. Klapper, I., Hagstrom, R., Fine, R., Sharp, K. & Honig, B. (1986). Focusing of Electrical Fields in the Active Site of Cu-Zn Superoxide Dismutase: Effects of Ionic Strength and Amino-Acid Modifications. *Proteins* 1, 47-59.

181. Yang, Q. & Sharp, K. A. (2009). Building alternate protein structures using the elastic network model. *Prot. Sci.* 74, 682-700.

182. Canutescu, A. A., Shelenkov, A. A. & Dunbrack Jr., R. L. (2003). A graph theory algorithm for protein side-chain prediction. *Prot. Sci.* 12, 2001-2014.

183. Kono, H. & Saven, J. G. (2001). Statistical theory for protein combinatorial libraries. packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J Mol Biol* 306, 607-628.

184. Shah, P. S., Hom, G. K., Ross, S. A., Lassila, J. K., Crowhurst, K. A. & Mayo, S. L. (2007). Full-sequence Computational Design and Solution Structure of a Thermostable Protein Variant. *J Mol Biol* 372, 1-6.

185. Lehmann, M. & Wyss, M. (2001). Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution *Curr Opin Biotech* 12, 371-375.

186. van der Burg, B. & Eijsink, V. G. H. (2002). Selection of mutations for increased protein stability *Curr Opin Biotech* 13, 333-337.

187. DiTursi, M. K., Kwon, S.-J., Reeder, P. J. & Dordick, J. S. (2006). Bioinformatics-driven, rational engineering of protein thermostability. *Prot Eng Des Sel* 19, 517-524.

188. Danciulescu, C., Ladenstein, R. & Nilsson, L. (2007). Dynamic Arrangement of Ion Pairs and Individual Contributions to the Thermal Stability of the Cofactor-

Binding Domain of Glutamate Dehydrogenase from Thermotoga maritima. *Biochem* 46, 8537-8549.

189. Henzler-Wildman, K. A., Lei, M., Thai, V., Kerns, S. J., Karplus, M. & Kern, D. (2007). A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* 450, 913-916.

190. Hollien, J. & Marqusee, S. (1999). Structural distribution of stability in a thermophilic enzyme. *Proc Natl Acad Sci USA* 96, 13674-13678.

191. Luke, K. A., Higgins, C. L. & Wittung-Stafshede, P. (2007). Thermodynamic stability and folding of proteins from hyperthermophilic organisms. *FEBS* 274, 4023-4033.

192. Alvarez, J. & Shoichet, B. K., Eds. (2005). Virtual Screening in Drug Discovery. New York: Taylor & Francis.

193. Kortagere, S., Krasowski, M. D. & Ekins, S. (2009). The importance of discerning shape in molecular pharmacology. *Trends in Pharma Sci* 30, 138-147.

194. Nobeli, I., Favia, A. D. & Thornton, J. M. (2009). Protein promiscuity and its implications for biotechnology. *Nat Biotech* 27, 157-167.

195. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J Mol Biol* 161, 269-288.

196. Brady Jr., G. P. & Stouten, P. F. W. (2000). Fast prediction and visualization of protein binding pockets with PASS. *J Comput Aided Mol Des* 14, 383-401.

197. Zhong, S. & MacKerell Jr, A. D. (2007). Binding Response: A Descriptor for Selecting Ligand Binding Site on Protein Surfaces. *J Chem Inf Model* 47, 2303-2315.

198. Harris, R., Olson, A. J. & Goodsell, D. S. (2008). Automated prediction of ligand-binding sites in proteins. *Proteins: Struct. Funct. Bioinf.* 70, 1506-1517.

199. Cheng, A. C., Coleman, R. G., Smyth, K. T., Cao, Q., Soulard, P., Caffrey, D. R., Salzberg, A. C. & Huang, E. S. (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotech* 25, 71-75.

200. Weisel, M., Proschak, E. & Schneider, G. (2007). PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem Central J* 1, 7.

201. Xie, L. & Bourne, P. E. (2007). A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics* 8 S9.

202. Hendlich, M., Rippman, F. & Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins *J Mol Graph Model* 15, 359-363.

203. Kalidas, Y. & Chandra, N. (2008). PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins *J Struct Biol* 161, 31-42.

204. Coleman, R. G., Salzberg, A. C. & Cheng, A. C. (2006). Structure-Based Identification of Small Molecule Binding Sites Using a Free Energy Model. *J Chem Inf Model* 46, 2631-2637.

205. Ming, D., Cohn, J. D. & Wall, M. E. (2008). Fast dynamics perturbation analysis for prediction of protein functional sites. *BMC Struct Biol* 8, 5.

206. Chen, B. Y., Bryant, D. H., Fofanov, V. Y., Kristensen, D. M., Cruess, A. E., Kimmel, M., Lichtarge, O. & Kavraki, L. E. (2007). Cavity Scaling: Automated Refinement of Cavity-Aware Motifs in Protein Function Prediction. *J Bioinf Comp Biol* 5, 353-382.

207. Joughin, B. A., Tidor, B. & Yaffe, M. B. (2005). A computational method for the analysis and prediction of protein:phosphopeptide-binding sites. *Prot Sci* 14, 131-139.

208. Pettit, F. K., Bare, E., Tsai, A. & Bowie, J. U. (2007). HotPatch: A Statistical Approach to Finding Biologically Relevant Features on Protein Surfaces. *J Mol Biol* 369, 863-879.

209. Schmitt, S., Kuhn, D. & Klebe, G. (2002). A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *J Mol Biol* 323, 387-406.

210. Kuhn, D., Weskamp, N., Schmitt, S., Hüllermeier, E. & Klebe, G. (2006). From the Similarity Analysis of Protein Cavities to the Functional Classification of Protein Families Using Cavbase. *J Mol Biol* 359, 1023-1044.

211. Weskamp, N., Hüllemmeier, E. & Klebe, G. (2009). Merging chemical and biological space: Structural mapping of enzyme binding pocket space. *Proteins: Struct. Funct. Bioinf.* 76, 317-330.

212. Powers, R., Copeland, J. R., Germer, K., Mercier, K. A., Ramanathan, V. & Revenz, P. (2006). Comparison of Protein Active Site Structures for Functional Annotation of Proteins and Drug Design. *Proteins: Struct. Funct. Bioinf.* 65, 124-135.

213. Petsalaki, E., Stark, A., García-Urdiales, E. & Russell, R. B. (2009). Accurate Prediction of Peptide Binding Sites on Protein Surfaces. *PLoS Comp Biol* 5, 1-10.

214. Kinnings, S. L. & Jackson, R. M. (2009). Binding Site Similarity Analysis for the Functional Classification of the Protein Kinase Family. *J Chem Inf Model* 49, 318-329.

215. Caffrey, D. R., Lunney, E. A. & Moshinsky, D. J. (2008). Prediction of specificity-determining residues for small-molecule kinase inhibitors. *BMC Bioinformatics* 9.

216. Gold, N. D. & Jackson, R. M. (2006). SitesBase: a database for structure-based protein–ligand binding site comparisons. *Nuc Acid Res* 34, 231-234.

217. Kinjo, A. R. & Nakamura, H. (2009). Comprehensive Structural Classification of Ligand-Binding Motifs in Proteins. *Structure* 17, 234-246.

218. Rosen, M., Lin, S. L., Wolfson, H. & Nussinov, R. (1998). Molecular shape comparisons in searches for active sites and functional similarity. *Prot Eng Des Sel* 11, 263-277.

219. Xie, L., Wang, J. & Bourne, P. E. (2007). In Silico Elucidation of the Molecular Mechanism Defining the Adverse Effect of Selective Estrogen Receptor Modulators. *PLoS Comp Biol* 3, e217.

220. Xie, L. & Bourne, P. E. (2008). Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments. *Proc Natl Acad Sci USA* 105, 5441-5446.

221. Halgren, T. A. (2009). Identifying and Characterizing Binding Sites and Assessing Druggability *J Chem Inf Model* 49, 377-389.

222. Hermann, J. C., Marti-Arbona, R., Fedorov, A. A., Fedorov, E., Almo, S. C., Shoichet, B. K. & Raushel, F. M. (2007). Structure-based activity prediction for an enzyme of unknown function *Nature* 448, 775-779.

223. Tarjan, R. E. (1975). Efficiency of a Good But Not Linear Set Union Algorithm. *J Assoc Comput Mach* 22, 215-225.

224. Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2, 559-572.

225. Cheng, A. C. (2008). Predicting Selectivity and Druggability in Drug Discovery. *Ann Rep Comp Chem* 4, 23-37.

226. Sitkoff, D., Sharp, K. & Honig, B. (1994). Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J. Phys. Chem.* 98, 1978-1988.

227. Graham, R. L. & Hell, P. (1985). On the History of the Minimum Spanning Tree Problem. *Ann History Computing* 7, 43-57.

228. Gansner, E. R., Koutsofios, E., North, S. C. & Vo, K. (1993). A technique for drawing directed graphs. *IEEE Transactions on Software Engineering* 19, 214-230.

229. Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C. & Woodhull, G. (2004). Graphviz and Dynagraph -- Static and Dynamic Graph Drawing Tools. In *Graph Drawing Software* (Junger, M. & Mutzel, P., eds.). Springer-Verlag.

230. (2009). aiSee 3.0.5 edit. AbsInt, Saarbruecken, Germany.

231. Wilmanns, M., Priestle, J. P., Neirmann, T. & Jansonius, J. N. (1992). Three-dimensional Structure of the Bifunctional Enzyme Phosphoribosylanthranilate

Isomerase : Indoleglycerolphosphate Synthase from Escherichia coli Refined at 2.0Å Resolution. *J Mol Biol* 223, 477-507.

232. Mosimann, S. C., Ardelt, W. & James, M. N. G. (1994). Refined 1·7 Å X-ray crystallographic structure of P-30 protein, an amphibian ribonuclease with anti-tumor activity *J Mol Biol* 236, 1141-1153.

233. Allen, K. N., Bellamacina, C. R., Ding, X., Jeffery, C. J., Mattos, C., Petsko, G. A. & Ringe, D. (1996). An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *J Phys Chem* 100, 2605-2611.

234. Teotico, D. G., Babaoglu, K., Rocklin, G. J., Ferreira, R., Giannetti, A. M. & Shoichet, B. K. (2009). Docking for fragment inhibitors of AmpC _-lactamase. *Proc Natl Acad Sci USA* 106, 7455-7460.

235. Hubbard, R. E., Chen, I. & Davis, B. (2007). Informatics and modeling challenges in fragment-based drug discovery. *Curr Opin Drug Discovery Dev* 10, 289-297.

236. Chen, Y. & Shoichet, B. K. (2009). Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat Chem Biol* 5, 358-364.

237. Verlinde, C. L. M. J., Rudenko, G. & Hol, W. G. J. (1992). In search of new lead compounds for trypanosomiasis drug design: A protein structure-based linked-fragment approach. *J Comp-Aided Molec Des* 6, 131-147.

238. Hubbard, R. E., Davis, B., Chen, I. & Drysdale, M. J. (2007). The SeeDs Approach: Integrating Fragments into Drug Discovery. *Curr Topics Med Chem* 7, 1568-1581.

239. Verdonk, M. L. & Hartshort, M. J. (2004). Structure-guided fragment screening for lead discovery. *Curr Opin Drug Discov Devel* 7, 404-410.

240. Müller, C. W., Schlauderer, G. J., Reinstein, J. & Schulz, G. E. (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* 4, 147-156.

241. Müller, C. W. & Schulz, G. E. (1992). Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor Ap5A refined at 1.9 Å resolution: A model for a catalytic transition state. *J Mol Biol* 224, 159-177.

242. Vonrhein, C., Schlauderer, G. J. & Schulz, G. E. (1995). Movie of the structural changes during a catalytic cycle of nucleoside monophosphate kinases. *Structure* 3, 483-490.

243. Beckstein, O., Denning, E. J. & Woolf, T. B. (2009). The Closed <-> Open Transition of Adenylate Kinase From Crystal Structures and Computer Simulations [abstract]. *Biophys J* 96, 70a-71a.

244. Henzler-Wildman, K. A., Thai, V., Lei, M., Ott, M., Wolf-Watz, M., Fenn, T., Pozharski, E., Wilson, M. A., Petsko, G. A., Karplus, M., Hübner, C. G. & Kern, D. (2007). Intrinsic motions along an enzymatic reaction trajectory. *Nature* 450, 838-844.

245. Weiss, D. R. & Levitt, M. (2009). Can Morphing Methods Predict Intermediate Structures? *J Mol Biol* 385, 665-674.

246. Fonzé, E., Charlier, P., To'th, Y., Vermeire, M., Raquet, X., Dubus, A. & Frère, J.-M. (1995). TEM1 beta-lactamase structure solved by molecular replacement and refined structure of the S235A mutant. *Acta Cryst D* 51, 682-694.

247. Wang, X., Minasov, G. & Shoichet, B. K. (2002). Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-offs. *J Mol Biol* 320, 85-95.

248. Marciano, D. C., Pennington, J. M., Wang, X., Wang, J., Chen, Y., Thomas, V. L., Shoichet, B. K. & Palzkill, T. (2008). Genetic and Structural Characterization of an L201P Global Suppressor Substitution in TEM-1 _-Lactamase. *J Mol Biol* 384, 151-164.

249. Thomas, V. L., Golemi-Kotra, D., Kim, C., Valulenko, S. B., Mobashery, S. & Shoichet, B. K. (2005). Structural Consequences of the Inhibitor-Resistant Ser130Gly Substitution in TEM _-Lactamase. *Biochem* 44, 9330-9338.

250. Wang, X., Minasov, G., Blásquez, J., Caselli, E., Prati, F. & Shoichet, B. K. (2003). Recognition and Resistance in TEM _-Lactamase. *Biochem* 42, 8434-8444.

251. Wang, X., Minasov, G. & Shoichet, B. K. (2002). The Structural Bases of Antibiotic Resistance in the Clinically Derived Mutant _-Lactamases TEM-30, TEM-32, and TEM-34. *J Biol Chem* 277, 32149-32156.

252. Swarén, P., Golemi, D., Cabantous, S., Bulychev, A., Maveyraud, L., Mobashery, S. & Samana, J.-P. (1999). X-ray Structure of the Asn276Asp Variant of the Escherichia coli TEM-1 _-Lactamase: Direct Observation of Electrostatic Modulation in Resistance to Inactivation by Clavulanic Acid. *Biochem* 38, 9570-9576.

253. Barr, A. J., Ugochukwu, E., Lee, W. H., King, O. N. F., Filippakopouls, P., Alfano, I., Savitsky, P., Burgess-Brown, N. A., Müller, S. & Knapp, S. (2009).

Large-Scale Structural Analysis of the Classical Human Protein Tyrosine Phosphatome. *Cell* 136, 352-363.

254. Andersen, J. N., Mortensen, O. H., Peters, G. H., Drake, P. G., Iversen, L. F., Olsen, O. H., Jansen, P. G., Andersen, H. S., Tonks, N. K. & Møller, N. P. H. (2001). Structural and Evolutionary Relationships among Protein Tyrosine Phosphatase Domains *Mol Cell Biol* 21, 7117-7136.

255. Villa, F., Deak, M., Bloomberg, G. B., Alessi, D. R. & van Aalten, D. M. (2004). Crystal structure of the PTPL1/FAP-1 human tyrosine phosphatase mutated in colorectal cancer: evidence for a second phosphotyrosine substrate recognition pocket. *J Biol Chem* 280, 8180-8187.

256. Hof, P., Pluskey, S., Dhe-Paganon, S., Eck, M. J. & Shoelson, S. E. (1998). Crystal structure of the tyrosine phosphatase SHP-2. *Cell* 92, 441-450.

257. Puius, Y. A., Zhao, Y., Sullivan, M., Lawrence, D. S., Almo, S. C. & Zhang, Z. Y. (1997). Identification of a second aryl phosphate-binding site in protein-tyrosine phosphatase 1B: a paradigm for inhibitor design. *Proc Natl Acad Sci USA* 94, 13420-13425.

258. Liu, G., Xin, Z., Hajduk, P. J., Abad-Zapatero, C., Hutchins, C. W., Zhao, H., Lubben, T. H., Ballaron, S. J., Haasch, D. L., Kaszubska, W., Rondinone, C. M., Trevillyan, J. M. & Jirousek, M. R. (2003). Fragment screening and assembly: a highly efficient approach to a selective and cell active protein tyrosine phosphatase 1B inhibitor. *J Med Chem* 46, 4232-4235.

259. Szczepankiewicz, B. G., Liu, G., Hajduk, P. J., Abad-Zapatero, C., Pei, Z., Xin, Z., Lubben, T. H., Trevillyan, J. M., Stashko, M. A., Ballaron, S. J., Liang, H., Huang, F., Hutchins, C. W., Fesik, S. W. & Jirousek, M. R. (2003). Discovery

of a potent, selective protein tyrosine phosphatase 1B inhibitor using a linked-fragment strategy. *J Am Chem Soc* 125, 4087-4096.

260. Liu, G., Szczepankiewicz, B. G., Pei, Z., Janowick, D. A., Xin, Z., Hajduk, P. J., Abad-Zapatero, C., Liang, H., Hutchins, C. W., Fesik, S. W., Ballaron, S. J., Stashko, M. A., Lubben, T., Mika, A. K., Zinker, B. A., Trevillyan, J. M. & Jirousek, M. R. (2003). Discovery and structure-activity relationship of oxalylarylaminobenzoic acids as inhibitors of protein tyrosine phosphatase 1B. *J Med Chem* 46, 2093-2103.

261. Wells, J. A. & McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* 450, 1001-1009.

262. Zhong, S., Macias, A. T. & MacKerell Jr., A. D. (2007). Computational Identification of Inhibitors of Protein-Protein Interactions. *Curr Topics Med Chem* 7, 63-82.

263. Fuller, J. C., Burgoyne, N. J. & Jackson, R. M. (2009). Predicting druggable binding sites at the protein-protein interface. *Drug Discovery Today* 14, 155-161.

264. Chen, T., Kablaoui, N., Little, J., Timofeevski, S., Tschantz, W. R., Chen, P., Feng, J., Charlton, M., Stanton, R. & Bauer, P. (2009). Identification of small-molecule inhibitors of the JIP–JNK interaction. *Biochem J* 420, 283-294.

265. Dömling, A. (2008). Small molecular weight protein–protein interaction antagonists—an insurmountable challenge? *Curr Opin Chem Biol* 12, 281-291.

266. Eyrisch, S. & Helms, V. (2007). Transient Pockets on Protein Surfaces Involved in Protein-Protein Interaction. *J Med Chem* 50, 3457-3464.

267. Kahraman, A., Morris, R. J., Laskowski, R. A. & Thornton, J. M. (2007). Shape Variation in Protein Binding Pockets and their Ligands *J Mol Biol* 368, 283-301.

268. Hert, J., Keiser, M. J., Irwin, J. J., Oprea, T. I. & Shoichet, B. K. (2008). Quantifying the Relationships among Drug Classes. *J Chem Inf Model* 48, 755-765.

269. Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J. & Shoichet, B. K. (2007). Relating protein pharmacology by ligand chemistry. *Nat Biotech* 25, 197-206.

270. Rush III, T. S., Grant, J. A., Mosyak, L. & Nicholls, A. (2005). A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J Med Chem* 48, 1489-1495.

271. Stockwell, G. R. & Thornton, J. M. (2005). Conformational Diversity of Ligands Bound to Proteins. *J Mol Biol* 356, 928-944.

272. Koch, M. A., Schuffenhauer, A., Scheck, M., Wetzel, S., Casaulta, M., Odermatt, A., Ertl, P. & Waldmann, H. (2005). Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *Proc Natl Acad Sci USA* 102, 17272-172777.

273. Favia, A. D., Nobeli, I., Glaser, F. & Thornton, J. M. (2007). Molecular Docking for Substrate Identification: The Short-Chain Dehydrogenases/Reductases. *J Mol Biol* 375, 855-874.

274. Hardy, J. A., Lam, J., Nguyen, J. T., O'Brien, T. & Wells, J. A. (2004). Discovery of an allosteric site in the caspases. *Proc Natl Acad Sci USA* 101, 12461-12466.

275. Hardy, J. A. & Wells, J. A. (2004). Searching for new allosteric sites in enzymes. *Curr Opin Struct Biol* 14, 706-715.

276. White, S. H. & Wimley, W. C. (1998). Hydrophobic interactions of peptides with membrane interfaces. *BBA - Reviews on Biomembranes* 1376, 339-352.

277. Robertson, J. L., Palmer, L. G. & Roux, B. (2008). Long-pore Electrostatics in Inward-rectifier Potassium Channels. *J Gen Physiol* 132, 613-632.

Appendix A

Table A-1 Travel Depth of the PDBbind Dataset

PDB Code	Affinity (pKd)	Average depth (overall) (Å)	Average depth (binding site) (Å)	p-value	Buried Surface Area (Å^2)	# of ligand heavy atoms
10gs	6.4	6.4	14.1	0.00E+00	858	22
11gs	5.82	6.0	14.0	0.00E+00	1027	22
16pk	5.22	5.7	17.9	0.00E+00	910	23
1a07	6.4	4.4	4.5	4.25E-01	591	7
1a08	5.62	3.0	4.8	2.00E-07	740	7
1a0q	7.57	5.8	10.2	2.58E-04	684	17
1a1b	6.4	4.5	5.5	7.35E-02	724	8

1a1c	6.4	4.5	5.5	6.16E-02	750	8
1a1e	6	4.4	4.9	2.59E-01	728	7
1a30	4.3	3.4	9.7	0.00E+00	790	15
1a42	9.89	4.0	12.1	0.00E+00	710	19
1a4k	8	6.3	9.1	1.73E-03	702	14
1a4m	13	4.5	15.7	0.00E+00	564	19
1a4w	5.92	4.0	8.7	0.00E+00	1070	15
1a50	6.7	5.5	0.0	1.00E+00	623	0
1a69	5.3	6.3	16.7	0.00E+00	571	21
1a7t	1.64	3.8	10.9	0.00E+00	453	14
1a7x	9.7	4.3	11.8	0.00E+00	1295	18
1a94	7.85	3.5	8.8	0.00E+00	1559	15
1a99	5.7	4.1	16.2	0.00E+00	330	19

1a9m	6.92	3.9	10.0	0.00E+00	1288	15
1aaq	8.4	3.6	10.1	0.00E+00	1264	15
1abf	5.42	3.9	19.2	0.00E+00	379	21
1add	6.74	4.2	15.7	0.00E+00	577	19
1adl	5.36	5.0	13.0	0.00E+00	828	18
1ado	6	6.2	22.1	0.00E+00	329	25
1af6	1.82	9.3	30.5	0.00E+00	536	34
1afk	6.62	3.1	7.1	0.00E+00	662	12
1afl	6.28	3.1	7.4	0.00E+00	673	12
1agm	12	4.8	9.9	0.00E+00	853	18
1ai4	2.5	7.5	32.8	0.00E+00	426	36
1ai5	3.72	7.4	32.9	0.00E+00	451	37
1ai7	4.09	7.5	35.2	0.00E+00	297	38

1aj6	5.92	4.4	7.0	1.64E-04	780	13
1aj7	3.87	5.4	10.4	6.40E-06	590	16
1ajn	2.63	7.4	32.1	0.00E+00	440	36
1ajp	2.23	7.4	33.2	0.00E+00	400	37
1ajq	4.31	7.5	33.5	0.00E+00	367	38
1ajv	7.72	3.4	11.1	0.00E+00	1167	15
1ajx	7.91	3.3	11.0	0.00E+00	1126	16
1alw	6.52	6.1	9.1	3.39E-02	408	13
1anf	5.46	4.5	15.9	0.00E+00	639	20
1apb	5.82	3.8	19.0	0.00E+00	391	21
1apv	9	4.0	13.1	0.00E+00	1011	18
1apw	8	3.9	13.1	0.00E+00	1001	18
1at6	4.07	2.6	4.9	0.00E+00	702	12

1atl	6.28	5.5	9.4	2.86E-04	662	15
1avn	3.9	3.9	10.3	6.10E-05	247	14
1awi	4.05	4.8	12.3	0.00E+00	1317	18
1ax0	3.13	4.0	5.2	6.49E-02	432	8
1ax1	3.29	4.0	5.2	5.72E-02	408	9
1ax2	3.99	4.0	4.8	1.04E-01	439	9
1axz	3.2	6.1	6.4	3.77E-01	349	9
1b05	7.12	4.6	0.0	1.00E+00	960	0
1b0h	6.7	6.4	22.8	0.00E+00	1148	26
1b1h	7.03	6.3	22.3	0.00E+00	1097	29
1b2h	4.54	6.7	24.5	0.00E+00	1019	29
1b32	7.1	6.5	24.1	0.00E+00	1044	27
1b3f	6.89	6.7	24.2	0.00E+00	1048	28

1b3g	6.7	6.2	24.4	0.00E+00	1001	31
1b3h	6.21	6.2	22.9	0.00E+00	1087	27
1b3l	5.89	6.2	22.9	0.00E+00	874	27
1b40	7.28	6.0	22.8	0.00E+00	1070	29
1b42	4.01	6.7	8.9	1.09E-01	396	12
1b46	5.28	6.0	22.6	0.00E+00	933	28
1b4h	5.46	6.3	23.2	0.00E+00	992	27
1b4z	5.23	6.2	23.7	0.00E+00	983	27
1b51	7.37	6.4	23.0	0.00E+00	929	27
1b52	7.12	6.3	23.0	0.00E+00	966	30
1b55	7.4	6.2	6.9	3.39E-01	589	10
1b58	6.59	6.1	23.4	0.00E+00	1096	29
1b5h	6.01	6.3	23.1	0.00E+00	944	27

1b5i	7.05	6.3	23.5	0.00E+00	1009	31
1b5j	7.43	5.6	22.6	0.00E+00	1035	27
1b6h	7.82	6.4	23.5	0.00E+00	992	28
1b6j	7.92	4.0	9.4	0.00E+00	1317	15
1b6k	8.74	3.7	9.2	0.00E+00	1317	14
1b6l	8.3	4.0	10.1	0.00E+00	1114	15
1b6m	8.4	3.8	9.7	0.00E+00	1243	14
1b6n	8.4	4.0	10.4	0.00E+00	962	15
1b6o	9.22	3.7	10.0	0.00E+00	955	14
1b6p	8.52	3.8	9.4	0.00E+00	1307	14
1b74	1.3	4.7	18.8	0.00E+00	359	21
1b7h	8.02	6.0	23.4	0.00E+00	1029	28
1b8o	10.64	3.6	12.3	0.00E+00	547	15

1b8y	7.85	3.4	10.4	0.00E+00	784	13
1b9j	5.96	6.3	22.2	0.00E+00	1014	25
1bai	7.7	4.4	10.9	0.00E+00	1643	17
1bap	6.85	3.9	18.9	0.00E+00	348	21
1bcd	8.7	4.1	17.8	0.00E+00	299	19
1bcj	3.7	8.5	6.8	7.94E-01	365	10
1bcu	5	4.4	10.3	0.00E+00	473	15
1bdq	6.34	3.9	11.4	0.00E+00	1170	15
1bgq	8.57	5.6	13.3	0.00E+00	675	18
1bhf	4.38	3.3	5.0	6.00E-07	891	8
1bhx	6.84	4.7	11.5	0.00E+00	870	16
1bky	3.84	6.6	7.8	2.55E-01	367	10
1bm7	7.52	5.3	17.9	0.00E+00	647	23

1bma	4.59	4.2	9.6	0.00E+00	860	17
1bn1	9.34	3.8	11.6	0.00E+00	586	18
1bn3	9.89	4.0	11.7	0.00E+00	601	19
1bn4	9.31	3.9	11.5	0.00E+00	588	18
1bnn	10	3.9	11.7	0.00E+00	604	19
1bnq	9.49	3.9	12.5	0.00E+00	694	19
1bnt	9.8	3.9	11.9	0.00E+00	604	18
1bnu	9.7	3.9	12.8	0.00E+00	629	19
1bnv	8.77	3.8	11.3	0.00E+00	675	18
1bnw	9.08	4.2	12.7	0.00E+00	564	19
1bq4	5.22	8.5	11.9	8.65E-02	616	16
1br5	2.7	4.7	13.1	0.00E+00	504	16
1br6	3.22	4.6	13.2	0.00E+00	565	17

1bra	1.82	6.1	18.9	0.00E+00	368	21
1bv7	9.3	3.8	10.1	0.00E+00	1361	15
1bv9	8.96	3.6	9.9	0.00E+00	1368	15
1bwa	7.6	3.7	10.0	0.00E+00	1368	15
1bwb	7.42	3.6	9.3	0.00E+00	1475	15
1bxo	10	4.1	12.6	0.00E+00	1166	19
1bxq	7.38	3.9	12.0	0.00E+00	1163	18
1byk	5	6.5	16.7	0.00E+00	815	21
1bzc	4.92	3.8	7.3	0.00E+00	699	14
1bzh	6.77	4.3	7.2	0.00E+00	826	14
1bzj	4.66	3.9	9.5	0.00E+00	531	14
1bzy	8.34	7.2	13.6	8.73E-04	684	18
1c1r	7.63	3.5	8.8	2.00E-07	546	14

1c1u	8.25	4.3	12.9	0.00E+00	682	18
1c1v	7.64	4.4	12.8	0.00E+00	772	18
1c2d	8.28	3.5	9.1	0.00E+00	616	13
1c3x	3.68	8.5	18.3	1.30E-03	495	20
1c4u	10.37	4.7	11.9	0.00E+00	914	17
1c4v	10.8	4.4	11.2	0.00E+00	965	17
1c5c	6.96	6.1	9.8	3.88E-03	633	16
1c5n	4.7	4.7	13.2	0.00E+00	498	16
1c5o	3.49	4.4	14.8	0.00E+00	362	17
1c5p	4.68	3.5	10.5	0.00E+00	365	13
1c5q	6.36	3.3	9.7	0.00E+00	482	13
1c5s	6	3.6	9.9	6.00E-07	417	13
1c5t	4.1	3.6	9.9	1.20E-06	425	14

1c6y	9.51	3.7	10.3	0.00E+00	1327	14
1c70	10.3	3.7	10.4	0.00E+00	1382	15
1c83	4.85	4.0	8.7	0.00E+00	496	12
1c84	5	3.8	8.9	0.00E+00	524	13
1c86	4.7	3.9	7.8	0.00E+00	528	13
1c87	4.2	4.0	10.3	0.00E+00	496	14
1c88	5.29	3.8	8.5	0.00E+00	519	14
1caq	7.72	3.7	10.1	0.00E+00	946	13
1cbx	6.35	4.5	12.5	1.60E-06	514	18
1ce5	4.74	3.5	10.6	0.00E+00	378	14
1cea	4.96	2.2	3.2	1.46E-02	344	5
1ceb	6	2.2	5.1	0.00E+00	379	8
1cet	2.89	6.2	7.1	2.49E-01	520	13

1cil	9.43	4.0	12.9	0.00E+00	602	19
1cim	8.82	4.1	13.1	0.00E+00	544	18
1cin	8.73	3.9	13.2	0.00E+00	570	19
1ciz	7.44	3.5	9.8	0.00E+00	946	15
1clu	8.27	3.5	7.3	0.00E+00	789	11
1cnw	7.72	4.1	11.2	0.00E+00	629	19
1cnx	7.37	4.0	11.9	0.00E+00	638	19
1cny	7.85	4.1	11.6	0.00E+00	619	19
1cps	6.66	4.0	10.9	2.00E-07	594	16
1cru	2.3	8.4	21.7	0.00E+00	592	25
1ct8	6.52	6.9	18.0	0.00E+00	934	24
1ctt	4.52	5.6	19.3	0.00E+00	495	23
1ctu	11.92	5.5	19.0	0.00E+00	502	22

1d09	7.57	8.4	25.3	2.00E-07	506	28
1d1p	3.6	3.3	6.7	0.00E+00	452	13
1d3d	9.09	4.5	10.4	0.00E+00	1047	18
1d3p	6.54	4.6	10.7	0.00E+00	1048	17
1d4k	9.22	3.9	9.3	0.00E+00	1383	14
1d4l	8.77	3.8	10.1	0.00E+00	1213	14
1d4p	6.3	4.4	12.2	0.00E+00	829	18
1d4s	9	3.6	10.4	0.00E+00	1163	15
1d4y	11.1	3.3	10.2	0.00E+00	1207	15
1d5r	1.82	5.2	16.6	0.00E+00	343	19
1d6v	6.17	6.5	12.8	1.80E-06	679	21
1d6w	5.96	4.5	11.6	0.00E+00	939	16
1d7i	3.6	2.7	8.4	0.00E+00	326	11

1d7j	3.3	2.5	9.4	0.00E+00	293	11
1d9i	9.11	4.8	12.2	0.00E+00	868	17
1db1	9.26	5.1	22.1	0.00E+00	1065	30
1df8	9.7	5.8	14.9	0.00E+00	557	19
1dg9	2.74	3.4	7.4	0.00E+00	473	13
1dhi	7.26	5.5	15.9	0.00E+00	832	23
1dhj	6.55	5.6	16.6	0.00E+00	845	24
1dif	10.66	3.5	9.6	0.00E+00	1478	15
1dl7	6.49	4.0	5.5	1.42E-02	635	10
1dmp	9.55	3.7	11.2	0.00E+00	1098	15
1dqn	8	6.0	14.2	0.00E+00	668	18
1dqx	11.05	5.9	20.0	0.00E+00	686	26
1drj	7.4	4.1	17.8	0.00E+00	310	20

1drk	6.82	4.0	17.1	0.00E+00	309	19
1dud	4.82	6.0	11.3	2.37E-04	617	15
1duv	11.8	7.7	26.2	0.00E+00	584	29
1dy4	4.36	4.7	18.1	0.00E+00	676	24
1e1v	4.92	4.8	13.2	0.00E+00	618	19
1e1x	5.89	4.9	13.3	0.00E+00	626	20
1e2k	4.94	7.4	18.9	0.00E+00	585	24
1e2l	4.29	7.5	17.7	2.00E-07	586	23
1e2n	4.51	7.4	22.5	0.00E+00	648	26
1e2p	4.57	7.2	17.6	2.00E-07	526	23
1e3v	4.34	4.5	12.0	0.00E+00	684	18
1e4h	8.41	5.3	15.7	6.00E-07	721	21
1e5a	7.64	5.9	16.2	5.50E-05	657	21

1e66	9.89	5.5	19.6	0.00E+00	698	23
1e6q	3.15	4.7	22.8	0.00E+00	398	25
1e6s	3.22	4.7	22.7	0.00E+00	388	25
1e70	3.05	5.0	23.1	0.00E+00	349	26
1ebg	10.82	7.3	0.0	1.00E+00	310	0
1ec9	3.1	5.1	26.4	0.00E+00	389	29
1ecq	3	5.2	23.6	0.00E+00	386	26
1ecv	4.85	4.0	10.0	0.00E+00	523	13
1eed	4.79	4.3	12.2	0.00E+00	1142	17
1efy	8.22	5.5	18.4	0.00E+00	608	23
1egh	5.7	8.3	18.7	5.76E-04	327	22
1eix	11.06	5.3	20.0	0.00E+00	654	25
1ejn	5.62	4.1	11.6	0.00E+00	690	15

1ela	6.36	4.1	9.2	0.00E+00	749	15
1elb	7.15	4.1	10.3	0.00E+00	716	16
1elc	6.66	4.0	9.9	0.00E+00	893	17
1eld	6.7	3.8	9.2	0.00E+00	725	14
1ele	6.85	4.0	8.2	0.00E+00	742	14
1elr	4.96	3.6	9.6	0.00E+00	937	14
1els	10.82	7.0	22.1	0.00E+00	350	24
1ent	6.96	4.3	12.2	0.00E+00	1248	17
1eoc	6.05	5.4	15.4	0.00E+00	384	19
1epo	7.96	4.3	13.0	0.00E+00	1208	18
1epp	7.16	4.3	11.6	0.00E+00	1263	17
1epq	8.19	4.4	12.4	0.00E+00	1044	18
1epv	6.89	7.4	23.4	0.00E+00	705	27

1erb	7.05	4.8	15.5	0.00E+00	916	23
1ets	8.22	5.2	12.6	0.00E+00	959	18
1ett	5.89	4.7	12.6	0.00E+00	845	18
1evh	3.22	3.4	4.8	3.66E-04	749	8
1ex8	6.33	3.7	10.6	0.00E+00	1235	15
1ez9	5.1	4.9	13.0	0.00E+00	837	19
1ezq	9.05	5.1	9.2	2.00E-07	959	16
1f0r	7.66	4.7	9.3	6.00E-07	818	18
1f0s	7.74	4.7	8.6	0.00E+00	778	16
1f0t	6	3.4	8.0	0.00E+00	736	15
1f0u	7.16	3.4	7.2	0.00E+00	843	14
1f2o	1.91	3.1	8.5	0.00E+00	389	12
1f2p	1.9	2.9	9.3	0.00E+00	447	12

1f3e	6.7	4.7	17.6	0.00E+00	487	21
1f3f	4.67	8.6	11.0	7.82E-02	704	15
1f4e	2.96	5.7	15.9	0.00E+00	563	20
1f4f	4.62	6.0	13.3	0.00E+00	835	18
1f4g	6.48	5.5	12.2	0.00E+00	973	17
1f4x	5.59	5.7	4.4	8.77E-01	522	8
1f57	5.64	3.8	12.9	0.00E+00	315	15
1f5k	3.74	4.2	13.3	0.00E+00	363	17
1f5l	5.28	4.0	12.3	0.00E+00	495	16
1f73	2.39	9.0	28.7	0.00E+00	566	34
1f74	3.05	6.2	20.7	0.00E+00	577	24
1f8a	5	4.3	9.4	0.00E+00	1001	15
1f8b	5.4	4.2	12.1	0.00E+00	594	17

1f8c	7.4	4.3	12.2	0.00E+00	602	16
1f8d	3.4	4.1	12.2	0.00E+00	608	16
1f8e	4.82	4.3	12.5	0.00E+00	614	16
1f9g	1.28	6.3	20.8	0.00E+00	358	24
1fao	7.37	3.6	4.9	5.52E-02	569	8
1fch	7.15	5.0	14.0	0.00E+00	1160	20
1fcx	7.19	4.8	16.2	0.00E+00	886	21
1fcy	8.52	4.2	15.4	0.00E+00	880	20
1fcz	9.22	4.5	16.4	0.00E+00	859	20
1fd0	8.4	4.8	16.5	0.00E+00	898	21
1fdq	7.27	4.0	11.5	0.00E+00	898	18
1fh7	5.24	3.5	12.3	0.00E+00	491	18
1fh8	6.89	3.7	12.8	0.00E+00	472	17

1fh9	6.43	3.5	12.0	0.00E+00	523	17
1fhd	6.82	3.7	12.3	0.00E+00	531	17
1fj4	4.59	6.8	17.1	4.20E-06	543	19
1fjs	9.96	4.9	9.7	0.00E+00	889	17
1fkb	9.7	2.9	6.4	0.00E+00	906	12
1fkf	9.4	4.1	13.2	0.00E+00	1017	18
1fkg	8	2.9	7.7	0.00E+00	727	12
1fkh	8.15	2.8	8.0	0.00E+00	740	12
1fki	7	4.2	7.4	1.20E-06	689	13
1fkn	8.8	5.1	11.8	0.00E+00	1476	18
1fkw	5.05	4.1	15.6	0.00E+00	563	19
1fkx	2.22	4.4	16.1	0.00E+00	567	21
1fl3	6.8	6.4	16.4	0.00E+00	802	25

1flr	7	5.9	8.3	1.90E-02	691	13
1fm9	9	7.2	23.2	0.00E+00	1285	30
1fmb	10	4.4	10.6	0.00E+00	1018	17
1fpc	7	4.3	8.8	0.00E+00	978	13
1fpu	7.43	5.3	15.9	0.00E+00	906	20
1fq5	8.4	5.2	12.1	0.00E+00	1469	19
1ftm	7.61	4.5	20.0	0.00E+00	456	22
1fv0	5.93	3.2	7.8	0.00E+00	724	13
1fwu	3.7	2.6	2.5	5.19E-01	416	5
1fwv	3.72	2.8	2.4	8.20E-01	431	5
1fzj	8.1	6.4	10.9	0.00E+00	1664	17
1fzk	8.4	6.8	11.2	0.00E+00	1668	18
1fzm	7.7	6.7	10.2	0.00E+00	1871	19

1fzo	7.89	6.6	10.3	4.00E-07	1638	17
1g1d	9.44	4.3	13.3	0.00E+00	569	20
1g2k	7.96	3.6	10.8	0.00E+00	1287	15
1g2l	7.24	4.8	9.2	0.00E+00	931	16
1g2o	10.55	8.0	52.7	0.00E+00	532	56
1g30	6.85	4.4	12.0	0.00E+00	904	18
1g32	6.11	4.4	12.3	0.00E+00	858	17
1g35	8.14	3.7	10.4	0.00E+00	1335	15
1g36	7.17	3.2	8.4	0.00E+00	753	14
1g3b	5.74	3.5	9.5	1.00E-06	476	14
1g3d	5.55	3.7	9.4	4.60E-05	473	15
1g3e	5.38	3.1	10.2	0.00E+00	541	14
1g45	8.64	4.1	13.2	0.00E+00	531	19

1g46	8.8	4.0	13.8	0.00E+00	518	19
1g48	8.41	4.0	13.0	0.00E+00	549	20
1g4j	8.7	4.0	13.0	0.00E+00	516	19
1g4o	8.25	4.1	13.3	0.00E+00	519	20
1g52	9.54	4.1	13.0	0.00E+00	576	20
1g53	9.04	4.1	12.9	0.00E+00	579	20
1g54	8.82	4.0	13.1	0.00E+00	586	20
1g7f	5.47	4.2	8.8	0.00E+00	858	15
1g7g	6.6	4.1	7.1	0.00E+00	956	15
1g7q	6.06	6.4	9.9	5.00E-06	1494	16
1g7v	6.4	4.9	11.4	0.00E+00	798	17
1g98	5.7	8.0	20.4	5.20E-06	497	24
1gaf	8	6.1	10.7	5.25E-04	630	17

1gar	10	5.4	10.1	0.00E+00	1228	18
1gca	6.7	4.2	25.5	0.00E+00	380	28
1gcz	5.13	5.6	11.2	9.61E-04	487	17
1gdo	4.82	8.2	0.0	1.00E+00	362	0
1ghv	4.35	4.8	12.8	0.00E+00	566	17
1ghw	4.2	4.6	13.2	0.00E+00	606	17
1ghy	8.1	4.3	13.2	0.00E+00	615	17
1ghz	4.8	4.2	8.7	1.06E-04	477	14
1gi1	4.77	3.1	9.5	0.00E+00	507	14
1gi4	7.19	3.6	10.1	0.00E+00	456	14
1gi6	5.31	3.5	9.2	0.00E+00	484	14
1gi7	4.51	4.2	11.8	0.00E+00	487	16
1gi8	5.05	4.3	11.6	0.00E+00	489	15

1gi9	5.22	4.1	11.2	0.00E+00	511	15
1gj4	4.07	4.4	12.7	0.00E+00	748	17
1gj5	4.96	4.3	12.9	0.00E+00	721	17
1gj6	6.11	3.3	9.2	0.00E+00	643	14
1gj7	7.89	4.4	11.9	0.00E+00	655	17
1gj8	6.96	4.3	10.7	0.00E+00	631	15
1gj9	7.48	4.3	11.6	0.00E+00	697	16
1gja	5.42	4.3	12.5	0.00E+00	532	16
1gjb	6.35	4.1	11.9	0.00E+00	612	16
1gjc	8.1	4.2	11.5	0.00E+00	613	15
1gjd	5.22	4.1	11.4	0.00E+00	644	16
1gni	8.07	8.3	20.3	0.00E+00	869	25
1gny	4.14	3.4	4.4	2.70E-02	621	9

1gpk	5.37	5.2	20.5	0.00E+00	621	26
1gpn	6.48	5.3	20.9	0.00E+00	569	26
1gpy	4.7	11.8	17.8	2.56E-02	493	20
1grp	3.72	7.5	23.7	0.00E+00	373	28
1gu3	4.32	3.0	4.8	1.00E-06	838	10
1gui	6.28	2.9	5.3	0.00E+00	879	12
1gvu	9	4.6	12.4	0.00E+00	1344	19
1gvw	6.96	4.4	12.1	0.00E+00	1290	18
1gvx	7.22	4.8	11.3	0.00E+00	1446	17
1gwm	4.1	3.7	3.5	7.31E-01	893	8
1gyx	2.48	3.8	7.1	3.15E-04	314	10
1gyy	3.64	3.8	7.4	3.50E-05	371	10
1gz9	3.57	3.5	4.4	2.63E-02	574	8

1gzc	3.28	3.6	5.8	2.08E-03	411	9
1h0a	5.44	4.2	8.3	0.00E+00	581	12
1h1h	5.22	3.7	9.5	0.00E+00	401	14
1h1p	4.92	6.5	11.9	6.27E-04	627	18
1h1s	8.22	6.3	11.1	1.55E-04	822	17
1h22	9.1	5.1	14.7	0.00E+00	1033	25
1h23	8.35	5.3	16.0	0.00E+00	1072	26
1h46	3.57	5.0	17.2	0.00E+00	660	24
1h4n	4.92	3.9	16.0	0.00E+00	314	18
1h4w	4.66	3.2	11.5	0.00E+00	375	15
1h6h	5.3	4.9	6.5	5.31E-02	602	12
1h9z	5.42	8.9	25.7	0.00E+00	701	31
1ha2	5.54	8.7	25.0	0.00E+00	689	29

1hbv	6.37	3.3	9.4	0.00E+00	1255	15
1hef	9	3.6	9.8	0.00E+00	1354	15
1heg	7.74	4.2	10.6	0.00E+00	1210	16
1hfs	8.7	5.5	14.4	0.00E+00	1458	20
1hi3	4.19	3.3	9.2	0.00E+00	540	13
1hi4	4.49	3.1	8.3	0.00E+00	662	13
1hih	8.05	3.5	10.1	0.00E+00	1236	15
1hii	7.28	3.5	10.6	0.00E+00	1244	14
1hiv	9	3.4	9.2	0.00E+00	1495	15
1hk4	5.31	8.6	20.5	0.00E+00	993	24
1hmr	6.55	4.2	11.5	0.00E+00	810	17
1hms	6.37	4.3	11.9	0.00E+00	832	18
1hmt	5.79	4.2	12.2	0.00E+00	801	18

1hn2	6	6.8	28.3	0.00E+00	547	31
1hn4	5.3	6.1	21.1	0.00E+00	1161	27
1hos	8.55	3.6	9.6	0.00E+00	1164	15
1hp0	6.7	5.1	19.2	0.00E+00	570	23
1hpo	9.22	3.7	10.7	0.00E+00	995	15
1hps	9.22	3.4	9.5	0.00E+00	1165	15
1hpv	9.22	3.7	11.0	0.00E+00	1107	15
1hpx	9.3	3.5	9.4	0.00E+00	1221	14
1hsg	9.42	3.5	10.2	0.00E+00	1320	15
1hsh	8.61	3.3	9.5	0.00E+00	1329	15
1hsl	7.19	6.0	17.1	0.00E+00	405	19
1hvh	7.96	4.2	11.6	0.00E+00	1174	16
1hvi	10.08	3.6	9.2	0.00E+00	1472	15

1hvj	10.46	3.7	9.5	0.00E+00	1459	15
1hvk	10.11	3.5	9.4	0.00E+00	1475	15
1hvl	9	3.6	9.3	0.00E+00	1462	15
1hvr	9.51	3.5	10.8	0.00E+00	1241	15
1hvs	10.3	3.3	9.5	0.00E+00	1420	15
1hwr	8.33	3.4	10.9	0.00E+00	910	15
1hxb	9.92	3.5	10.1	0.00E+00	1267	16
1hxw	10.82	3.5	9.4	0.00E+00	1363	15
1hyo	4.07	7.2	23.2	0.00E+00	410	27
1hyx	7.72	6.3	6.0	6.12E-01	946	13
1i1e	5.03	8.6	4.4	9.24E-01	660	7
1i5r	8.52	5.2	14.9	0.00E+00	1410	19
1i7z	6.4	5.9	9.6	2.94E-04	674	15

1i80	6.41	8.0	17.8	1.33E-03	352	21
1i9n	8.66	4.1	13.7	0.00E+00	528	19
1i9p	8.41	4.2	13.0	0.00E+00	590	19
1icj	2.22	7.3	9.2	8.80E-02	805	14
1if7	10.52	4.1	11.8	0.00E+00	650	19
1if8	9.64	4.3	11.8	0.00E+00	633	19
1igb	6.4	5.4	10.9	2.46E-05	566	14
1igj	10	6.1	8.4	8.24E-03	697	17
1ii5	6.62	6.2	17.3	6.00E-07	384	20
1iih	2.89	5.4	13.4	2.00E-07	410	17
1ik4	7.41	8.4	18.6	5.93E-04	358	22
1ikt	3.4	3.6	9.4	0.00E+00	957	16
1imx	3.52	2.8	3.7	3.57E-03	556	7

1is0	7	3.0	5.4	0.00E+00	841	8
1it6	8.39	3.7	6.3	0.00E+00	1356	10
1iup	2.54	6.2	6.7	3.72E-01	299	8
1ivp	7.52	3.7	9.2	0.00E+00	1527	15
1iy7	6.19	3.9	11.1	1.40E-06	572	15
1izh	7.7	3.4	9.4	0.00E+00	1369	14
1izi	6.59	3.3	9.2	0.00E+00	1314	15
1j01	6.47	3.4	12.1	0.00E+00	478	17
1j14	4.49	3.3	11.3	0.00E+00	368	14
1j16	3.84	3.6	11.8	0.00E+00	375	14
1j17	5.22	3.9	8.0	6.00E-07	840	15
1j4r	7.72	2.9	7.1	0.00E+00	827	12
1jao	5.92	3.2	7.3	0.00E+00	695	13

1jaq	4.48	3.1	7.1	0.00E+00	620	11
1jcx	5.15	5.8	18.1	0.00E+00	723	24
1jd5	6.62	2.9	5.3	0.00E+00	951	10
1jet	7.25	5.6	24.0	0.00E+00	909	30
1jeu	6.82	6.6	23.5	0.00E+00	1042	28
1jev	6.89	6.5	25.0	0.00E+00	1117	30
1jgl	8.7	6.1	9.3	3.26E-03	567	14
1jkx	4.7	5.8	12.0	0.00E+00	1309	16
1jlx	5.55	6.5	8.0	1.01E-01	659	10
1jn2	4.09	4.0	2.2	9.96E-01	414	6
1jn4	4.95	3.1	7.6	0.00E+00	810	11
1joc	4.62	5.6	6.5	2.42E-01	421	8
1jq8	6	3.3	6.8	0.00E+00	1046	13

1jq9	8.45	3.4	6.1	0.00E+00	1113	11
1jqd	5.16	4.3	13.6	0.00E+00	782	19
1jqy	4.92	5.5	6.7	1.06E-01	691	10
1jr0	4.92	5.4	5.2	5.81E-01	579	8
1jt1	3.4	3.8	10.9	0.00E+00	402	14
1jwt	7.85	4.7	12.7	0.00E+00	917	18
1jyq	8.7	3.6	4.2	4.38E-02	1038	7
1jys	3.52	4.9	12.4	9.84E-05	373	16
1jzs	6.6	7.7	19.8	0.00E+00	954	27
1k1i	6.58	3.1	8.6	0.00E+00	631	14
1k1j	7.55	3.3	8.1	0.00E+00	737	14
1k1l	6.9	3.5	8.3	0.00E+00	701	14
1k1m	7.4	3.5	8.0	4.00E-07	726	13

1k1n	6.82	3.1	7.9	0.00E+00	761	14
1k1y	3.22	6.0	19.2	0.00E+00	988	26
1k21	8.38	4.5	11.5	0.00E+00	906	17
1k22	8.4	4.3	11.5	0.00E+00	846	17
1k4g	5.85	5.1	17.5	0.00E+00	681	21
1k4h	5.11	5.7	18.4	0.00E+00	686	21
1k6c	7.48	4.1	10.4	0.00E+00	1359	15
1k6p	7.36	3.6	10.5	0.00E+00	1417	15
1k6t	7.62	4.0	10.5	0.00E+00	1387	15
1k6v	6.92	3.9	10.3	0.00E+00	1419	15
1k9s	6.52	4.0	10.6	0.00E+00	613	14
1kav	5.82	4.5	7.6	5.18E-05	607	14
1kc7	5.52	8.5	23.2	1.18E-03	321	26

1kdk	9.05	4.3	0.0	1.00E+00	665	0
1kel	7.28	6.3	8.7	9.75E-03	637	15
1kf0	2.55	6.5	16.8	0.00E+00	764	20
1kll	5.2	4.5	4.5	4.98E-01	400	8
1koj	6.7	8.0	21.7	0.00E+00	515	25
1kpm	5.8	3.4	6.5	0.00E+00	975	13
1kr3	5	3.8	9.0	0.00E+00	581	14
1ksn	9.4	4.6	8.9	0.00E+00	939	17
1kts	8.35	4.1	11.2	0.00E+00	982	18
1kug	3.8	3.5	8.5	0.00E+00	786	14
1kui	3.77	3.5	8.3	0.00E+00	786	13
1kuk	3.91	3.4	7.5	0.00E+00	823	13
1kv1	5.94	5.3	20.0	0.00E+00	774	24

1kv5	4.22	5.1	13.9	6.80E-06	351	18
1kyv	5.92	6.9	14.9	0.00E+00	814	18
1kzk	10.39	3.6	10.8	0.00E+00	1211	15
1kzn	8.92	4.2	7.2	6.80E-06	963	17
1l2s	4.59	4.2	12.4	0.00E+00	517	14
1l5q	3.97	11.4	26.7	0.00E+00	823	33
1l5r	4.77	10.0	30.7	0.00E+00	826	35
1l5s	3.26	11.2	25.5	0.00E+00	815	32
1l7s	9.52	6.0	10.6	1.06E-05	655	16
1l7x	4.04	9.9	30.0	0.00E+00	828	35
1l83	3.4	3.5	15.8	0.00E+00	285	19
1l8b	6.85	4.0	7.5	6.00E-07	699	11
1laf	7.85	4.4	17.7	0.00E+00	444	20

1lag	6.3	4.0	17.3	0.00E+00	385	20
1lah	7.52	3.9	18.8	0.00E+00	360	23
1lan	7.22	12.5	47.0	0.00E+00	373	51
1lbf	7.85	3.8	12.1	0.00E+00	743	17
1lbl	7.85	4.1	11.1	0.00E+00	687	16
1lcp	6.64	12.9	47.4	0.00E+00	405	50
1lee	7.74	5.1	14.0	0.00E+00	1142	20
1lf2	7.52	4.8	13.7	0.00E+00	1135	18
1lf9	12	4.8	13.4	0.00E+00	791	21
1lgt	6.1	4.1	15.5	0.00E+00	485	20
1lgw	4	3.5	0.0	1.00E+00	329	0
1li2	4.04	3.3	0.0	1.00E+00	301	0
1li3	4.25	3.4	0.0	1.00E+00	348	0

1li6	3.8	3.3	0.0	1.00E+00	298	0
1lke	6.53	4.6	13.7	0.00E+00	809	20
1lkk	6.85	3.2	5.0	0.00E+00	936	9
1lkl	5.81	3.1	4.4	1.38E-04	832	7
1lnm	8.7	4.9	13.8	0.00E+00	786	22
1loq	3.7	6.1	11.9	7.00E-06	616	15
1lor	11.06	5.4	19.0	0.00E+00	673	24
1los	7.19	4.8	13.2	0.00E+00	642	17
1lox	5.52	6.4	22.5	0.00E+00	672	26
1lpg	7.09	4.6	8.9	0.00E+00	987	16
1lpk	7.55	4.3	9.8	0.00E+00	866	17
1lqe	5.82	3.7	8.0	0.00E+00	796	14
1lrh	6.82	5.7	14.1	0.00E+00	524	18

1lyb	11.42	7.0	13.5	0.00E+00	1319	19
1lyx	4.54	6.8	16.5	1.80E-06	361	19
1lzq	8.39	3.9	9.6	0.00E+00	1359	15
1m0n	2.22	6.8	25.9	0.00E+00	775	30
1m0o	2.31	10.4	27.6	0.00E+00	774	32
1m0q	2.96	9.5	27.7	0.00E+00	724	34
1m13	7.57	5.6	20.0	0.00E+00	1193	25
1m1b	4.66	6.0	0.0	1.00E+00	348	0
1m21	6.52	5.0	12.8	0.00E+00	959	20
1m2p	6.11	4.9	13.2	0.00E+00	610	19
1m2q	6.1	4.9	14.4	0.00E+00	570	19
1m2r	6.46	4.9	13.2	0.00E+00	593	19
1m2x	4.15	3.6	9.4	0.00E+00	434	13

1m48	5.09	3.6	4.4	4.89E-02	719	8
1m4h	9.52	5.1	11.6	0.00E+00	1602	19
1m5j	3.52	4.5	6.1	4.42E-02	354	9
1m5w	4.27	5.4	21.2	0.00E+00	498	25
1m6p	5.1	5.2	7.1	4.08E-02	488	9
1m7d	6.23	6.0	9.1	2.00E-03	828	14
1m7i	5.4	6.0	7.9	1.31E-02	1073	14
1m9n	6.92	9.3	14.1	1.96E-02	671	17
1mai	6.68	3.0	6.6	0.00E+00	513	10
1me8	7.19	7.7	26.4	0.00E+00	623	31
1mes	7.7	3.9	10.8	0.00E+00	1124	15
1met	9.4	3.5	11.1	0.00E+00	1128	15
1meu	6.1	3.3	10.8	0.00E+00	1151	15

1mfa	5.04	3.7	5.3	1.33E-03	594	10
1mfd	5.31	5.9	5.3	7.73E-01	620	10
1mfi	5.59	5.6	10.6	2.42E-03	414	14
1mfl	3.89	3.2	5.6	0.00E+00	745	11
1mh5	9.21	5.2	11.9	0.00E+00	834	21
1mj7	8.35	5.9	10.1	5.60E-06	837	19
1mjj	8.74	5.4	10.8	2.00E-07	841	20
1mmp	6.07	3.4	6.5	0.00E+00	715	14
1mmq	7.52	3.2	6.6	0.00E+00	742	13
1mmr	5.4	3.3	7.1	0.00E+00	777	13
1moq	3.46	7.0	14.8	7.60E-04	555	18
1mq5	9	4.6	8.6	0.00E+00	939	16
1mq6	11.15	4.7	8.5	6.00E-07	947	18

1mtr	8.4	4.1	10.1	0.00E+00	1239	15
1mx1	4	5.7	21.1	0.00E+00	575	25
1n1m	5.7	11.1	36.5	0.00E+00	456	40
1n1t	3.85	5.8	12.8	4.00E-07	636	17
1n2v	4.08	4.8	17.1	0.00E+00	509	22
1n3i	8.89	7.5	20.2	2.00E-07	584	25
1n43	10.55	5.7	14.3	0.00E+00	549	19
1n46	10.52	6.9	20.7	0.00E+00	817	25
1n4h	6.55	4.7	16.8	0.00E+00	858	21
1n4k	10.05	5.6	9.2	4.59E-04	569	13
1n5r	5.66	5.7	7.6	2.64E-02	714	12
1n9m	10.96	6.8	15.2	0.00E+00	568	20
1nc1	6.12	4.2	13.1	0.00E+00	620	17

1nc3	6.12	4.7	14.1	0.00E+00	566	17
1nc9	11.17	6.2	15.1	0.00E+00	576	19
1ndj	12.16	6.0	15.2	0.00E+00	547	19
1nf8	7.82	4.7	18.8	0.00E+00	523	22
1nfu	7.74	4.6	8.4	0.00E+00	792	17
1nfw	8.96	4.7	9.5	2.00E-07	825	17
1nfx	8.52	4.8	9.3	0.00E+00	846	18
1nfy	8.89	4.8	9.6	0.00E+00	836	18
1nhu	5.66	6.8	6.7	4.87E-01	635	12
1niu	7.1	8.1	24.4	0.00E+00	720	29
1nja	6.31	6.5	19.8	0.00E+00	572	25
1njc	5.55	6.5	19.6	0.00E+00	554	24
1nje	3.8	6.7	20.0	0.00E+00	542	24

1njj	2.1	5.2	8.2	1.33E-03	739	12
1njs	7.82	4.0	10.0	0.00E+00	976	15
1nl9	5.96	4.3	8.8	0.00E+00	860	13
1nli	3.59	3.7	13.4	0.00E+00	358	17
1nm6	10.05	4.5	9.3	0.00E+00	903	14
1nms	6.7	5.6	10.7	8.60E-06	805	16
1nny	7.66	4.1	8.4	0.00E+00	1177	14
1no6	7.41	4.3	10.2	0.00E+00	604	13
1nq7	6.8	4.7	16.8	0.00E+00	978	24
1nt1	8.89	4.3	10.2	0.00E+00	862	16
1nu3	4	5.1	13.8	0.00E+00	475	17
1nvq	8.25	5.2	12.1	0.00E+00	895	18
1nvr	8.11	5.1	11.8	0.00E+00	877	18

1nvs	7.82	5.2	12.3	0.00E+00	787	18
1nw5	5.21	6.1	13.2	0.00E+00	826	18
1nw7	5.09	6.0	12.2	0.00E+00	748	14
1nwl	2.39	4.2	6.8	1.20E-06	787	12
1nz7	7.12	4.4	9.3	0.00E+00	1099	14
1o0f	5.3	3.2	7.3	0.00E+00	560	12
1o0m	5.15	3.2	9.3	0.00E+00	470	13
1o0n	4.09	3.4	9.7	0.00E+00	468	12
1o0o	5.1	3.2	7.4	0.00E+00	427	12
1o1s	7.31	7.1	27.4	0.00E+00	1022	34
1o2g	6.12	4.4	11.0	0.00E+00	765	16
1o2h	6.15	3.6	8.7	0.00E+00	636	14
1o2k	5.7	3.1	8.6	0.00E+00	749	13

1o2n	4.38	3.5	9.3	0.00E+00	705	14
1o2o	5	3.8	8.8	0.00E+00	749	14
1o2p	4.85	3.5	8.9	0.00E+00	760	13
1o2q	5.64	3.5	8.8	0.00E+00	667	14
1o2r	5.21	3.3	8.4	0.00E+00	690	13
1o2t	6.89	3.4	9.2	0.00E+00	646	14
1o2x	5.12	3.5	8.6	2.00E-07	620	14
1o2z	5.1	3.7	8.0	0.00E+00	726	14
1o30	5.54	3.3	8.4	0.00E+00	700	14
1o35	5	3.3	8.9	0.00E+00	512	14
1o36	5.07	3.4	8.3	0.00E+00	726	13
1o3c	5.3	3.3	8.9	0.00E+00	626	14
1o3g	6.72	3.3	8.6	0.00E+00	628	14

1o3h	6.34	3.5	8.9	2.00E-07	592	13
1o3k	5.57	3.3	8.8	0.00E+00	581	14
1o3l	6.54	3.5	9.1	0.00E+00	609	14
1o3p	6.66	4.0	11.5	0.00E+00	588	16
1o86	9.57	7.5	27.8	0.00E+00	844	32
1o8b	2.68	5.8	12.6	2.00E-06	457	15
1o9d	5.6	4.7	11.3	0.00E+00	973	16
1oai	5	2.4	3.5	1.34E-05	808	9
1obx	5.72	2.7	5.3	0.00E+00	621	8
1ocq	5.19	3.4	7.9	0.00E+00	550	12
1ody	8.1	3.5	9.3	0.00E+00	1520	15
1oe7	5.52	7.2	16.8	0.00E+00	646	23
1oe8	5.52	7.5	17.3	0.00E+00	630	23

1ogx	6.09	4.5	12.6	0.00E+00	596	18
1ohr	8.7	3.3	10.5	0.00E+00	1146	15
1oif	7.72	4.4	20.0	0.00E+00	353	22
1oim	5.32	4.4	20.6	0.00E+00	366	24
1okl	6.03	3.9	14.5	0.00E+00	511	18
1okn	8.64	4.0	11.4	0.00E+00	601	18
1oko	4.54	5.1	3.7	8.98E-01	339	6
1ony	6.77	4.3	9.7	0.00E+00	935	15
1onz	5.1	4.3	10.4	0.00E+00	615	14
1ork	8.81	6.6	15.7	0.00E+00	1077	22
1os0	6.03	4.3	12.0	0.00E+00	891	19
1oss	4.79	3.7	10.7	2.00E-07	363	15
1ow4	5.68	5.1	15.2	0.00E+00	670	21

1oxn	5.68	6.8	3.7	9.96E-01	635	7
1oxq	6.3	6.7	4.6	9.61E-01	672	8
1oyq	6.96	3.3	7.7	0.00E+00	780	14
1oyt	7.24	4.9	12.3	0.00E+00	860	18
1oz0	7.7	9.9	17.5	1.00E-06	1381	23
1p6d	2.94	3.1	8.0	0.00E+00	750	14
1p6e	2.92	3.1	8.0	0.00E+00	782	14
1pa9	4.6	3.7	8.0	3.40E-06	436	12
1pb8	5.15	5.3	16.7	0.00E+00	268	18
1pb9	3.62	5.2	17.0	0.00E+00	265	18
1pbk	9.05	3.3	7.2	0.00E+00	933	14
1pbq	6.27	7.1	18.7	0.00E+00	529	22
1pdz	3.7	6.6	19.8	2.00E-05	307	23

1pgp	5.7	7.7	23.7	0.00E+00	540	26
1ph0	6.92	4.4	9.6	0.00E+00	1004	14
1pip	5	3.5	5.3	1.13E-04	728	10
1pme	9.4	5.2	15.5	0.00E+00	751	20
1pot	5.49	4.5	24.0	0.00E+00	446	29
1ppc	6.16	3.5	8.7	0.00E+00	751	14
1pph	5.92	3.3	8.2	0.00E+00	638	13
1ppi	5.01	4.6	14.4	0.00E+00	1142	22
1ppk	7.66	3.9	12.9	0.00E+00	1034	16
1ppl	8.55	4.0	12.6	0.00E+00	1186	18
1ppm	5.8	3.9	12.0	0.00E+00	1166	18
1pr1	5.3	6.3	14.8	0.00E+00	567	19
1pr5	3.92	6.9	15.6	0.00E+00	573	20

1pro	11.3	3.9	11.6	0.00E+00	1154	16
1ps3	2.28	7.2	17.3	7.53E-04	420	20
1pvn	9.3	7.4	0.0	1.00E+00	679	0
1pxh	8.74	4.4	6.5	4.00E-07	1047	14
1pyn	5.49	4.3	9.5	0.00E+00	980	15
1pz5	5.4	5.7	7.7	3.75E-03	1158	14
1q54	5.85	5.4	15.1	0.00E+00	582	20
1q8t	4.76	5.5	19.8	0.00E+00	648	26
1q8u	5.96	5.5	20.2	0.00E+00	677	26
1q8w	5.24	5.6	19.5	0.00E+00	630	25
1qan	4.48	4.3	9.0	0.00E+00	844	14
1qaw	5.12	6.0	12.3	3.02E-04	502	16
1qb1	6.77	3.3	7.3	0.00E+00	812	14

1qb6	6.06	3.6	7.9	0.00E+00	693	14
1qb9	7.44	3.3	7.4	0.00E+00	834	14
1qbn	5.85	3.2	8.2	0.00E+00	747	14
1qbo	7.74	3.6	7.6	0.00E+00	818	14
1qbq	8.3	6.6	27.4	0.00E+00	751	35
1qbr	10.57	3.8	10.0	0.00E+00	1342	15
1qbs	9.47	4.3	11.4	0.00E+00	1123	16
1qbt	10.62	3.4	9.4	0.00E+00	1422	15
1qbu	10.24	3.8	10.9	0.00E+00	1184	15
1qbv	5.39	4.4	10.7	0.00E+00	805	15
1qca	5.27	7.6	14.4	0.00E+00	990	21
1qf0	7.38	3.8	12.7	0.00E+00	914	19
1qf1	7.32	3.9	13.4	0.00E+00	788	19

1qf2	5.92	3.8	13.0	0.00E+00	742	19
1qft	8.77	4.2	13.4	2.00E-07	343	16
1qhc	7.57	3.1	8.2	0.00E+00	971	13
1qi0	2.35	3.4	5.3	1.32E-03	481	10
1qin	8	5.0	11.4	0.00E+00	993	18
1qiw	7.74	4.7	12.6	0.00E+00	1151	18
1qjb	6.38	7.0	17.8	0.00E+00	1169	24
1qji	4.85	3.2	9.2	0.00E+00	1065	16
1qk3	5.15	9.9	10.7	2.99E-01	686	14
1qk4	4.21	9.3	13.7	2.00E-02	610	18
1qka	5.92	6.3	22.4	0.00E+00	1100	26
1qkb	7.35	5.8	23.5	0.00E+00	955	30
1qpb	1.36	8.5	26.0	0.00E+00	896	33

1qq9	2.06	3.1	9.8	0.00E+00	406	12
1qsc	3.68	8.3	4.8	1.00E+00	957	8
1qxk	5.05	4.2	9.0	0.00E+00	943	13
1qy5	6.7	4.6	14.1	0.00E+00	637	19
1r0p	8	5.3	13.2	0.00E+00	860	20
1rbo	10.55	5.3	12.0	0.00E+00	672	16
1rbp	6.72	4.6	15.5	0.00E+00	839	23
1rdi	2.06	4.1	2.4	9.51E-01	216	4
1rdj	1.66	3.7	2.2	9.04E-01	199	3
1rdl	2.24	3.8	2.3	8.73E-01	208	3
1rdn	1.84	3.8	2.4	8.86E-01	283	3
1rgk	4.31	2.6	5.0	2.00E-07	449	8
1rpj	6.48	4.0	16.1	0.00E+00	358	18

1sbg	7.74	3.8	10.7	0.00E+00	1181	15
1siv	8.08	3.7	9.8	0.00E+00	1123	15
1sld	6.57	6.9	11.5	0.00E+00	937	18
1sle	6.17	6.6	11.3	2.00E-07	887	19
1slg	3.9	6.3	11.3	0.00E+00	1086	19
1sln	6.64	3.5	7.9	0.00E+00	829	13
1srg	5.3	5.6	14.3	0.00E+00	592	19
1sri	6.08	5.8	15.0	0.00E+00	618	19
1stc	8.1	5.0	17.7	0.00E+00	878	23
1str	4.77	6.6	11.8	0.00E+00	953	19
1sts	5	6.9	11.4	2.00E-06	938	18
1swg	7.36	5.8	12.8	2.00E-07	529	16
1swk	12	6.2	15.5	0.00E+00	548	19

1swn	12	6.9	15.0	0.00E+00	577	19
1swp	11	6.2	13.6	0.00E+00	554	19
1swr	6.92	6.0	14.6	0.00E+00	550	19
1tcw	6.02	3.8	8.0	0.00E+00	918	14
1tcx	6.95	3.5	10.8	0.00E+00	1174	15
1tet	6.08	5.7	4.8	6.46E-01	291	7
1thl	6.25	6.5	18.7	0.00E+00	765	24
1tkb	6.1	8.3	27.3	0.00E+00	844	37
1tlp	7.55	3.9	12.4	0.00E+00	816	18
1tmn	7.3	4.1	12.2	0.00E+00	833	18
1tmt	6.24	4.9	12.4	0.00E+00	916	19
1tng	2.93	3.5	10.3	0.00E+00	350	13
1tnh	3.37	3.2	10.6	0.00E+00	351	13

1tni	4	3.6	9.6	8.00E-07	400	13
1tnj	1.96	3.5	11.5	0.00E+00	368	14
1tnk	1.49	3.6	11.0	2.00E-07	390	14
1tnl	1.88	3.5	10.9	2.00E-07	377	14
1tom	8.3	4.1	10.0	0.00E+00	784	13
1trd	5.4	4.4	11.1	4.98E-05	392	14
1tsl	6.15	6.9	13.3	4.98E-05	716	18
1tyr	7	3.9	5.6	2.69E-03	541	9
1ugx	5.91	6.5	4.7	9.32E-01	531	8
1uio	4.35	4.5	16.4	0.00E+00	554	20
1umw	6.55	4.2	5.8	4.13E-04	1019	11
1upf	4.6	8.9	13.8	1.01E-01	303	15
1usn	7.74	3.3	7.4	0.00E+00	602	11

1uvt	7.64	4.6	10.5	0.00E+00	821	17
1vfn	5.6	7.3	19.1	3.80E-06	350	22
1vot	6.6	4.9	20.1	0.00E+00	612	24
1vwf	5.54	6.9	9.7	3.37E-04	998	17
1vwl	5.63	5.7	13.8	0.00E+00	887	18
1vwn	5.82	6.3	11.3	0.00E+00	930	19
1wdn	6.3	4.2	16.6	0.00E+00	374	20
1wht	3.7	6.4	19.1	0.00E+00	483	22
1xka	6.88	5.2	9.5	2.80E-06	824	16
1xug	7.05	3.3	8.5	0.00E+00	606	14
1yda	6.55	3.9	14.1	0.00E+00	461	19
1ydb	8.24	4.0	14.2	0.00E+00	438	18
1ydd	7.07	4.1	14.9	0.00E+00	469	19

1ydr	5.52	5.1	18.1	0.00E+00	619	22
1yds	5.92	5.0	19.0	0.00E+00	597	24
1ydt	7.32	5.1	18.2	0.00E+00	895	24
1yei	7.46	5.7	11.8	1.00E-06	725	19
1yej	7.46	5.7	10.2	2.60E-05	814	19
1zsb	0.6	3.7	14.0	0.00E+00	471	18
2aac	2.22	4.2	12.2	0.00E+00	383	15
2ada	13	4.2	16.2	0.00E+00	555	21
2adm	5.7	5.7	11.8	0.00E+00	808	15
2amv	8.8	9.1	11.3	9.48E-02	724	14
2ans	5.92	5.8	17.5	0.00E+00	672	25
2bpv	7.67	3.5	9.7	0.00E+00	1290	14
2bpy	7.4	3.4	9.9	0.00E+00	1314	15

2bza	2.8	3.6	11.2	0.00E+00	344	14
2cgr	7.28	6.1	8.6	1.10E-02	787	15
2cht	5.52	4.5	11.5	4.44E-05	448	14
2csn	4.41	7.3	25.0	0.00E+00	592	28
2ctc	3.89	3.7	12.2	0.00E+00	434	16
2drc	9.89	5.7	16.9	0.00E+00	828	24
2dri	6.89	3.8	17.5	0.00E+00	309	19
2er6	7.22	4.5	11.0	0.00E+00	1462	18
2er9	7.4	4.5	11.7	0.00E+00	1528	17
2fgi	7.34	7.0	13.8	0.00E+00	966	20
2fmb	8.7	4.3	11.0	0.00E+00	1510	16
2gss	4.94	6.0	11.0	1.43E-03	527	17
2gst	6.07	7.2	18.3	0.00E+00	930	26

2h4n	8.7	3.8	14.2	0.00E+00	463	19
2izl	6	6.1	15.2	0.00E+00	573	20
2jxr	7.05	4.6	13.6	0.00E+00	1205	19
2mas	4.52	5.2	16.3	0.00E+00	515	20
2olb	5.54	5.8	23.9	0.00E+00	1039	30
2pcp	8.7	5.7	10.5	1.16E-05	609	15
2pri	2.91	12.0	18.0	4.19E-02	508	21
2qwb	2.74	5.5	13.0	0.00E+00	594	17
2qwc	3.55	5.5	13.2	2.00E-07	591	17
2qwd	4.85	5.6	13.1	4.00E-07	591	17
2qwe	7.48	5.4	13.3	0.00E+00	638	17
2qwf	5.67	5.6	12.5	0.00E+00	704	18
2qwg	8.4	5.8	14.0	1.20E-06	636	18

2rkm	3.9	6.2	21.9	0.00E+00	780	26
2sim	3.42	4.5	13.0	0.00E+00	587	16
2std	9.85	6.0	0.0	1.00E+00	747	0
2tct	9	6.8	15.8	0.00E+00	871	21
2tmn	5.89	6.6	21.3	0.00E+00	505	27
2tpi	4.31	4.3	8.9	0.00E+00	589	13
2usn	6.51	3.2	7.6	0.00E+00	686	11
2ypi	4.82	6.4	12.0	1.78E-02	367	16
3aid	6.86	4.6	11.2	0.00E+00	1162	15
3amv	7.97	11.8	17.5	8.01E-03	851	22
3er3	7.09	4.3	11.4	0.00E+00	1389	18
3gss	5.82	5.9	13.5	0.00E+00	953	24
3gst	6.72	6.8	19.3	0.00E+00	925	25

3jdw	3.6	6.3	18.7	0.00E+00	426	22
3kiv	4.7	2.2	3.6	2.31E-04	355	6
3mag	4.07	6.6	6.6	4.89E-01	377	9
3mbp	6.8	4.9	14.6	0.00E+00	874	20
3mct	4.07	6.6	7.6	2.93E-01	352	10
3pcb	2.4	9.6	23.4	1.79E-02	371	24
3pcc	3.62	9.3	24.0	8.56E-03	373	25
3pce	2	9.6	25.1	1.74E-05	410	28
3pcf	6.05	9.4	25.6	1.22E-05	375	27
3pcg	2.3	9.7	25.8	1.34E-05	399	28
3pch	5.4	9.6	25.3	1.50E-04	369	28
3pcj	7.22	9.6	25.9	1.16E-05	380	28
3pck	6.7	9.5	25.0	9.42E-05	384	27

3pcn	3.66	9.7	25.2	1.08E-04	423	27
3std	11.11	7.2	20.1	0.00E+00	802	24
3tmk	6.87	6.2	12.3	0.00E+00	1332	24
43ca	6	3.8	13.0	0.00E+00	348	15
456c	9.77	5.3	9.2	3.12E-05	724	14
4apr	6.7	4.3	13.0	0.00E+00	1234	18
4er1	6.62	4.5	11.7	0.00E+00	1402	17
4er2	9.3	4.3	11.3	0.00E+00	1228	18
4fiv	6.52	3.3	9.0	0.00E+00	1672	15
4mbp	5.64	5.1	13.2	0.00E+00	1077	20
4rsk	4.32	3.5	9.7	0.00E+00	468	13
4sga	7.3	2.6	7.5	0.00E+00	752	11
4std	10.33	6.1	0.0	1.00E+00	756	0

4tim	2.16	5.5	13.9	9.60E-06	408	18
4tln	3.72	4.1	14.6	0.00E+00	422	18
4tmk	7.7	4.4	11.7	0.00E+00	1216	20
4tmn	10.17	6.5	21.3	0.00E+00	1013	27
4ts1	4.94	6.2	24.8	0.00E+00	436	28
5abp	6.64	3.8	18.8	0.00E+00	396	21
5apr	7.77	4.6	12.7	0.00E+00	1363	17
5er1	6.02	4.7	10.1	0.00E+00	986	18
5er2	6.57	4.3	10.7	0.00E+00	1571	17
5hvp	7.7	3.4	9.2	0.00E+00	1356	14
5std	10.49	7.2	19.9	0.00E+00	833	25
5tln	6.37	4.0	13.6	0.00E+00	658	18
5tmn	8.04	6.6	21.5	0.00E+00	921	29

5tmp	7.47	4.5	11.3	0.00E+00	1398	20
5upj	7.12	3.8	11.9	0.00E+00	714	15
5yas	3.26	3.9	19.1	0.00E+00	345	22
6abp	6.36	4.0	19.6	0.00E+00	356	22
6apr	7.77	4.2	12.5	0.00E+00	1253	19
6cpa	11.52	4.0	8.7	4.00E-06	761	18
6rnt	2.37	2.6	5.2	0.00E+00	491	8
6std	8.64	6.9	20.9	0.00E+00	790	25
6tim	3.21	5.8	14.6	1.40E-06	395	18
6upj	6.32	3.8	11.8	0.00E+00	700	15
7abp	6.46	4.1	19.1	0.00E+00	385	21
7cpa	13.96	3.9	8.1	3.40E-06	835	16
7hvp	9.62	3.8	9.1	0.00E+00	1547	14

7kme	4.4	4.2	10.7	0.00E+00	1006	17
7std	10.72	6.8	20.1	0.00E+00	743	24
7upj	8.49	4.3	10.3	0.00E+00	983	15
830c	9.28	3.8	8.5	0.00E+00	755	13
8abp	8	3.9	18.6	0.00E+00	393	20
8cpa	9.15	4.0	9.7	2.00E-07	780	17
8hvp	9	3.9	8.7	0.00E+00	1522	14
966c	7.64	3.4	8.0	0.00E+00	701	12
9abp	8	3.8	17.1	0.00E+00	405	19