Published in final edited form as:

Proteins. 2016 February ; 84(2): 201–216. doi:10.1002/prot.24966.

An Amino Acid Code to Define a Protein's Tertiary Packing Surface

Keith J. Fraga, Hyun Joo, and Jerry Tsai*

Department of Chemistry, University of the Pacific 3601 Pacific Ave, 95211

Abstract

One difficult aspect of the protein-folding problem is characterizing the non-specific interactions that define packing in protein tertiary structure. To better understand tertiary structure, this work extends the knob-socket model by classifying the interactions of a single knob residue packed into a set of contiguous sockets, or a pocket made up of 4 or more residues. The knob-socket construct allows for a symbolic two-dimensional mapping of pockets. The two-dimensional mapping of pockets provides a simple method to investigate the variety of pocket shapes in order to understand the geometry of protein tertiary surfaces. The diversity of pocket geometries can be organized into groups of pockets that share a common core, which suggests that some interactions in pockets are ancillary to packing. Further analysis of pocket geometries displays a preferred configuration that is right-handed in α -helices and left-handed in β -sheets. The amino acid composition of pockets illustrates the importance of non-polar amino acids in packing as well as position specificity. As expected, all pocket shapes prefer to pack with hydrophobic knobs; however, knobs are not selective for the pockets they pack. Investigating side-chain rotamer preferences for certain pocket shapes uncovers no strong correlations. These findings allow a simple vocabulary based on knobs and sockets to describe protein tertiary packing that supports improved analysis, design and prediction of protein structure.

Keywords

protein packing; protein tertiary structure; non-specific protein interactions; knob-socket analysis; packing pocket

INTRODUCTION

The manner in which a protein's amino acid sequence encodes its three-dimensional (3D) structure remains a challenge in structural biology.^{1–9} One step towards addressing this problem is to understand the rules for predicting how amino acids non-local in sequence pack together locally in space. It is well established that the hydrophobic effect is the driving force for protein folding.^{10,11} Resulting from increased solvent entropy,^{11,12} the burial of hydrophobic side-chains produces non-specific packing interactions. While our understanding of how the regular patterns of backbone hydrogen bonds restrict torsion angles to produce the fundamental principles of secondary structure,^{5,13–16} corresponding

*corresponding author: jtsai@pacific.edu.

patterns in tertiary structure have been difficult to codify due to the hydrophobic driven non-specific packing of residues.¹⁷ To address this problem, this work characterizes the higher order patterns of residue packing in protein tertiary structure based on the knob-socket model.^{18–20}

Previous investigations of protein packing have made significant advances in our understanding of protein structure. Many have been efforts to analyze and predict packing between regular elements of secondary structures.^{21–25} The most notable examples primarily describe the packing between helical secondary structures such as the helical wheel,²⁶ knobs-into-holes,^{27–29} and ridges-into-grooves.²⁵ While these approaches identify canonical residue packing patterns, their applicability is limited to design and prediction of α -helices. The formative work by Chothia *et. al.* and Janin *et. al.* are some of the first approaches that attempted to generalize the packing in proteins.^{23–25} More recently, protein tertiary structure has been analyzed using pairwise amino acid contacts.^{30,31} Such efforts have been formative in offering new perspectives on the nature of protein structure. We seek to build off this work to consider the topology of tertiary packing surfaces.

As an abstraction of protein tertiary structure, the knob-socket model has not only been able to identify the canonical packing patterns between regular secondary structure,^{18,19} but also accurately describes the general packing between all types of secondary structures, including the non-repetitive coil elements.²⁰ The 2 motifs of the knob-socket model provide a basic unit of protein packing that directly relates amino acid composition to 3D configuration (Figure 1). These motifs decompose the complicated residue interaction networks into simpler representations, where packing between secondary structures is projected onto the two-dimensional (2D) lattices in terms of the knobs packing into sockets. In many instances, a knob packs into several contiguous sockets, which is designated as a pocket of 4 or more residues. In the literature, there are references to notions of pockets in protein structure^{32–34}. Much of the literature concerning protein pockets defines pockets in terms of protein-protein interaction surfaces. This work is significantly different because pockets are defined in terms of protein tertiary structure, not quaternary structure. The goal of this work is to provide a qualitative assessment of the manner in which the knob-socket model can organize and describe protein tertiary packing surfaces. With the knob-socket model, the diversity of pocket geometries is surveyed that are possible in α -helix, β -sheet, coil, and turn. The propensities for amino acid sequences to form specific pocket geometries identified through the knob-socket analysis are computed. Further, the relationships of knob amino acid composition and knob side-chain rotamer distributions with pocket geometry are investigate. By abstracting protein tertiary structure, the knob-socket analysis can simplify the representation of the complex packing interactions and furthermore, provide unique insights on protein tertiary structures, and folding mechanisms.

MATERIALS AND METHODS

Knob-Socket Model

As characterized previously for each type of secondary structure,^{18–20} the knob-socket model describes the tetrahedral packing of two motifs: a three residue socket and a single residue knob packing into the socket (Figure 1). A socket is comprised of three residues on

the same secondary structure that are mostly local in sequence and all contact each other. Beginning from the atomic coordinates of a protein structure, a Voronoi polyhedra³⁵ calculation³⁶ is performed on the heavy atoms. A 6 Å cutoff was used to determine atomic contacts with the Voronoi scheme based on atomic radii and volumes developed from Tsai et. al.^{17–19,37} So, contacts between heavy atoms are those that share a polyhedral face, which defines a Delaunay tessellation³⁸ between atoms. From these atom contacts, residue contacts are collated to create an interaction network. Building the residues contact graph from atom contacts helps to avoid errors that occur from almost-Delaunay simplices.³⁹ Cliques are identified using the maximal clique method⁴⁰, and classified according to the residue contact order⁴¹ based on secondary structure. As shown consistently in previous analysis across all types of secondary structure,^{18–20} the 3-residue socket and the 4-residue knob-socket cliques account for packing within secondary structure and the tertiary packing between secondary structure elements. For regular secondary structure, all residues are involved in either sockets or knob-socket motifs.^{18,19} Sockets from regular secondary structures of α -helices and β -sheets are designated with the **XY:H** convention, where **XY** are two amino acids connected through one or two peptide bonds, and **H** is an amino acid on the same secondary structure as **XY** and in contact with both.^{18,19} Additionally, less regular coil and turn secondary structure contain **XYZ** sockets, depicted in Figure 1h, where most residues are involved in knobs and some in sockets.²⁰ The tetrahedral knob-socket motif's symbolic representation is four amino acids labeled as **XY:H+B**. The **XY:H** in **XY:H+B** is the socket discussed above. The **B** in **XY:H+B** is the knob residue packing into the socket formed by **XY:H**. In this way there are free sockets, **XY:H**, and filled sockets, **XY:H+B**. Figure 1 depicts the **XY:H+B** knob-socket representation for α -helix (Fig 1a), β -sheet (Figure 1d, 1e), coil and turn (Figure 1h). Because the knob **B** residue packing into the **XY:H** socket provides a simplified decomposition of complex packing networks, the **XY:H+B** serves as a fundamental unit of protein packing. Therefore, another important aspect of the knob-socket model is the simplified description of regular secondary structure elements as 2D lattices of repeating sockets. The 2D α -helix lattice is presented in Figure 1c: the 2D β -sheet lattice is Figure 1f, the coil lattice is Figure 1i, and the turn lattice is Figure 1j.

Defining Pockets

A pocket is defined as a set of contiguous sockets that pack with the same knob. Since a socket consists of 3 residues, 2 contiguous sockets consist of 4 residues and 3 contiguous sockets consist of 5 residues (Figure 1). Because a single knob residue packs into these pockets, the geometrical properties and amino acid compositions of tertiary packing surfaces can be investigated in a definitive manner. Furthermore, the knob-socket model's regular lattices that identify socket relationships provides a limit to the types of pockets that a knob can pack into, so that the pocket space can be explored exhaustively. To identify the various arrangements of residues in pockets, the data collection for this work required a method to mine knob-pocket data that was generated from the 16,673 domains in the ASTRAL SCOP 1.75 set filtered at 95% sequence similarity.⁴² The list of PDB⁴³ codes is available upon request. Secondary structures were defined using DSSP.⁴⁴ The pocket patterns present in α -helices, β -sheets, coils, and turns were investigated in the following manner.

Previous work^{18–20} discussed the analysis of single cliques classified by contact order⁴¹: the 3 residue sockets. For the regular secondary structure of α -helices and β -sheets, the 2:1+1 cliques are the dominant knob-socket interaction.^{18,19} Therefore, the pocket analysis is restricted to pockets constructed from combinations of 2:1+1 cliques. For coil and turn secondary structures, the 3+1 cliques are the most dominant knob-socket interaction, thus pocket construction was restricted to combinations of contiguous 3+1 cliques⁴⁵. The interpretive power of knob-socket model is that the 3D arrangement of residues in a pocket can be directly decomposed into 2D based on the socket lattices for each secondary structure as shown in Figure 1. The essential approach was to collate the socket amino acids that all pack with the same knob, then determine the geometrical arrangement of the pocket in terms of the 2D lattices for α -helices, β -sheets, coils, and turns. Statistics of pockets from each type of secondary structure is shown in Table 1.

To enumerate over all structures of pockets in the dataset, a new language needed to be developed to represent the geometrical information of pockets in a one-dimensional (1D) string from the 2D lattice. The method to construct the 1D string incorporates rules so that a translation of a 1D code produces a single unique 2D pocket shape and vice versa. The topological differences between the four main secondary structures led to the use of two different systems of naming pockets. Examples of the major pockets with their 1D code is shown in Table 2. Although these codes are useful for computational enumeration, they are unwieldy and nondescriptive, so throughout the manuscript these sheet pockets are referred to using their more descriptive names presented in Table 2. Since the residues in α -helices, coils, and turns are local in primary sequence, their relative sequence distance from the lowest residue sequence is sufficient information to project these pockets onto a 2D lattice. For example, the diamond shaped pocket in the 2D α -helix lattice formed by 4 residues at positions, i , $i+3$, $i+4$, and $i+7$ has sequence separations of 0, 3, 4, and 7 from the lowest sequence anchor residue. Combining these numbers produces a 1D name **0347** for this pocket, which can be seen in Figure 1c. For more examples, the 1D codes for α -helix pockets in row A in Figure 2 from left to right are **01458**, **03467**, **0347**, **03478**, and **01347**, respectively.

Building a language that can describe β -sheet pockets is more complicated because of the non-local and bidirectional nature of strands in sheet structures (Figure 1g). Knowing only the relative sequence separation between residues in a β -sheet pocket does not provide enough information on the spatial organization of the residues in a sheet that construct pockets. There are two requirements to describe pockets in sheets: relative pocket orientation and appropriate sheet register information. Although DSSP identifies the relative strand orientation between pairs of strands as parallel or anti-parallel, either strand on the edge of a pocket can be assigned as the bottom. Therefore, relative pocket orientation requires defining the strand at the bottom of the pocket and the sequence direction relative to the knob. The approach to make this designation is outlined in Supplementary Figure S1. First, the strand lowest in sequence between the edge strands of the pocket is designated as the bottom edge of the pocket. Then, a dot product is taken between a vector normal to the plane of the pocket and the $\text{Ca-C}\beta$ vector of the knob. The vector normal to the pocket is found by computing the cross product between the vector from the lowest residue on the bottom

strand to the next residue in sequence and the vector from the lowest residue in sequence on the bottom strand of the pocket to the residue highest in sequence on the top strand of the pocket. In this way, the approach applies to pockets of any size. If the dot product is negative, then the sequence on the bottom strand is increasing to the right relative to the knob. It is important to note that as the number of amino acids increases in the sheet pocket, or the number of strands in the sheet pocket increases, there will be a considerable curve to the β -sheet. Therefore it is important to know that the method we use to determine the orientation of the bottom strand in the pocket relative to the knob is an approximation. Sheet register information is determined from the DSSP secondary structure assignments⁴⁴ and indicates which residues lay across from each other in the 2D sheet lattice. Register number is referenced against the residue on the bottom strand, which creates an effective coordinate system on the 2D lattice. So, all residues in the other strands that are in register with the reference anchor residue are assigned the 0 value. Orientation on the other strands is given by “+” or “-” values for the corresponding registers. A “.” is used to indicate the separation between strands. Once the register and orientations are determined, the 2D geometry of the pocket from the lattice can be encoded into a 1D string. For example, the pocket packing with the B1 knob in the β -sheet lattice shown in Figure 1g consists of residue 2 on the i strand, residues 4 and 2 on the anti-parallel j strand, and residues 2 and 4 on the parallel k strand. The residues 3 on both the j and k strands are not involved as they face the other side of the sheet. Relative to the bottom strand with the lowest sequence, the residues are i , j , $j-2$, k , and $k+2$, so the 1D code is **0:0-2:02**. These rules allow us to read 1D codes directly from 2D representation of any sheet pockets shown in Figure 2. For example, the 1D codes for β -sheet pockets in row E are **024:02**, **02:02:2**, **02:02**, **02:02:0**, and **02:024** respectively.

Rotamer Analysis

To investigate the side-chain orientations of knobs packing into specific pocket geometries, the specific rotamer orientations for the knob residues were calculated. A rotamer is the set of amino acid side chain torsion angles (χ 's) describing the conformation of the amino acid side-chain. These angles were calculated using the MMTSB tool set⁴⁶ and were assigned to a rotamer according to the scheme utilized by Lovell et. al.⁴⁷ Inconsistent sets of angles and planar χ angles were not used.

Kullback-Leibler Divergence

For the amino acid composition of pockets, the Kullback-Leibler divergence (D_{KL})⁴⁸ was used as a comparison metric to quantify the difference between two probability distributions. Standard statistical tests, such as the chi-square, cannot be used due to the interdependence of amino acids within the data set. The D_{KL} between two different probability distributions, P and Q , is defined as:

$$D_{KL}(P||Q) = \sum_i P_i \log \left(\frac{P_i}{Q_i} \right) \quad (1)$$

Using a base-2 logarithm sets the unit of information in bits. The D_{KL} between general distributions P and Q is a suitable means to discriminate between two probability

distributions P and Q.⁴⁸ For example, the distribution P is the data collected on amino acid composition of pocket cores, while the distribution Q is the rates that each amino acid is found in α -helix/ β -sheet. The D_{KL} quantifies how different these distributions are from each other. The asymmetry of the D_{KL} is not relevant in the context of our problem and does not make it an improper metric. An investigation of the new information that can be gained when considering amino acid compositions of pockets versus amino acid compositions of α -helices and β -sheets is not affected by asymmetry of the D_{KL} .

RESULTS AND DISCUSSION

Pocket Analysis

Table 1 summarizes pocket frequencies in α -helices, β -sheets, coil, and turns classified by the secondary structures of the knobs: α -helices, β -sheet, and coil. The total number of instances a knob was observed packing into a pocket and the total number of unique pocket types is presented for each type of secondary structure. Based on the predominance of helical structure in the Protein Data Bank (PDB),⁴³ it was expected to find that most pockets are filled from helical knobs. However, a rather surprising result was the abundance of coil knobs packing into pockets. The percentages for secondary structure of knob packing into α -helix and β -sheet pockets are 60%, 18%, 22% and 43%, 22%, 35% (helix, sheet, coil), respectively. Helix knobs tend to pack into helix pockets, but helix and coil knobs pack significantly into sheet pockets. Additionally, 49% and 47% of coil and turn pockets are packed with coil knobs. Coil knobs have been implicated in playing a vital role in protein packing by providing stabilizing hydrophobic filler interactions between the more rigid elements of secondary structure.⁴⁵

This table also indicates the difference in the number of unique pocket shapes. There is less diversity in α -helix pocket types than in β -sheet pocket types. An explanation for this difference is the rigid and regular conformation of helical structures in proteins compared to the flexibility of sheet structure.^{6,25,49} This rigidity constrains the different modes in which knobs can pack into pockets on α -helices. In contrast, β -sheets exhibit greater structural freedom that allows for a wide variety of pockets to interact with knobs. In addition, sheets are found to wrap around other secondary structures, producing more diverse knob-pocket interactions.^{6,23,24} A surprising result was the low number of unique pocket types observed in coil and turn pockets. The irregularity of these secondary structures did not allow for higher order socket patterns to form, as seen in the limited number of unique pocket types from Table 1.

Although many different types of pocket shapes are possible (Table 1), five pocket types for both α -helix and β -sheet account for the large majority (96% and 95%, respectively) of the pockets observed in this study. Similarly, three pockets for coil and one turn pocket account for large majorities (93% and 83%, respectively) of the pocket occurrences of coil and turn pockets. The distribution of pocket size in sockets shows an exponential decrease with increasing number of residues in a pocket for both α -helices and β -sheet (Supplementary Figure S2), which is consistent with larger pocket sizes being more complex to form. For the major pocket sizes and shapes, the frequency of the major pockets found in α -helix, β -sheet,

coil, and turn are shown in Table 2 classified by secondary structure of the knob. For comparison, the α -helix sockets and the β -sheet main-chain socket are included.

Focusing first on the α -helix pockets, the most prevalent helix pocket is the **0347** diamond pocket. The diamond pocket consists of two sockets and is commonly known as the “hole” in helical packing.²⁷ The diamond pocket accounts for roughly 70% of pockets found in α -helices. The preferences for helix, sheet, and coil knobs packing into helix diamond pockets are roughly the same as the preferences seen in Table 1. The **0145** Low-High and **0134** High-Low pockets are also combinations of two sockets in two different orientations. The Low-High and High-Low helix pockets accounts for 12% and 4% of the pockets in helix, respectively. The Low-High pocket is of greater abundance than the High-Low due to residues 1 and 3 facing opposite sides of the helical cylinder and requiring specific amino acid compositions to achieve packing. The Low-High and High-Low α -helix pockets are found packed with a higher fraction of knobs from coil (~27% each) than coil knob preference in general for helix pockets (~22%), as shown in Table 1. This could be due to the greater flexibility of coil to pack into these arrangements.

The High (**03478**) and Low (**01458**) diamonds are pockets that have a diamond core with an extra interaction at the i+8 or i-1 positions, respectively. These pockets contribute less than 4% to the total number of pockets in helix. These extra, or ancillary, interactions in comparison with core interactions will be discussed further in the next section. The secondary structure preferences for knobs packing into High and Low diamonds also indicate a higher fraction of coil (26% and 24%, respectively). The increased preference for coil knobs could again signify the importance of flexibility of coil for knobs to pack into these pockets.

Additionally, Table 2 describes the top seven packing shapes found in β -sheets that accounts for approximately 95% of the sheet pocket observations. The side chain (SC) sockets are the preferred packing surface in β -sheets accounting for 49% of sheet pockets. We include the SC socket in our pocket discussion because the SC socket serves as a core for many pockets. The concept of a core pocket will be discussed in the next section. The main chain (MC) sockets account for 15% of packing into β -sheets. It is important to note that MC sockets are primarily packed with coil knobs, which also might be due to the flexibility of coil. The SC diamond pocket accounts for 12% of sheet pockets.

The SC diamond is a pocket that spans three strands. It was found that SC diamond pockets pack with knobs primarily from α -helices. With slightly fewer counts, the next populated pocket is the SC Square, accounting for 11% of sheet pockets. The SC square pocket also favors helical knobs. Each pocket after the SC square accounts for less than 4% of the sheet pockets. The MC square accounts for approximately 4% of sheet pockets. We again observe that MC square pockets are packed significantly with helix and coil knobs. Next we observe the 2-strand-diamond and the 3-strand-triangle responsible for 2% and 1% of sheet pockets, respectively. It is interesting to note how un-represented these pocket geometries are in sheet pockets. The pocket analysis we present here shows that sheets pockets are primarily organized into SC sockets, SC diamonds, and SC squares.

Finally, Table 2 presents the most significant pockets in coil and turn. The top three coil pockets are a diamond pocket, the ancillary diamond pocket, and the off pocket. Respectively, these three pockets constitute 72%, 11%, and 10% of observations. Coil diamond pockets are simply two overlapping 3+1 coil sockets. The ancillary coil diamond pocket is a coil diamond pocket with an extra interaction with residue 4. The off coil pocket is a pocket that skips the first residue (and by symmetry the fourth position). These last two coil pockets do not significantly contribute to the population of pockets found in coil as compared to the coil diamond pocket. Similarly, the diamond pocket in turn forms a dominant 83% of the time. The rest of the 13 turn pocket types accounts for the remaining 17% of the turn pocket observations. Again, it is interesting to note that coil and turn secondary structures do not exhibit the diversity of pockets as seen in β -sheets. For additional analysis of coil and turn pockets, the top ~95% coil and turn pockets are shown in Table S1.

Core Pockets and Ancillary Packing

Analysis of the larger, less populated pockets for both α -helices and β -sheets indicate that there are variations upon the more common pockets shown in Table 2. To provide consistency to the analysis, these larger pockets are classified based on the common pocket cores, and the extra residues outside of the core are considered ancillary to the overall knob-pocket interaction. The pockets in Table 1 are the most consistently observed pocket cores. Figure 2 introduces a variety of α -helix, β -sheet, coil, and turn pockets identified in protein tertiary structure and shows examples of core pockets with their ancillary variations.

While the larger pockets could be thought of deriving from a number of core geometries, for consistency, the choice of core pockets are based on the most prevalent and consistent geometries. The other important factor in determining core pockets is centrality. The shape in the center of a pocket is the core; the surrounding elements of a pocket are ancillary. Row A of Figure 2 shows all the prevalent helical pockets. The diamond pocket is the most prevalent pocket in α -helix packing interfaces (87.8%) as seen in Table 2. For this reason, the other larger helical pockets in Row A can be seen as sharing a common diamond core and the extra residues are considered ancillary. For the sheet pockets shown in Figure 2, these principles are used to demarcate between the square and diamond pockets in β -sheets. Row B and E in Figure 2 contain sheet pockets with a square core and ancillary sockets around the core. Rows C and D present variations of a core diamond sheet pocket with surrounding ancillary interactions.

Figure 2 illustrates the chiral nature of pocket patterns, and therefore an aspect of the chiral nature of protein packing. As diagrammed, the helix pockets favor left-handed conformations, whereas sheet pockets clearly favor right-handed orientations. This specific handedness originates naturally from the inherent secondary structure. For α -helices, the pattern of side-chain placement along the helical axis not only favors the diamond pocket, which has been characterized in the well-known “knob into hole” motif,²⁷ but also favors the left-handed shape of helical packing pockets.^{25,49} For β -sheets, the individual strands exhibit a well documented $\sim 30^\circ$ twist in secondary structure that conventionally is designated as right handed.^{23,50,51} This twist in β -sheets produces the right-handed chirality of sheet

pockets. It is encouraging that the knob-socket motif is sensitive to these structural features of protein secondary structure. Again, the objective of core pocket identification is the organization of the diverse packing found in protein structure. The perspective acquired from the knob-socket allows a unique organizational principle to protein tertiary structure.

Figure 2 also shows the ancillary expansion from diamond core pockets depends on the right-handed α -helix topology. The diamond core pocket is overlap between ± 4 and ± 3 grooves.²⁵ The expansion along the ± 4 grooves is observed more often than expansion along the ± 3 grooves. The **01458** and **03478** pockets (± 4 groove) are 4.6 % and 4.8% of total pockets related to diamond core pockets. On the contrary, the **03467** and **01347** pockets (± 3 groove) are observed 0.8% and 2.0% of total diamond core pockets. This is due to the helical topology that ± 4 grooves simply expand on the same plane, but the ± 3 groove expansion is involved with the residues on another plane that is almost perpendicular to the diamond plane. This also shows that topological characteristics of α -helices are also accounted for by the pocket analyses.

Core vs. Ancillary Amino Acid Composition

The organizational principle of core versus ancillary is further investigated by computing the frequency each amino acid was observed in a core or ancillary position of a pocket. Figure 3 presents histograms of the frequency each amino acid was observed in core and ancillary positions of the helical diamond and top four sheet pockets. Additionally, Figure 3 presents the distribution of amino acid composition in α -helix and β -sheets, the total frequency histograms. Amino acid compositions of α -helix are frequently composed of Leu, Ala, and Glu, while Val, Leu, and Ile are the most frequently observed amino acids composing β -sheets.

In general, core amino acid frequency distributions are distinct from the total and ancillary amino acid distributions. In α -helices, Leu, Ala, Ile, and Val are most common amino acid residues observed in core diamond pockets, but their relative proportions are reduced to about a third in ancillary positions of diamond pockets. For example in the helix diamond core, Leu is about 3.5 times more than Phe, but in ancillary helix diamond pockets the former is only about 1.2 times more than the latter. Similar trends can be found in β -sheet pockets. For example, Ala is about 2.5 times more common than Tyr in SC socket cores but is observed at the almost same frequencies as the Tyr in ancillary positions of SC sockets. However Val is equally prevalent in ancillary and core positions of SC sockets. This illustrates that there are amino acid preferences for the identified core geometries, which once again supports that the knob-socket model produces an amino acid structural code.^{18,19}

The results of the D_{KL} analyses are presented in Table S2. The quantitative results illustrate that amino acid composition of core pockets in α -helices are consistent with the amino acid composition of α -helices. The results also signify the uniqueness of the SC socket and SC diamond in β -sheets. The helical diamond, Low-High, and High-Low core amino acid frequency distributions exhibit D_{KL} with helical amino acid compositions of less than 0.10 bits. Where the core amino acid compositions of helical pockets are very similar to the compositions of helices, the ancillary compositions of helical diamonds are very different than the composition of helices. The D_{KL} between the core and ancillary amino acid

compositions of diamond pockets in α -helix was 0.186 bits, and between the ancillary diamond and total α -helix amino acid composition was 0.129 bits. These differences are caused by ancillary interactions in diamonds that require long residues to bend around the helix to interact with the knob. This is clearly the case with Trp, Phe, and Leu presenting higher propensity in ancillary helical diamonds (Figure 3).

The β -sheet SC Socket and SC Diamond core amino acid and β -sheet compositions exhibit D_{KL} of 0.16 and 0.35 bits, respectively. These two exhibit the greatest divergence from the β -sheet amino acid composition, suggesting that the amino acids constructing these packing geometries are specific. More importantly, the amino acid composition of the ancillary portions of these two pockets exhibits great similarity with the composition of sheets. The D_{KL} between the SC Socket and SC Diamond ancillary amino acid compositions with β -sheet amino acid compositions are 0.027 and 0.065 bits, respectively. The small divergences indicate that the pockets identified with the SC Socket and SC Diamond core display amino acids in ancillary positions at rates similar to the amino acid compositions of β -sheets. In fact, all pockets exhibit ancillary compositions very consistent with the composition of sheets, except for the MC Square ancillary composition, which could be due to small sample size. The SC, MC Square core amino acid compositions all exhibit small divergences with the β -sheet amino acid composition with all divergences less than 0.100 bits. These small differences suggest a non-specificity for the amino acid composition for the two core pockets. Their amino acid compositions all maintain similarities with the amino acid composition of sheets.

Position Specific Heat Maps

For the pockets presented in Figure 3, the position specific preferences of amino acid pocket compositions can be characterized. To further establish the sequence specificity found in the pocket cores, position specific amino acid heat maps were calculated. Figure 4 illustrates the propensity for each of the twenty amino acids at specific core positions in α -helix and β -sheet pockets. Column (1) illustrates the position that a particular amino acid will occupy in the core and the column (2) illustrates the amino acid that each position in the pocket prefers. First, the α -helix diamond pocket displays position specific sequence preferences. From the heat map of Figure 4a(1), amino acids Trp, Tyr, Phe, Ile, Leu, and Met all prefer to be found at position 0 and/or 7 of the pocket. This preference could be considered as a result of their side chain length; however, the amino acids from Ala to Asn in the map do not prefer position 7 even though Lys, and Arg have long side chains.

The Low-High pocket establishes position specificity by way of minimizing steric hindrance. Referencing the Figure 4b(1) heat map, Tyr occupies almost exclusively at position 0 of the Low-High pocket to possibly allow enough room for knob packing. In a similar way, the propensities of amino acids Gly, Tyr, Phe, Leu, and Ala all exhibit behavior indicative of how length and size of side-chain influences position specificity in the Low-High pocket.

For the High-Low pocket (Figure 4c(1)), the side chain length and the helical topology again affect amino acid preferences at pocket positions. Once again, Trp is selective for the position in the pocket and highly prefers the 3 position, which can be again attributed to its

size. Amino acids Gly, Tyr, and Leu also exhibit behavior that further establishes the importance of length of side-chain in pocket composition. The heat maps of Figure 4a(2), 4b(2), and 4c(2) all show that mainly non-polar residues dominate the sequence compositions found in the major α -helix pockets. While there may be some details that are not conserved between these α -helix pocket heat maps, it is still clear that non-polar residues play important roles in α -helix packing interactions.

The next four rows of heat maps in Figure 4 present the position specificity for β -sheet pockets. For clarity in this section, the XY:HG sheet pocket notations will be used with underline to distinguish them from the single letter amino acid code. A significant trend across the last four heat maps in column 2 is the strong preference for non-polar amino acids to populate each position of the four major packing shapes in β -sheets. There is some sequence specificity in the MC Square; however, non-polar residues still compose the majority of residues found in those pockets. The position specific heat maps in column (1), however, display significant specificity influenced by the topology of the sheet, side chain size and chemical nature of amino acids.

The position specific heat map for the SC socket in Figure 4d (1) displays two significant trends. The non-polar amino acids from Tyr through Leu prefer the positions Y and H of the SC Socket. This could mainly be due to the fact that positions Y and H in the SC socket involve non-covalent, van der Waals interactions, which non-polar residues best satisfy. The other important signal in Figure 4d(1) is the preference for the charged and polar residues to be found at the X position of the SC Socket. This trend could be due to a negative design principle. The charged and polar residues are more likely to occupy the X position because they cannot satisfy the non-covalent interactions necessary to be placed at positions Y or H.

Next, the sheet diamond pocket (Figure 4e(1)) shows that the length and size are important factors for amino acid placement in sheet diamond pockets. Amino acids Trp, Gly, Tyr, Phe, Asp, and Thr show preferences for certain positions which can be rationalized by their size and length. The SC Square position specific histogram at Figure 4f(1) is an exception. There is no real observable pattern in the heat map indicating non-specificity. For example, both the largest amino acid Trp and the smallest Gly have large propensities to populate the Y position of the square pocket, suggesting little dependence on length of residue.

The MC Square exhibits great position specificity as observed in the position specific heat map at Figure 4g(1). Mainly the non-polar residues Tyr through Ala all prefer to pack at the X and H positions, but the polar residues prefer the Y and G positions. This specificity is because of the topology of the β -sheet. The side chains of the amino acids at the positions X and H are pointing towards the knob. Therefore, in order to present the most attractive environment to pack, the amino acids at these positions should be non-polar. The charged/polar amino acids Glu through Asn pack at the Y and G positions that do not interact with the knob.

Overall, Figure 4 provides an amino acid code that specifies the amino acid composition of sockets and pockets in α -helices and β -sheets. Therefore, if a protein structure design places amino acids at positions of pockets that data does not support, the packing should be

adjusted accordingly in order to effectively pack. As somewhat expected, the histograms in column (2) indicate that each pocket prefers mainly non-polar amino acids. This suggests that there is no discernible signal between amino acid sequences that form specific pockets. Yet, some specificity is found with the MC Square and the High-Low and Low-High helix pockets. However, for the rest of the pockets, non-polar residues are most often found in those shapes. This analysis allows protein design efforts to rationally specify positions for amino acid placement in pockets as shown in column (1) of Figure 4.

Figure 5 further establishes the importance of non-polar residues in pocket compositions, presents the skewedness of the data, and provides a more global perspective on the sensitivity of the pocket to amino acid type. Because the $20^4/20^3$ combinatorics of a four/three residue pocket produce such sparse data across the pockets types, the data was collapsed based on polarity and charge of the 20 amino acid residues: polar (P), non-polar (N), and charged (C). A 75% cutoff of the observed counts was used to display the data in Figure 5 for each pocket shape. All non-polar compositions are the most prevalent pocket composition regardless of pocket shape. Furthermore, all of the most frequent pockets are composed of at least one non-polar residue.

This figure also presents the skewedness in this data. For example, only 55% of the 81 amino acid compositions are observed for the helix diamond to account for 75% of the counts. The SC socket skewedness is more drastic with 36% of sequences accounting for 75% of the data. This skew implies that many possible sequences can form these pockets, but are not well represented in the data set.

The data in this figure also presents a more global view of the sensitivity to amino acid type in pocket composition. For example, comparing the NNCN and NNNC compositions in the helix diamond bar graph shows that NNCN and NNNC are not equivalent and the NNCN is better represented in the data than NNNC. This is attributed to the position specificity found in helix diamond pockets; charge residues are not very commonly found at position 7 in helical diamond pockets.

Knob Analysis

Figure 6 shows the normalized frequency of the 20 amino acid knobs packing into the prevalent α -helix and β -sheet pockets. The data is presented in two different ways: the knob preferences for pockets (Figures 6a and 6c), and the pocket preferences for knobs (Figures 6b and 6d). As shown in Figures 6a and 6c, the non-polar residues are well represented as knobs for the five types of helix and sheet pockets. An exception is Tyr, which is well represented as a knob for helical pocket but not for sheet pockets (Figures 6a, and 6c). In the same vein, Figures 6b and 6d illustrate that the 20 amino acids all prefer a similar set of pockets. Figure 6b illustrates that the 20 amino acids prefer the helical diamond pocket at a high rate. For amino acids packing into sheets, all residues pack strongly into the SC Socket as shown in Figure 6d. The SC Socket has been shown to be a fundamental packing motif for sheets, making this unsurprising¹⁹. The other pockets are packed at a very minimal rate. Hence specific pockets do not prefer specific knobs at any significant margin.

While specific amino acid knobs do not pack into specific pocket shapes, Figure 7 investigates if the non-specificity extends also to residue conformation by comparing the rotamer distributions of knobs in different pocket types. Figure 7 illustrates that the rotamer distributions does not significantly differ for each amino acid within the different pocket types. In addition, the rotamer distributions for Ile, Leu, Glu, and Lys maintain similar trends: the rotamer distributions for each knob maintain many similarities across pocket shapes. This result is consistent with previous studies.⁵² This indicates that the forces involved in packing are very weak and non-specific and provides a unique insight into the nature of packing. Regularity may emerge out of the shape of pockets; however, this regularity does not enforce spatial restrictions on the protein structure.

Pocket Mapping in Real Structures

As an example of the utility of pocket analysis in the knob-socket framework, Figure 8 provides some illustrative examples of pockets and ancillary pockets. Ribbon diagrams for the pockets and their representations on 2D lattices produce simple topological shapes for tertiary structure mapping. The first part of Figure 8 shows four types of α -helix pockets, whereas the second part show examples of five types of β -sheet pockets. Figure 8a shows an example of the most common type of helix packing: knob packing into a diamond pocket and its 2D representation is shown on the helix lattice to the right. Figure 8b presents an ancillary helix diamond pocket, the low helix diamond. It clearly shows the specific diamond pocket contacts with the knob and additional contact with the ancillary portion of the pocket. The less common low-high (Figure 8c) and high-low (Figure 8d) pocket examples are also shown.

In a similar fashion, Figures 8e–i presents examples of β -sheet pockets. Figure 8e is a pocket with the SC socket as a core with an ancillary side-chain interaction. This extra side-chain interaction is caused by the curvature of the sheet. This pocket is depicted on the sheet lattice packing with knob e. The canonical sheet diamond pocket is seen in Figure 8f. Figure 8g displays a SC diamond pocket with a main chain contact. The ribbon diagram clearly shows that the knob primarily packs with the residues forming the core diamond pocket, but the interaction with main chain forms the ancillary pocket. Figure 8h presents a SC square pocket with an ancillary side-chain interaction. The knob primarily packs with the residues forming the square while the ancillary interaction is caused by the extra side-chain bending over to interact with the knob. Lastly, an interesting discovery is the sheet MC square interaction, shown in Figure 8i. In many cases, the knob with the MC Square pocket can be considered a β -sheet stabilizing interaction. Despite the low abundance of MC Square pockets, these pockets may still be structurally significant to the integrity of sheets. While all the packing interactions are complex, the abstraction into the repetitive lattices demonstrates the intuitive strength of the knob-socket model by simplifying the packing representation and helping further our understanding of protein packing.

CONCLUSION

The analyses of pockets, clusters of sockets packing with a knob, provide deep insights into the nature of protein tertiary structure. Even though a majority of packing interactions are

knob-socket interactions, pockets play significant roles in characterizing the packing interfaces and relative orientations of the secondary structure units. On α -helix, the diamond is the most prevalent pocket shape. The preferred pockets in β -sheets are the diamond and the square. Coil and turn secondary structures do not form extensive pockets. When they do, both prefer the diamond configuration, which accounts for 72% of coil and 83% of turn pocket observations. Further, separating pockets into core and ancillary segments, we have shown that many pockets can be classified according to the core geometry of a socket, diamond, or square. In Figure 3, with D_{KL} analyses of the amino acid compositions, we show that the ancillary regions of pockets follow the amino acid compositions of α -helix and β -sheet. These results indicate that the amino acid compositions of core pockets are specific to packing but ancillary interactions are more context dependent. The chiral nature of protein structure has been recognized and plays an important role⁵³. The α -helix is innately right-handed and combination of ± 3 and ± 4 grooves can only make left-handed diamond pockets, and the β -sheet's right hand twists result in right-handed diamond pockets. The pocket analyses can characterize the chirality at the level of packing between amino acid residues.

While most of the analyses reiterate the non-specific nature of packing, the knob-socket model shows specific packing preferences of amino acids. As shown in Figure 4, amino acids display preferences to certain positions in packing pockets. These preferences depend upon the secondary structure topology, length/size, and chemical nature of amino acids. The non-specificity of packing, however, is consistent with preference for hydrophobic amino acid compositions of all pockets (Figures 3, 4, 5, 6), prevalence of hydrophobic knobs packing with each pocket (Figure 6), and rotamer distributions for each amino acid packing into specific pockets (Figure 7).

The knob-socket model is able to generalize packing into simple geometrical shapes, yet is also able to identify specific details of protein packing, such as the MC Square pocket. Expanding knob-socket into knob-pocket interaction show many promising benefits. The cooperative action of amino acids to form a pocket from sockets involves one or more additional residue interactions on the packing interface. This will enhance both strength and specificity of the interactions. The projection of protein packing into the 2D socket lattices allows for an intelligent interrogation of structure. Identification of pockets in the tertiary structure packing surfaces helps to better understand mutational effects of individual residues in terms of packing as a knob into a pocket as well as functioning as a residue in a pocket. This descriptive way to analyze protein packing provides perspective and insight into the stability, structure, and function of proteins.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank members of the Tsai laboratory and Sayandeb Basu for helpful discussions. We thank Alfonso Gonzalez for his dedication in verifying the placement of MC Square pockets. This work was supported by an NIH NIGMS grant R01-GM104972.

References

1. Pauling L, Corey RB. Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. *Proceedings of the National Academy of Sciences of the United States of America*. 1951; 37(11):729–740. [PubMed: 16578412]
2. Pauling L, Corey RB. The structure of fibrous proteins of the collagen-gelatin group. *Proceedings of the National Academy of Sciences of the United States of America*. 1951; 37(5):272–281. [PubMed: 14834150]
3. Pauling L, Corey RB. The structure of hair, muscle, and related proteins. *Proceedings of the National Academy of Sciences of the United States of America*. 1951; 37(5):261–271. [PubMed: 14834149]
4. Pauling L, Corey RB. The structure of feather rachis keratin. *Proceedings of the National Academy of Sciences of the United States of America*. 1951; 37(5):256–261. [PubMed: 14834148]
5. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*. 1951; 37(4):205–211. [PubMed: 14816373]
6. Levitt M, Chothia C. Structural patterns in globular proteins. *Nature*. 1976; 261(5561):552–558. [PubMed: 934293]
7. Richardson JS. The anatomy and taxonomy of protein structure. *Advances in protein chemistry*. 1981; 34:167–339. [PubMed: 7020376]
8. Chothia C. Principles that determine the structure of proteins. *Annual review of biochemistry*. 1984; 53:537–572.
9. Deleage G, Blanchet C, Geourjon C. Protein structure prediction. Implications for the biologist *Biochimie*. 1997; 79(11):681–686. [PubMed: 9479451]
10. Dill KA. Dominant forces in protein folding. *Biochemistry*. 1990; 29(31):7133–7155. [PubMed: 2207096]
11. Tanford C. The hydrophobic effect and the organization of living matter. *Science*. 1978; 200(4345):1012–1018. [PubMed: 653353]
12. Kauzmann W. Some factors in the interpretation of protein denaturation. *Advances in protein chemistry*. 1959; 14:1–63. [PubMed: 14404936]
13. Pauling L, Corey RB. The polypeptide-chain configuration in hemoglobin and other globular proteins. *Proceedings of the National Academy of Sciences of the United States of America*. 1951; 37(5):282–285. [PubMed: 14834151]
14. Pauling L, Corey RB. The pleated sheet, a new layer configuration of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America*. 1951; 37(5):251–256. [PubMed: 14834147]
15. Pauling L, Corey RB. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proceedings of the National Academy of Sciences of the United States of America*. 1951; 37(5):235–240. [PubMed: 14834145]
16. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*. 1963; 7:95–99. [PubMed: 13990617]
17. Day R, Lennox KP, Dahl DB, Vannucci M, Tsai JW. Characterizing the regularity of tetrahedral packing motifs in protein tertiary structure. *Bioinformatics*. 2010; 26(24):3059–3066. [PubMed: 21047817]
18. Joo H, Chavan AG, Phan J, Day R, Tsai J. An amino acid packing code for alpha-helical structure and protein design. *Journal of molecular biology*. 2012; 419(3–4):234–254. [PubMed: 22426125]
19. Joo H, Tsai J. An amino acid code for beta-sheet packing structure. *Proteins*. 2014; 82(9):2128–2140. [PubMed: 24668690]
20. Joo H, Fraga K, Chavan A, Tsai J. An Amino Acid Code for Irregular and Mixed Secondary Structure Packing. *Proteins: Struct Funct & Bioinform*. 2015 accepted.
21. Wu GA, Coutsiadis EA, Dill KA. Iterative Assembly of Helical Proteins by Optimal Hydrophobic Packing. *Structure (London, England : 1993)*. 2008; 16(8):1257–1266.

22. Hu C, Koehl P. Helix-sheet packing in proteins. *Proteins*. 2010; 78(7):1736–1747. [PubMed: 20186972]
23. Chothia C, Levitt M, Richardson D. Structure of proteins: packing of alpha-helices and pleated sheets. *Proceedings of the National Academy of Sciences of the United States of America*. 1977; 74(10):4130–4134. [PubMed: 270659]
24. Janin J, Chothia C. Packing of alpha-helices onto beta-pleated sheets and the anatomy of alpha/beta proteins. *Journal of molecular biology*. 1980; 143(1):95–128. [PubMed: 7003165]
25. Chothia C, Levitt M, Richardson D. Helix to helix packing in proteins. *Journal of molecular biology*. 1981; 145(1):215–250. [PubMed: 7265198]
26. Schiffer M, Edmundson AB. Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys J*. 1967; 7(2):121–135. [PubMed: 6048867]
27. Crick FHC. The Packing of α -Helices: Simple Coiled-Coils. *Acta Cryst*. 1953; 6(8–9):689–697.
28. Walshaw J, Woolfson DN. Socket: a program for identifying and analysing coiled-coil motifs within protein structures. *Journal of molecular biology*. 2001; 307(5):1427–1450. [PubMed: 11292353]
29. Walshaw J, Woolfson DN. Extended knobs-into-holes packing in classical and complex coiled-coil assemblies. *J Struct Biol*. 2003; 144(3):349–361. [PubMed: 14643203]
30. Petersen SB, Neves-Petersen MT, Henriksen SB, Mortensen RJ, Geertz-Hansen HM. Scale-free behaviour of amino acid pair interactions in folded proteins. *PloS one*. 2012; 7(7):e41322. [PubMed: 22848462]
31. Jha AN, Vishveshwara S, Banavar JR. Amino acid interaction preferences in proteins. *Protein Sci*. 2010; 19(3):603–616. [PubMed: 20073083]
32. Connolly ML. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. *Biopolymers*. 1986; 25(7):1229–1247. [PubMed: 3741993]
33. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci*. 1998; 7(9):1884–1897. [PubMed: 9761470]
34. Cazals, F., Chazal, F., Lewiner, T. Molecular shape analysis based upon the morse-smale complex and the connolly function. *Proceedings of the nineteenth annual symposium on Computational geometry*; San Diego, California, USA: ACM; 2003. p. 351-360.
35. Voronoi GF. Nouvelles applications des paramètres continus à la théorie des formes quad-ratiques. *J Reine Angew Math*. 1908; 134:198–287.
36. Harpaz Y, Gerstein M, Chothia C. Volume changes on protein folding. *Structure*. 1994; 2(7):641–649. [PubMed: 7922041]
37. Tsai J, Taylor R, Chothia C, Gerstein M. The packing density in proteins: standard radii and volumes. *Journal of molecular biology*. 1999; 290(1):253–266. [PubMed: 10388571]
38. Delaunay B. Sur la sphère vide. *Bull Acad Sci USSR (VII) Classe Sci Mat Nat*. 1934:783–800.
39. Bandyopadhyay D, Snoeyink J. Almost-Delaunay Simplices : Nearest Neighbor Relations for Imprecise Points. 2004:403–412.
40. Bron C, Kerbosch J. Finding All Cliques of an Undirected Graph [H]. *Communications of the ACM*. 1973; 16:575–577.
41. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. *Journal of molecular biology*. 1998; 277(4):985–994. [PubMed: 9545386]
42. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. *Nucleic Acids Res*. 2004; 32(Database issue):D189–192. [PubMed: 14681391]
43. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Research*. 2000; 28(1):235–242. [PubMed: 10592235]
44. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983; 22(12):2577–2637. [PubMed: 6667333]
45. Joo H, Fraga K, Chavan A, Tsai J. Simple and Regular Code or Irregular and Mixed Protein Packing. 2015 In Press.

46. Feig M, Karanicolas J, Brooks CL 3rd. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal of molecular graphics & modelling*. 2004; 22(5):377–395. [PubMed: 15099834]
47. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. *Proteins*. 2000; 40(3):389–408. [PubMed: 10861930]
48. Kullback S, Leibler RA. On Information and Sufficiency. 1951:79–86.
49. Chothia C, Levitt M, Richardson D. Structure of proteins: packing of α -helices and pleated sheets. *Proceedings of the National Academy of Sciences of the United States of America*. 1977; 74(10): 4130–4134. [PubMed: 270659]
50. Chothia C. Conformation of twisted beta-pleated sheets in proteins. *Journal of molecular biology*. 1973; 75(2):295–302. [PubMed: 4728692]
51. Schulz GE, Elzinga M, Marx F, Schirmer RH. Three dimensional structure of adenyl kinase. *Nature*. 1974; 250(462):120–123. [PubMed: 4367210]
52. Shapovalov MV, Dunbrack RL Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*. 2011; 19(6):844–858. [PubMed: 21645855]
53. Chothia C. Asymmetry in protein structures. *Ciba Foundation symposium*. 1991; 162:36–49. discussion 49–57. [PubMed: 1802649]
54. Egloff MP, Uppenberg J, Haalck L, van Tilbeurgh H. Crystal structure of maltose phosphorylase from *Lactobacillus brevis*: unexpected evolutionary relationship with glucoamylases. *Structure*. 2001; 9(8):689–697. [PubMed: 11587643]
55. Fishman R, Ankilova V, Moor N, Saftro M. Structure at 2.6 Å resolution of phenylalanyl-tRNA synthetase complexed with phenylalanyl-adenylate in the presence of manganese. *Acta Crystallogr D Biol Crystallogr*. 2001; 57(Pt 11):1534–1544. [PubMed: 11679717]
56. Pesce A, Couture M, Dewilde S, Guertin M, Yamauchi K, Ascenzi P, Moens L, Bolognesi M. A novel two-over-two α -helical sandwich fold is characteristic of the truncated hemoglobin family. *EMBO J*. 2000; 19(11):2424–2434. [PubMed: 10835341]
57. Milani M, Savard PY, Ouellet H, Ascenzi P, Guertin M, Bolognesi M. A TyrCD1/TrpG8 hydrogen bond network and a TyrB10/TyrCD1 covalent link shape the heme distal site of *Mycobacterium tuberculosis* hemoglobin O. *Proceedings of the National Academy of Sciences of the United States of America*. 2003; 100(10):5766–5771. [PubMed: 12719529]
58. Bolognesi M, Rosano C, Losso R, Borassi A, Rizzi M, Wittenberg JB, Boffi A, Ascenzi P. Cyanide binding to *Lucina pectinata* hemoglobin I and to sperm whale myoglobin: an x-ray crystallographic study. *Biophys J*. 1999; 77(2):1093–1099. [PubMed: 10423453]
59. Zhu X, Larsen NA, Basran A, Bruce NC, Wilson IA. Observation of an arsenic adduct in an acetyl esterase crystal structure. *J Biol Chem*. 2003; 278(3):2008–2014. [PubMed: 12421810]
60. Sandgren M, Gualfetti PJ, Shaw A, Gross LS, Saldajeno M, Day AG, Jones TA, Mitchinson C. Comparison of family 12 glycoside hydrolases and recruited substitutions important for thermal stability. *Protein Sci*. 2003; 12(4):848–860. [PubMed: 12649442]
61. Ding K, Ramelot TA, Cort JR, Jiang M, Xiao R, Swapna GVT, Montelione GT, Kennedy MA. NMR Structure of *E. Coli* YehR Protein. *Northeast Structural Genomics Target*. 2007; ER538
62. Ingelman M, Bianchi V, Eklund H. The three-dimensional structure of flavodoxin reductase from *Escherichia coli* at 1.7 Å resolution. *Journal of molecular biology*. 1997; 268(1):147–157. [PubMed: 9149148]
63. Luan P, Heine A, Zeng K, Moyer B, Greasely SE, Kuhn P, Balch WE, Wilson IA. A new functional domain of guanine nucleotide dissociation inhibitor (α -GDI) involved in Rab recycling. *Traffic*. 2000; 1(3):270–281. [PubMed: 11208110]

Figure 1.

Knob socket (KS) essentials. (a) XY:H representation of the socket in the knob-socket theory for helix. Solid black lines indicate the backbone between two amino acids local in sequence. Dashed black lines indicate non-covalent van der Waals interactions. Red dashed lines are hydrogen bonds. These conventions are the same for the remaining knob-socket representations. (b) XY:H+B representation of the knob B into the XY:H socket. (c) 2D lattice of contiguous sockets representing the α -helix. Presented in the lattice is an XY:H+B knob-socket between residues 7, 8, and 11 on the lattice, and a knob-pocket interaction between residues 1, 4, 5, and 8. (d) Tertiary packing example of knob-socket between a sheet Leu knob and helical socket of a Lys at X, a Leu at Y, and a Val at H from 1JJC.⁵⁴ (e)

XY:H representation of side-chain (SC) socket for sheets (f) XY:H representation of main-chain (MC) socket for sheets. (g) 2D lattice of sheet. The white circles symbolize residues with side chains directed out of the paper. Black circles symbolize residues with side chains directed into the paper. Represented in the lattice are a SC socket, MC socket, and a pocket. The SC socket is represented with residues packed with knob B3. The MC socket is represented by knob B2. The pocket is the combination of residues packing with knob B1. (h) Tertiary packing example from 1H45⁵⁵ of a square pocket in a sheet consisting of a Phe, Tyr, Val and Ile packed with a Ile knob from a helix.

Figure 2.

Examples of Core and Ancillary pockets. This figure introduces the variety of α -helix, β -sheet, coil, and turn pockets one will find in protein packing. These are 2D representations of the pockets. Each row is labeled using a letter from A to F and the each of the pockets are enumerated within the row. The count of each pocket is given under each pocket representation.

Figure 3.

Amino acid composition in α -helix and β -sheet sheet pockets. The frequencies of amino acids in pockets are given in 10^3 counts. The total counts for α -helix (left) and for β -sheet (right) pockets are given on the top panel (above the black solid line). On the lower panel, the frequencies of the twenty amino acids in core (left) and ancillary (right) portion of five representative pockets in α -helices and β -sheets are shown. The From top, the amino acid compositions of helix diamond pocket, sheet SC socket, sheet diamond pocket, SC square pocket, and MC square pocket are presented.

Figure 4.

Position specific heat maps of amino acid distribution in the representative pocket types. The propensities for each of the twenty amino acids to occupy a specific core position of helix and sheet pockets are plotted on a gray scale. The left column (1) show the frequencies normalized for each amino acid to represent amino acid preference for a position in a pocket. The right column (2) shows the frequencies normalized for each position in the pocket to represent the position preference for amino acid.

Figure 5. Generalized amino acid compositions of pockets. The distribution of generalized amino acid compositions for significant pockets - three α -helix (top) and four β -sheet (bottom) - are presented. Each of the twenty amino acids are grouped into non-polar (N), polar (P), and charged (C) according to hydrophobicity and electrostatic property. Each plot contains amino acid compositions that represent the least number of sequence compositions that can account for 75% of the counts observed for each pocket core. For each, the number unique sequences found forming the specific pocket, and the percentage of sequences required to account for 75% of the counts are summarized in the accompanying table.

Figure 6.

Knob distributions on the packing pocket types. Only top 5 most frequent pockets for α -helices and β -sheets are chosen. (a) Amino acid knob preferences for α -helix pockets (b) α -helix pocket preferences for amino acid knob (c) Amino acid knob preferences for β -sheet pockets (d) β -sheet pocket preferences for amino acid knob.

Figure 7.

Rotamer distribution of knobs in different types of pockets in α -helices and β -sheets. (a) Histograms of rotamers (gauche(-) (m), gauche(+) (p) and trans (t)) distribution for various amino acids packing into the top five pockets in β -sheets and α -helices. The amino acids with one side chain torsion angle are presented. The three amino acids, Phe, Val, and Ser, are chosen as representatives of the rotamer distributions. Thorough rotamer distribution of all the amino acids are presented in Supplementary data (Supplementary Figures S3a-f). (b) Set of histograms illustrating some of the exceptions to the general patterns described in (a). Trp rotamer distribution packing into SC Socket have less *trans* rotamer than the Phe rotamer. Additionally, the rotamer distribution for Trp packing into a SC Square has a significantly diminished amount of *trans* rotamer. The other amino acid that did not follow the trend was Cys.

Cys packing into SC Socket and MC Socket where the *trans* rotamer was more heavily populated.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 8.

Utility of the Knob-Socket and Pockets. The illustrative examples of core pockets and ancillary pockets are presented for four α -helix (**a** to **d**) and five β -sheet (**e** to **i**) pockets. The 4 α -helix pockets **a** to **b** originate from 1dly⁵⁶; **c** comes from 1ngk⁵⁷, and **d** from 1b0b⁵⁸. The **e** to **i** β -sheets were taken from 1lzl,⁵⁹ 1oa4,⁶⁰ 2joe,⁶¹ 1fdr,⁶² and 1d5t,⁶³ respectively. The ribbon diagrams for the pockets are given and their 2D representations are shown on the α -helix (top right) and β -sheet (bottom right) lattices. In the lattices, the knobs are labeled to identify the corresponding pocket shapes.

Table 1

Summary of results acquired through the pocket analysis. We provide the number of pockets identified in α -helix (N_H), β -sheet (N_S), coil (N_C), and turn (N_T) secondary structures packing with a knob from α -helix, β -sheet, and coil. Furthermore, we include the number of unique pocket types found in these four secondary structures; U_H , U_S , U_C , U_T .

Knob	N_H	N_S	N_C	N_T	U_H	U_S	U_C	U_T
Helix	77814	61100	7011	331	43	823	50	15
Sheet	23911	32072	7674	153	34	450	43	10
Coil	28899	50299	13919	437	43	774	57	14
Total	130624	143471	28604	921	51	1279	63	14

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Here we present the top ~90% most populated pockets found in α -helix (N_H), β -sheet (N_S), coil (N_C), and turn (N_T) secondary structures packed with α -helix, β -sheet, and coil knobs. We also include the counts for the two sockets in α -helices for comparison. The SC and MC sockets in β -sheets are included because they serve as core geometries for β -sheet pockets.

3

4

4

0

1

4

0

1

7

3

4

0

1

4

5

4

0

1

3

4

0

1

7

3

4

0

1

8

7

4

3

0

8

8

4

0

1

α -Helix
Pockets

Description	0-3-4	0-1-4	0-3-4-7	0-1-4-5	0-1-3-4	0-3-4-7-8	0-1-4-5-8
	High	Low	Diamond	Low high	High low	High diamond	Low diamond
Helix Knob	64568	51264	57066	8415	3245	3202	3082
Sheet Knob	27176	21623	18231	2827	834	814	707
Coil Knob	33611	41721	19322	4295	1534	1436	1238
Total	125355	114608	94619	15537	5613	5452	5027

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Description	SC Socket	SC Diamond	SC Square	MC Square	2 Strand Diamond	3 Strand Triangle	MC Socket
Helix Knob	27891	10684	7976	2142	1414	988	6864
Sheet Knob	16909	4038	3787	841	833m	570	3426
Coil Knob	25595	3111	4065	2319	856	756	10889
Total	70395	17833	15828	5302	3103	2314	21179

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Description	Coil Diamond Pocket	Ancillary Coil Diamond Pocket	Off Coil Pocket	Turn Diamond Pocket
Helix Knob	4,887	799	717	272
Sheet Knob	5,643	776	781	136
Coil Knob	10,133	1,469	1,305	357
Total	20,663	3,044	2,803	765