Published in final edited form as:

Proteins. 2015 December ; 83(12): 2147–2161. doi:10.1002/prot.24929.

An Amino Acid Code for Irregular and Mixed Protein Packing

Hyun Joo, Archana Chavan, Keith Fraga, and Jerry Tsai*

University of the Pacific, Department of Chemistry, Stockton, CA 95211

Abstract

To advance our understanding of protein tertiary structure, the development of the knob-socket model is completed in an analysis of the packing in irregular coil and turn secondary structure packing as well as between mixed secondary structure. The knob-socket model simplifies packing based on repeated patterns of 2 motifs: a 3 residue socket for packing within 2° structure and a 4 residue knob-socket for 3° packing. For coil and turn secondary structure, knob-sockets allow identification of a correlation between amino acid composition and tertiary arrangements in space. Coil contributes almost as much as α -helices to tertiary packing. Irregular secondary structure involves 3 residue cliques of consecutive contacting residues or **XYZ** sockets. In irregular sockets, Gly, Pro, Asp and Ser are favored, while Cys, His, Met and Trp are not. For irregular knobs, the preference order is Arg, Asp, Pro, Asn, Thr, Leu, and Gly, while Cys, His, Met and Trp are not. In mixed packing, the knob amino acid preferences are a function of the socket that they are packing into, whereas the amino acid composition of the sockets does not depend on the secondary structure of the knob. A unique motif of a coil knob with an **XYZ** β -sheet socket may potentially function to inhibit β -sheet extension. In addition, analysis of the preferred crossing angles for strands within a β -sheet and mixed α -helices/ β -sheets identifies canonical packing patterns useful in protein design. Lastly, the knob-socket model abstracts the complexity of protein tertiary structure into an intuitive packing surface topology map.

Keywords

Knob-socket motif; Helix-sheet packing; Coil packing; Protein 3° Structure

Introduction

Understanding the architecture of folded proteins has been an intriguing challenge for several decades.^{1–9} While backbone hydrogen bonds^{5,10–12} and torsion angles^{13,14} serve as fundamental principles for characterizing secondary (2°) structure, a general packing motif that allows similar classification of tertiary (3°) structure has been difficult to derive. The purpose of a 3° structure motif is to simplify the complexity of side-chain packing yet retain information about residues' roles in the protein fold. The challenge has been to identify a fundamental packing unit that is applicable across all types of structural classes. Prompted by a pseudo-heptameric sequence repeat, the first 3° packing motifs were identified between α -helices¹⁵. More recently, common side-chain arrangements have been found within

*Corresponding Author: University of the Pacific, Department of Chemistry, 3601 Pacific Avenue, Stockton, CA 95219, Tel: 1-209-946-2298, Fax: 1-209-946-2607.

families of proteins.^{16–20} However, these motifs only perform well for their specific 2° structure or fold family. For example, the most useful 3° structure motifs such as the knob-into-hole^{15,21,22}, helical wheel^{23,24} and ridges-into-groves²⁵ are limited to describing packing within one type of 2° structure. Broader analyses of 3° structure have dissected packing into smaller units. The most straightforward approaches have been to characterize local packing motifs in sequence along the peptide backbone^{26,27} or to describe 3° interactions based on the packing between regular 2° structure elements of α -helices, β -sheets, and turns.²⁸ Other approaches extracted 3° motifs from contact networks using graph theory approaches^{29,30} or identified repetitive 3° topologies^{31,32} from a rigorous definition of contacts determined by the Delauney tessellation.³³ While characterizing residue interactions, these identified 3° motifs lack the generality to perform as a basic unit of protein packing across all types of protein structure. In comprehensive analyses of residue packing within α -helices³⁴ and β -sheets³⁵, the knob-socket model clearly describes 3° packing as simple repetitive patterns of a single motif. To complete the demonstration of the knob-socket motif as a fundamental unit of 3° packing, the model is used to characterize the packing trends exhibited in the less regular turn and coil 2° structures as well as between different 2° structural elements mixed. For simplicity, coil and turn 2° structure is referred to as irregular secondary structure, while interactions between 2 elements of 2° structure is labeled as mixed 2° structure.

Turns and coil represent about half of all 2° structure in folded proteins, yet their contribution to protein packing has been a challenge to investigate. Both types of 2° structure are not easily described by a simple set of rules like the repetitive hydrogen bonding patterns of α -helices and β -sheets. The term “random” is often associated with coil structure. However, coil conformations are neither random nor disordered⁷, but are highly organized and firmly held in place. Because they form more irregular and non-repetitive configurations, previous analyses of turn and coil have focused mainly on classifying conformation. Turns have been categorized into certain types according to hydrogen bonding pattern and backbone conformation^{36,37}. Commonly considered a random polymer or denatured structure³⁸ that acts as a loop connecting two regular elements of 2° structure^{39,40}, coil conformations have been analyzed only in terms of the backbone (ϕ, ψ) torsion angles^{41–44}. These analyses demonstrate the difficulty in identifying coil and turn contributions to protein 3° packing. One goal of this work is to extend the knob-socket model and demonstrate that it provides a simple and intuitive motif to identify the role of coils and turns in 3° structure.

The only characterization of packing between mixed 2° structural elements has been between α -helices and β -sheets. Even though α -helices and β -sheets are regular, their interactions have been difficult to characterize. Unlike the repeated sequence cues of packing between α -helices¹⁵, the variation inherent in β -sheet structure⁴⁵ has made identifying packing patterns involving β -sheet difficult. Therefore, general geometrical features such as distance between axes and crossing angles have been the most common 3° structure analysis of β -sheet^{46–48} and α -helix/ β -sheet packing^{48–50}. A residue level accounting found individual amino acid preferences for packing between α -helices and β -sheets⁵¹, and such residue preferences correlate with the complementary hydrophobic patterns on the α -helices in the helix-sheet interface⁵². Applying the knob-socket model in

the characterization of helix-sheet 3° packing extends the concept of complementary hydrophobic pattern matching to relate the primary sequence to a 3° packing surface⁵³.

As developed previously,^{34,35} the knob-socket model directly relates amino acid sequence composition to structural packing state by identifying order in protein 3° structure as repetitive patterns of a basic unit. Although the approach takes advantage of the graph theory's reduction in complexity of protein packing⁵⁴ into residue interaction network,⁵⁵ these contact networks themselves are too complex and unique to be used as a general descriptor of residue packing structure. In contrast, clustering contacts around individual amino acid types⁵⁶ or even clustering residue cliques (groups of residues that all contact each other) from the residue interactions networks⁵⁷ did not reveal enough specificity to find any order to residue packing in 3° structure. The breakthrough came from a contact order classification and grouping of the residue packing cliques that identified the 3 residue socket and the 4 residue knob-socket as the primary constructs to describe packing structure in proteins (Figure 1). The other contact order types of residue cliques were found to result from and therefore be derivative of patterns of knob-socket packing, as pointed out in more detail in the discussion of Figure 2 below. Because the 3 residue sockets describes the portion of the residue interaction network that are not involved in protein 3° structure while patterns of the 4 residue knob-sockets define 3° protein packing, the amino acid composition preferences of these two constructs identifies a code between protein sequence and structure.

To complete the analysis of 3° packing structure begun on the regular 2° structures of α -helices³⁴ and β -sheets³⁵, the knob-socket model is applied to an analysis of packing in the irregular coil and turn 2° structure and between different types of or mixed 2° structure within the Protein Data Bank (PDB)⁵⁸. The knob-socket model's ability to clearly describe the packing between irregular and regular structures will be demonstrated. In particular, residue propensities of knob-socket motifs define an amino acid code relating the primary sequence to 3° conformation and also, new canonical packing patterns are characterized. As shown in Figure 1, the knob-socket model describes packing structure in 2 parts: sockets and knobs packing into sockets. In the knob-socket model, the 3 residue sockets describe the packing within 2° structure and the members are considered 'local' to the 2° structure, although the hydrogen bonded residue **H** is farther away in sequence, especially in β -sheets (Figure 1a, 2nd column). Indicating a 'nonlocal' interaction in the knob-socket model, the knob packing from another 2° structure element into a socket defines 3° structure packing (Figure 1, bottom). Because the knob-socket construct reduces the multiplicity of residue packing to a simple one to one correspondence between a knob and a socket, the knob-socket motif is shown to be a fundamental unit of packing that allows the intelligent interrogation of protein 3° structure.

Methods

All protein structures were obtained from the Protein Data Bank⁵⁸. Protein 2° structure was assigned by DSSP⁵⁹ using the 2006 CMBI version written in C, although a new C++ version of DSSP has been released.⁶⁰ Because the newer C++ version of DSSP only changed 0.15% of the 2° assignments, the CMBI 2006 version of DSSP was used to be consistent with the previous knob-socket analyses.^{34,35} The UCSF-Chimera molecular graphics package⁶¹ was

used to visualize protein structures. All the analyses were carried out by using PERL scripts and plots including heat maps were made using the R statistics package⁶².

Packing Clique Calculation

Contacts were calculated between non-bonded heavy atoms using an implementation^{63,64} of the Voronoi polyhedra⁶⁵. This Delaunay tessellation³³ of atom interactions allowed the construction of a contact graph at the residues level as done previously.^{34,35} So, in the knob-socket model, the definition of two residues packing against each other is that they have at least one atom in contact. Packing between all side-chains is considered, as well as packing between main-chain and side-chain atoms for all non-neighboring residues, where neighbors are less than 3 residues away in sequence. The maximal clique detection method⁶⁶ identified sets of residues that all contact each other: packing cliques. Packing cliques were found for all 16,673 protein domains in the SCOP ASTRAL 1.75 set⁶⁷ of structures filtered at 95% sequence identity. While the 95% cutoff includes a certain amount of redundancy, the goal has been to sample as deeply into the 8,000 possible 3 residue sockets and 160,000 possible 4 residue knob-sockets for each of the 102° structure types and their combinations. Therefore, this structure set is consistent with the previous knob-socket analyses on α -helices and β -sheets, so that the relative amino acid composition of knob-socket residues in different 2° structures can be directly compared. Instead of using direct B-factors and resolution values, the data set was originally filtered at 0.4 using the SPACI (summary PDB ASTRAL check index) score.⁶⁸ The structural data was downloaded directly from the SCOP website⁶⁹ that provides the atomic coordinate files.

Packing Clique Data Set

From the 16,673 domains, a total of 6,677,903 packing cliques are found. 6,258,849 (94%) packing cliques consist of 3 (1,505,208) or 4 (4,753,641) residue packing cliques. These cliques are further classified based on the 2° structure of their constituent residues (Table 1). Because the only α -helix packing cliques and the only β -sheet packing cliques were analyzed previously^{34,35}, a total of 4,492,432 packing cliques are analyzed in this work. For the analyses of irregular 2° structure, 385,825 coil and turn packing cliques were analyzed. This leaves a total of 4,106,067 packing cliques that are used in the analysis of mixed packing structure. As described previously⁵⁷, packing cliques are further classified based on contact order⁷⁰. For example, a 3 residue clique packing of all local residues is a **3** packing clique, while a 4 residue clique with 3 local residues and 1 non-local residue is a **3+1** packing clique.

Knob-socket definition

A knob-socket motif is the tetrahedral **3+1** packing clique with 3 local residues from one 2° structure and 1 non-local residue from another (Figure 1). For α -helices³⁴ and β -sheets³⁵, the knob-socket motif is defined as **XY:H+B**. The **XY:H** residues constitute the local socket, where the “:” indicates a hydrogen bonded residue, and **B** is the knob. For the irregular coil and turn 2° structure, the knob-socket motif is defined as **XYZ+B**. The **XYZ** socket is made up of all local residues packed into a knob **B** residue. This **XYZ** socket is further classified based on residue interactions (Figure 3): all residues in contact is a closed

socket (**K**); the first and last residues not in contact is an open socket (**O**); and any neighboring residues not in contact is classified as a strained socket (**S**). These **K**, **O**, and **S** socket frequencies and their amino acid compositions were calculated. In coils, 498,069 **K**, 340,324 **O**, and 114,249 **S** sockets are found and in turns 107,544 **K**, 161,034 **O**, and 136,480 **S** sockets were found.

Inter-strand and helix-strand crossing angles

From the knob-socket residue pairs, crossing angles between the inter-sheet strands of β -sheets and between α -helix and strands of β -sheets were calculated. The vectors for each strand were calculated over the region of interaction between the knob and socket. For residues i to j ($i < j$ in sequence) on one strand packing with l to m ($l < m$ in sequence) on another, the crossing angle is the torsion angle made by the two C α atom vectors. If on one strand only one residue k is involved with packing, this vector is drawn between the C α atoms of $k-1$ and $k+1$ residues. The helix-strand crossing angles were calculated by calculating the torsion angle between helical axis and the strand. The helical axes are calculated according to the HELANAL method^{71,72}. The angle frequencies were presented from -180° to $+180^\circ$ in 5° bins.

Result and Discussion

The Significance of Coil and Turn Packing

The substantial packing contribution of coil and turn argues for a rigorous knob-socket analysis of irregular 2° structure. As classified in Table 1, the total number of 3 and 4 residue cliques involved in packing of mixed 2° structure is 708,615 and 3,397,992, respectively. 81% of the 3 residue cliques and 77% of these 4 residue cliques are involved with at least 1 coil residue. Both of these percentages are surprisingly higher than the comparable packing cliques consisting of at least 1 helix or 1 sheet residue. With 35% 3 residue cliques and 26% 4 residue cliques, turn residues are less involved in packing. The significant involvement in packing cliques indicates that the irregular coil plays as an important role in protein 3° structure as α -helices and β -sheets. To be a major participant in 3° structure, coil residues should exhibit regular packing clique patterns, which is a departure from the common view of coil as random conformations.

The distribution of coil and turn packing cliques exhibits similar sockets and knob-sockets found previous for α -helices³⁴ and β -sheets.³⁵ Each clique is further classified based on contact order of residues, where residues on the same 2° structure element are grouped together. For both coil and turn the 3 and 4 residue packing cliques dominate over the other packing clique sizes (Figures 2a and b, respectively). The coil packing clique distribution (Figure 2a) displays a predominance of 3 (36%) and 4 (55%) residue packing cliques. Representing 2° structure packing, the “**3**” socket clique of all local residues is 58% of the 3 residue packing cliques in coil. For 3° interactions with coil, the four residue “**3+1**” knob-socket packing clique is most prevalent at 42%. Turn packing cliques (Figure 2b) mainly favor 3 residue cliques at 64% with 4 residue cliques at 16%, indicating that turns seldom interact with other 2° structure elements. Majority of 3 residue turn cliques are the local “**3**” sockets (63%) and only 29% of the 4 residue turn cliques are “**3+1**”. As has also been found

previously^{34,35}, the other types of contact order packing cliques are either derivatives of the “3+1” patterns like the “2+2” and “2+2” or are insignificant in number like the “1+1+1” and “1+1+1+1”. The knob-socket model simplifies protein 3° structure into 3 and 4 residue packing cliques, and together, the socket and knob-socket cliques provide constructs to investigate the 3° packing characteristics of coils and turns (Figure 1).

Knob-Socket Model of Irregular 2° Structure

From the packing cliques, the socket and knob-socket motifs describing coil and turn should be identifiable. The **XY:H** socket for regular 2° structure consists of two covalently attached residues **X** and **Y**, and one hydrogen bonded residue **H** (Figure 1). Because of a lack of repetitive hydrogen bonding, the irregular coil and turn sockets are different and are designated by their order in sequence of **X**, **Y**, and **Z** to make an **XYZ** socket. Of the 3 residue packing cliques, a predominant 74% (138,814 coil and 26,164 turn sockets) involves 3 consecutive residues *i*, *i*+1, *i*+2 (Figure 2b). While the consecutive socket is prevalent in both coil and turn, turn cliques exhibit a higher prevalence of the next 2 socket types due to their inherent bend and occasional hydrogen bonding. The next highest clique at 14% includes a separation of 2 residues (*i*, *i*+2, *i*+3 or *i*, *i*+1, *i*+3), while the sockets with a 3 residues separation (*i*, *i*+1, *i*+4 and *i*, *i*+3, *i*+4) make up 7%. The remaining 3 residue coil and turn sockets consisting of residues farther separated in sequence constitute ~5% of the population, and are therefore not considered as coil or turn sockets. As the main socket type in coil and turn, the consecutive *i*, *i*+1, *i*+2 socket is the primary representative of the **XYZ** coil and turn socket. On their own, these **XYZ** sockets are unfilled or free of 3° packing and shown in Figure 1 columns 3 and 4. Like the previously characterized sockets^{34,35}, the coil and turn knob-socket adds a knob **B** residue that packs into the surface made by an **XYZ** socket to describe 3° interactions. For mapping of 3° structure, this “3+1” four residue clique of irregular 2° structure can be projected into two dimensions (2D) as has been done for regular 2° structure (Figure 1, right column).

The contact pattern for these consecutive *i*, *i*+1, *i*+2 sockets exhibits three main types of variations (Figure 3). The closed **K** sockets occur when the three residues all contact each other in extended conformations of coil and in tighter portions of turns. Without non-covalent contact between the first and last residues, the open **O** sockets are observed before or after the bending of the coil and turn structure. In the last strained **S** sockets, consecutive residues' side chains do not interact, because a stressed peptide backbone occludes the side chains from each other. The closed **K** sockets are the most common in coil, while the open **O** socket is most common in turns. The strained **S** socket is the least common type for both.

For each of the **K**, **O**, and **S** sockets in coils and turns, the percent amino acid frequencies at each **X**, **Y**, and **Z** position are shown as heat maps in Figure 3. Although coil and turn sockets are depicted somewhat differently, their amino acid distributions suggest that they are more similar than different. Across all the plots, Cys, His, Met and Trp are rarely observed, as expected due to the low prevalence of Cys, Met, and Trp in protein sequences⁷³. As a slight exception, His occurs at the **X** and **Y** positions of strained **S** sockets. In the closed **K** socket, the population is relatively even and similarly distributed between coils and turns. The primary difference between them is the expected preference of

turn sockets for Gly at the **X** and **Y** positions and for Pro at the **Z** position. At the **X** position, Pro, Asp, and Ser are also commonly observed with Gly. Tyr is disfavored at the **Y** position, while Val and Glu are commonly observed at the **Z** position. In open **O** sockets, the striking feature for both coil and turn is that Gly favors all three positions, and especially the **X** and **Z** positions. In the strained **S** sockets, the distributions for coil and turn are comparable and relatively even over all amino acids. While both favors Pro in the **X** position and Ala in the **Y** and **Z** positions, Gly and Ile are not favored at the **Y** position.

To complete the modeling of coil and turn structure, the knobs that pack into the **XYZ** sockets to form the “**3+1**” packing cliques need to be characterized. Figure 4 shows the distribution of coil and turn knob preferences for the 20 amino acids. Both coil knobs packing into coil sockets and turn knobs packing into turn sockets exhibit uniform distributions, unlike α -helix and β -sheet which show clear tendencies for hydrophobic knob residues. For coil knobs packing into coil sockets, the amino acid preference in order is Arg, Asp, Pro, Asn, Thr, Leu, and Gly, while Met, Cys, Trp, and His are the least favored. For turn knobs packing into turn sockets, the amino acid preference orders is Arg, Gly, Glu, Lys, Ser, Asn, and Thr, while Trp, Met, His, Ile, Cys and Phe are the least favored. For both coil and turn, the favored amino acids are an inconsistent mix of small and long amino acids. They both like Arg and Gly, but coils prefer Asp and turns prefer Glu, which differ by only 1 methylene group. The disfavored amino acids are more easily grouped as primarily large and bulky, in spite of the inclusion of the hydrophobic Ile.

These differences and similarities at the position level are reflected in the amino acid composition of the top sockets and knob-sockets for coil and turn (Figure S1), although both exhibit certain unique combinations. For sockets, coil produces GVP, GIP, LGL, GAD, and GLP as the top 5 sockets, while turn has DGT, DGS, DGK PGK, and LGL. Both involve Gly, but coil prefers the Gly at the **X** position, while turn prefers Gly at the **Y** position. In fact, the preference of turns for Gly at the **Y** position is so strong that another amino acid is not found in turn sockets until an Asp in the 61st packing clique on the list. For complete knob-socket packing motifs, coil structures exhibit high frequencies for the LGL+L, AGV+L, LGV+L, VPD+R, GDS+G, and GAD+K. Turn knob-sockets favor PGK+R, PGQ+R, PSG+V, EDL+N, DGK+S, and DGT+N. These biases for certain amino acid combinations in the knob-socket motif indicate that they possess a definable packing structure. While not surprising for the more ordered turn, coil has been thought to exhibit more random conformations. These results show that the knob-socket model is able to discover an underlying order in the packing of coil and turn.

Knob-Socket Packing in Mixed 2° Structure

Figure 2c shows the distribution of the ~4.4 million packing cliques that occur between different elements of 2° structure. Clearly, the ~3.4 million (77%) 4 residue clique is most populated, and of these, the knob-socket “**3+1**” motif is the most prevalent at ~1.4 million (40%). Consistent with previous work and the coil and turn clique distributions, the predominance of the “**3+1**” knob-socket clique for interactions between 2° structure elements strongly supports the knob-socket motif as the fundamental unit of packing in protein 3° structure. Once more, the other types of cliques are found to be either derivative

or too small in number to be significant. Table 2 classifies the “3+1” mixed 2° structure packing cliques according to the various pair combinations of knobs and sockets. Because these results are influenced by 2° structure representation in the PDB, it is not surprising that α -helix cliques are prevalent and turn cliques are not. Most remarkably, α -helix/coil cliques (16%) are almost as prevalent as the interactions of α -helix/ β -sheets (22%).

Investigating deeper into the data, knob selection depends on the type of socket that it is packing into, whereas the composition of sockets depends on the type of 2° structure. The coil and turn knob dependence can be found by comparing the residue distributions based on socket type (Figure 4). The knob distributions group largely according to the socket type that the knob packs. Knobs packing into the regular α -helices and β -sheets exhibit similar distributions with high counts for the hydrophobic residues of Leu, Phe, Ile, Val, and Tyr, where β -sheets show more counts with the large hydrophobic Trp. These regular sockets also disfavor polar and charged amino acids. Knobs packing into irregular coil and turn produce more uniform distributions with higher values for Arg, Asn, Gly and Pro. For specific preferences, coil sockets like to pack with Asp, but do not pack well with Trp, Ile, Met, His, and Cys. Mirroring the preferences seen by the coil and turn knobs, the percent frequencies of β -sheet knobs packing into α -helix sockets, and α -helix knobs packing into β -sheet side chain (SC) and main chain (MC) sockets are presented in Figure 5a. For both α -helix and β -sheet, Leu is the most frequently observed knob packing into a socket, which is the same result found for both types of regular 2° structure packing^{34,35}. These distributions are equally dominated by the 3 non-polar residues Leu, Ile and Val that contribute ~56% of all knobs packing into sockets from α -helices and β -sheets, matching earlier work^{74,75}. Interestingly, the hydrophobic/aromatic side groups of Tyr and Phe also exhibit high preferences as β -sheet knobs in comparison to Ala. The residue Ala is especially more common in α -helices than in β -sheets. Knobs do affect sockets in more subtle ways. Figure S2 shows that knobs from the regular 2° structure sample tighter distributions of socket types, whereas a knob from irregular 2° structure samples a broader distribution of socket types.

By contrast, 2° structure determines the residue composition of sockets. As the most common mixed interaction (Table 2), the interactions between α -helices and β -sheets is used as an example socket analysis for mixed 2° packing. While a number of specific features are revealed, the discussion is limited to the general trends. Overall, in terms of **XY:H** sockets, the composition and distribution in Figure 5b are almost exactly like the results from our previous investigations of α -helix only packing³⁴ and β -sheet only packing³⁵, as illustrated by the plots in Figure S3. The same is generally found in the sockets for the other mixed 2° structure combinations (data not shown). In the same analysis done for irregular 2° structure (Figure 3), distributions per **X**, **Y** and **H** position were calculate for the mixed packing α -helical and β -sheet sockets. The composition of α -helical socket positions (Figure 5b, top) is similar to previous results with the dominance of non-polar, aliphatic side-chains³⁴. Leu and Ala are the most represented amino acids at over 17%, and Ile and Val are second at ~10% each. Arg, Glu, and Lys residues (5% each) exhibit a notable preference for the **Y** position, while Thr and Met (~4% each) prefer the **H** position. The compositions of β -sheet side-chain and main-chain sockets positions (Figure 5b, bottom) also exhibit distributions

similar to the packing between β -sheets³⁵, where Val, Ile, and Leu make-up the majority of residues in both socket types. For the side-chain sockets, the C β -branched Thr has slightly better frequency at the **Y**-position at 10%, and Lys and Arg prefers the **X** position. In β -sheet main-chain sockets, the composition are generally sparse and particularly sparser at the **Y** position, since the side-chain atoms at the **Y** position does not contribute to main-chain socket packing. Also, Gly at ~9% prefers the **X** and **Y** position over the **H** position at ~5% that creates a void space to accommodate the packing of the larger aromatic residues. In addition, residues forming sockets but not acting as knobs are defined as edge residues, and the edge residue frequencies are calculated by excluding knobs from the sockets. Figure S4 displays a clear difference between the residues in α -helices and β -sheets. Interestingly, mostly hydrophilic residues (Ala, Glu, Lys, Arg, Gln, and Asp) occupy the edge in α -helices, while β -branched (Val & Ile) and aromatics (Phe & Tyr) are edge residues in β -sheets. This sort of bias has direct implications on protein design.

This knob-socket analysis of mixed 2° structure packing also reveals some novel interactions. In Figure 6a, a helix/sheet junction occurs when a single α -helix packs perpendicularly along both sides of a β -sheet. A single knob residue from the β -sheet packs into an extended diamond pocket on the α -helix, while the helical diamond pocket contributes 2 knob residues that each pack into sockets on opposite sides of the β -sheet. While this arrangement is found rarely, a more common configuration involves a knob (usually from coil) packing into an **XYZ** β -sheet (Figure 6b) or β -sheet/coil (Figure 6c) socket. Both of these β -sheet **XYZ** sockets occur on the exposed edge strand of a β -sheet and seem to terminate β -sheet extension by inhibiting regular hydrogen bonding. Adding to the previously characterized β -sheet edge protection⁷⁶, the knob-socket model provides specific detail of the structure and amino acid composition of these edge-capping interactions. Figure S5 plots the knob preferences against the all β -sheet socket (E3) and the 2 β -sheet/1 coil socket (E2) in two left panels. Generally, the E3 composition prefers Gly, Ser, and Val residues that pack Val and Thr knobs, but also specifically, the Tyr, Cys, and Ala E3 socket likes Trp knobs. The E2 composition also favors Val, but exhibits a broader diversity that is specific for certain knobs. For example, the top Thr, Ile, and Ser E2 socket almost exclusively packs with Arg knobs, whereas the Val, Gln, and Leu E2 socket packs with Tyr knobs. To demonstrate the specificity of edge strand packing, amino acid distributions of sockets in middle β -strands are shown on the right panel of Figure S5 to compare with sockets on edge β -strands. The 2 panels clearly show that there are differences in socket amino acid composition between middle β -strands and edge β -strands as well as specific preferences toward knob residues. The middle β -strand sockets prefer to pack with Leu, Ile, and Val, but sockets on edge β -strands show little preference toward those aliphatic amino acids. Therefore, different combination of sockets in the β -strand can lead to β -sheet extension or termination. This insight into β -sheet edges could be helpful in analysis of protein mutations and design of protein structure.

Canonical Helix-Sheet and Sheet Packing

Canonical packing patterns between α -helices were identified with the help of a pseudo-repetitive 1° sequence pattern¹⁵, which led to further identification of canonical α -helical packing patterns based on crossing angles^{24,25}. The knob-socket model provides an elegant

and intuitive representation of not only the canonical but also the non-canonical packing between α -helices³⁴. Using the knob-socket model to produce a 2D projection of packing, a topological packing surface map is used to identify canonical patterns between the other combinations of regular 2° structure at preferred crossing angles within β -sheets and between α -helices and β -sheets.

A histogram of the crossing angle between two β -strands (Figure S6A) closely matches the mostly recently calculated distributions^{46–48}. The two dominant peaks at -30° and 150° represent the same packing orientation in parallel and antiparallel, respectively, while the two smaller peaks at -165° and 25° describe a different packing orientation in antiparallel and parallel, respectively. Although specific sequence preferences were not found with crossing angles, certain canonical knob-socket patterns related directly to crossing angle. Using the knob-socket model, Figure 7a maps the canonical packing topology for the dominant -30° and 150° β -strand crossing angles. The β -strand crossing angles at -165° and 25° are essentially the inversion of Figure 7a. A knob residue packing into 2 adjacent **XY:H** sockets forms the basic unit of a repetitive pattern. The 2 adjacent **XY:H** sockets consist of 4 residues spanning 3 strands: a 3 stranded diamond pocket⁷⁷. This β -sheet unit of a knob into pocket packing is reminiscent of Crick's α -helical coiled-coil¹⁵. Successive parallel (j and $j+2$) or anti-parallel ($k+2$ and k) knob residues from one β -sheet essentially form a ridge that packs into consecutive 3-stranded diamond pockets that create a groove between the diagonal **Y---H** lines on the other β -sheet. This type of packing echoes the ridge-into-groove parallel and anti-parallel packing of α -helices²⁵. As noted with α -helical packing, the ideal abstraction in Figure 7a is a useful reference point for understanding β -sheet packing, but realistic β -sheet packing deviates significantly. A genuine depiction of the intricacies of β -sheet packing is shown in Figure S7, where certain residues like W36 and I72 overshadow the canonical packing pattern. However, the consistent simplicity of the canonical pattern in Figure 7a provides a useful reference for rational design of β -sheet structure.

The packing between α -helices and β -sheets can be categorized into 2 types of canonical knob-socket patterns. The crossing angle histogram between α -helices and β -strands (Figure S6B) shows preferences with peaks at -160° , -35° , and 155° , where this asymmetric distribution is consistent with recent studies^{48–50}. The peaks result from 2 types of canonical knob-socket packing where the α -helix aligns with the β -sheet (Figure 7b) or the α -helix packs at an angle to the β -sheet (Figure 7c). The aligned α -helix/ β -sheet are the most common orientation with peaks closest to parallel -160° and anti-parallel 155° with each other. The packing can be explained as ridges of knobs that pack into grooves of sockets. Knobs from each strand in the β -sheet pack into adjacent high **H:YX** and low **XY:H** sockets forming an α -helical diamond pocket³⁴, essentially Crick's α -helical hole¹⁵. A ridge of β -sheet knobs packs into a groove of helical diamond pockets along the ± 4 residue lines. Concomitantly, the knobs from the α -helix are as 2 lines of ± 4 residue ridges packing into a β -sheet groove of successive three stranded diamond pockets along the **Y---H** line similar to β -sheet packing of Figure 7a. Alternatively, the α -helical knobs could be considered as 3 lines of ± 3 residue ridges that pack into a groove between the **X---H** lines. While Figure 7b is an ideal depiction, the mapping of the aligned α -helix and β -sheet packing from 1kjn⁷⁸ exhibits some variation but does not deviate significantly from this ideal (Figure S8A). The

–35° α -helix/ β -sheet packing angle produces the canonical pattern depicted in Figure 7c. The knobs along the central ± 4 ridge of the α -helix packs into the groove formed by a series of β -sheet **XY:HG** square pockets^{35,77}. The α -helical residues on the next ± 4 ridge pack in the groove formed by the 2 stranded diamond sharing **X---H** line. Conversely, the ridges of knobs from each β -strand pack into the grooves formed by a series of α -helical diamond pockets bordered by the ± 4 ridges. As shown in Figure S8B, the mapping of a real packing surface between an angled α -helix and β -sheet from 1uds⁷⁹ shows strong consistency with the ideal shown in Figure 7c. For both types of packing, the knob-socket mapping accurately reveals the canonical packing pattern and captures the counterintuitive packing due to the α -helix's 100° or 3.6 per turn residue periodicity⁵. For the aligned orientation in Figure 7b, the residue periodicity requires the α -helical knobs to pack at an angle to the β -sheet and for the angled orientation in Figure 7c, the α -helical knobs to pack aligned to the β -sheet. These knob-socket based canonical patterns provide a clear guide for protein design that matches closely with real packing patterns.

Knob-Socket Packing Surface Topology Map of Protein 3° Structure

One of the inherent difficulties in explaining or analyzing protein 3° structure is producing a representation that captures and simplifies the many-bodied interactions of packing. As shown in the top of Figure 8 for the Src homology 2 (SH2) domain⁸⁰, the detail of an all atom model obscures an understanding of packing structure, while a ribbon diagram or 2° structure topology⁶ provides too little information. The knob-socket model allows the projection of the three dimensional (3D) packing into an intuitive two dimensional (2D) map called a packing surface topology map (Figure 8, bottom). Overlaid on a depiction of the regular as well as irregular 2° structure, 3° packing is clearly represented by colored knobs on top of grey sockets, while the yellow sockets indicate 2° structure packing. This abstract mapping organizes the 3° structure into one to one knob-socket correlations. Residue packing patterns between elements of 2° structure can easily be discerned like the coil knobs V63, F70, L96, and L85 packing into the last α -helix. The packing surface topology also reveals the specificity and intricacy of 3° structure by identifying the regular 2° structure surfaces upon which knobs can pack. While the ribbon diagram suggests that the regular 2° structure dominates packing interactions, the topology map clearly shows the somewhat ubiquitous coil packing as filler around the more rigid regular 2° structure. Edge packing is found on both sides of the β -hairpin and one side of the three stranded β -sheet. Interestingly, like β -sheets, coil residues can also have 2 sides to their packing surface as shown by residues from G1 to S15. Figure 8 demonstrates that the knob-socket model simplifies protein 3° structure into digestible units of packing motifs, yet retains the intricacies of residue packing.

Conclusions

As a continuation of previous work^{34,35,77}, this study completes the development of the knob-socket model by characterizing the packing motifs of irregular coil and turn 2° structure as well as the interactions between mixed 2° structure. Because the knob-socket model is descriptive of protein packing across all protein 2° interactions, the knob-socket motif can function as a fundamental unit to describe and abstract protein 3° structure like the

covalent bond performs for 1° structure and main-chain hydrogen bonds perform for 2° structure. The knob-socket motifs of 2° structure shown in Figure 1 simplify the complexity of residue 3° packing networks into simple repetitive patterns of knobs and knob-into-sockets for all protein structures. Therefore, the knob-socket model advances our understanding of protein 3° structure by revealing a direct relationship between the sequence and 3° structure (Figures 3, 4 and 5). While these preferences derive from the protein folding in water solvent and largely follow the chemical nature of each amino acid, the distributions exhibit variation outside of a strict hydrophobic/hydrophilic definition. In particular, the knob-socket model identifies specific combinations of amino acid residues that prefer to pack with each to do not always fit a strict hydrophobic/hydrophilic definition (see Figures S1, S2, S3, and S5). Furthermore, these knob-socket motifs provide constructs and a language to intelligently describe and interrogate the complexity of protein 3° structure packing. In terms of description, the packing motifs clearly reveal canonical packing patterns within β -sheets as well as between mixed α -helices/ β -sheets (Figure 7). Moreover, the knob-socket characterization of the canonical packing for the 2 prevalent orientations of helix-sheet crossing angles displays a regular if not somewhat counterintuitive organization of side-chain packing. Using these motifs to project the 3D tetrahedral packing interactions into 2D space produces a simplified depiction of packing structure called a packing surface topology map (Figures 8). As these representations of protein 3° structure demonstrate, analysis using the knob-socket model allows a precise deconstruction of packing's role in determining conformation, and moreover allows rational interrogation the contributions of residues to protein 3 structure. The knob-socket model accomplishes this insight by simplifying the interpretation of packing from a complex many-bodied network of residue interactions to one to one correspondence between knobs and sockets. Because they are discrete yet general units of packing, the knob-socket motif is able to produce meaningful analysis of all protein 3° structure. As an example, the knob-socket analysis has identified specific knob-sockets compositions and configurations that limit β -sheet growth (Figure 6). Also, the analysis clearly points to the significant contribution of coil interactions to the overall 3° structure (Tables 1 and 2 and Figure 8). These results indicate that the knob-socket model offers a productive method to better understand protein 3° structure.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by an NIH NIGMS grant GM104972.

References

1. Pauling L, Corey RB. Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. *Proc Natl Acad Sci U S A*. 1951; 37(11):729–740. [PubMed: 16578412]
2. Pauling L, Corey RB. The structure of fibrous proteins of the collagen-gelatin group. *Proc Natl Acad Sci U S A*. 1951; 37(5):272–281. [PubMed: 14834150]
3. Pauling L, Corey RB. The structure of hair, muscle, and related proteins. *Proc Natl Acad Sci U S A*. 1951; 37(5):261–271. [PubMed: 14834149]

4. Pauling L, Corey RB. The structure of feather rachis keratin. *Proc Natl Acad Sci U S A*. 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. *Proc Natl Acad Sci U S A*. 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. Pauling L, Corey RB. The polypeptide-chain configuration in hemoglobin and other globular proteins. *Proc Natl Acad Sci U S A*. 1951; 37(5):282–285. [PubMed: 14834151]
11. Pauling L, Corey RB. The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A*. 1951; 37(5):251–256. [PubMed: 14834147]
12. Pauling L, Corey RB. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proc Natl Acad Sci U S A*. 1951; 37(5):235–240. [PubMed: 14834145]
13. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*. 1963; 7:95–99. [PubMed: 13990617]
14. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. *Advances in protein chemistry*. 1968; 23:283–438. [PubMed: 4882249]
15. Crick FHC. The Packing of α -Helices: Simple Coiled-Coils. *Acta Cryst*. 1953; 6(8–9):689–697.
16. Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of molecular biology*. 1987; 193(4):775–791. [PubMed: 2441069]
17. Bandyopadhyay D, Huan J, Prins J, Snoeyink J, Wang W, Tropsha A. Identification of family-specific residue packing motifs and their use for structure-based protein function prediction: I. Method development *Journal of computer-aided molecular design*. 2009; 23(11):773–784. [PubMed: 19543979]
18. Huan, J.; Bandyopadhyay, D.; Prins, J.; Snoeyink, J.; Tropsha, A.; Wang, W. Distance-based identification of structure motifs in proteins using constrained frequent subgraph mining. *Computational systems bioinformatics/Life Sciences Society Computational Systems Bioinformatics Conference*; 2006; p. 227-238.
19. Huan J, Bandyopadhyay D, Wang W, Snoeyink J, Prins J, Tropsha A. Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *Journal of computational biology: a journal of computational molecular cell biology*. 2005; 12(6): 657–671. [PubMed: 16108709]
20. Ku SY, Hu YJ. Discovery of structural motifs using protein structural alphabets and 1D motif-finding methods. *Advances in experimental medicine and biology*. 2010; 680:117–123. [PubMed: 20865493]
21. 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]
22. 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]
23. Schiffer M, Edmundson AB. Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophysical journal*. 1967; 7(2):121–135. [PubMed: 6048867]
24. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. *Science*. 1998; 282(5393):1462–1467. [PubMed: 9822371]
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. Jonassen I, Eidhammer I, Conklin D, Taylor WR. Structure motif discovery and mining the PDB. *Bioinformatics* (Oxford, England). 2002; 18(2):362–367.
27. Tyagi M, de Brevern AG, Srinivasan N, Offmann B. Protein structure mining using a structural alphabet. *Proteins*. 2008; 71(2):920–937. [PubMed: 18004784]
28. Shi S, Zhong Y, Majumdar I, Sri Krishna S, Grishin NV. Searching for three-dimensional secondary structural patterns in proteins with ProSMoS. *Bioinformatics* (Oxford, England). 2007; 23(11):1331–1338.
29. Basu S, Bhattacharyya D, Banerjee R. Mapping the distribution of packing topologies within protein interiors shows predominant preference for specific packing motifs. *BMC bioinformatics*. 2011; 12:195. [PubMed: 21605466]
30. Hleap JS, Susko E, Blouin C. Defining structural and evolutionary modules in proteins: a community detection approach to explore sub-domain architecture. *BMC structural biology*. 2013; 13:20. [PubMed: 24131821]
31. Wako H, Yamato T. Novel method to detect a motif of local structures in different protein conformations. *Protein engineering*. 1998; 11(11):981–990. [PubMed: 9876918]
32. Bostick DL, Shen M, Vaisman II. A simple topological representation of protein structure: implications for new, fast, and robust structural classification. *Proteins*. 2004; 56(3):487–501. [PubMed: 15229882]
33. Delauney B. Sur la sphère vide. *Bull Acad Sci USSR (VII) Classe Sci Mat Nat*. 1934:783–800.
34. 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]
35. Joo H, Tsai J. An amino acid code for beta-sheet packing structure. *Proteins*. 2014; 82(9):2128–2140. [PubMed: 24668690]
36. Koch O, Klebe G. Turns revisited: a uniform and comprehensive classification of normal, open, and reverse turn families minimizing unassigned random chain portions. *Proteins*. 2009; 74(2):353–367. [PubMed: 18704930]
37. Panasik N Jr, Fleming PJ, Rose GD. Hydrogen-bonded turns in proteins: the case for a recount. *Protein Sci*. 2005; 14(11):2910–2914. [PubMed: 16251367]
38. Smith LJ, Fiebig KM, Schwalbe H, Dobson CM. The concept of a random coil. *Residual structure in peptides and denatured proteins. Folding & design*. 1996; 1(5):R95–106. [PubMed: 9080177]
39. Joo H, Chavan AG, Day R, Lennox KP, Sukhanov P, Dahl DB, Vannucci M, Tsai J. Near-native protein loop sampling using nonparametric density estimation accommodating sparsity. *PLoS Comput Biol*. 2011; 7(10):e1002234. [PubMed: 22028638]
40. Choi Y, Agarwal S, Deane CM. How long is a piece of loop? *PeerJ*. 2013; 1:e1. [PubMed: 23638343]
41. Fitzkee NC, Fleming PJ, Rose GD. The Protein Coil Library: a structural database of nonhelix, nonstrand fragments derived from the PDB. *Proteins*. 2005; 58(4):852–854. [PubMed: 15657933]
42. Perskie LL, Street TO, Rose GD. Structures, basins, and energies: a deconstruction of the Protein Coil Library. *Protein Sci*. 2008; 17(7):1151–1161. [PubMed: 18434497]
43. Perskie LL, Rose GD. Physical-chemical determinants of coil conformations in globular proteins. *Protein Sci*. 2010; 19(6):1127–1136. [PubMed: 20512968]
44. Schweitzer-Stenner R. Conformational propensities and residual structures in unfolded peptides and proteins. *Molecular bioSystems*. 2012; 8(1):122–133. [PubMed: 21879108]
45. Zhang N, Ruan J, Duan G, Gao S, Zhang T. The interstrand amino acid pairs play a significant role in determining the parallel or antiparallel orientation of beta-strands. *Biochemical and biophysical research communications*. 2009; 386(3):537–543. [PubMed: 19540200]
46. Zhang N, Duan G, Gao S, Ruan J, Zhang T. Prediction of the parallel/antiparallel orientation of beta-strands using amino acid pairing preferences and support vector machines. *Journal of theoretical biology*. 2010; 263(3):360–368. [PubMed: 20035768]
47. Subramani A, Floudas CA. beta-sheet topology prediction with high precision and recall for beta and mixed alpha/beta proteins. *PloS one*. 2012; 7(3):e32461. [PubMed: 22427840]

48. Woetzel N, Karakas M, Staritzbichler R, Muller R, Weiner BE, Meiler J. BCL::Score--knowledge based energy potentials for ranking protein models represented by idealized secondary structure elements. *PloS one*. 2012; 7(11):e49242. [PubMed: 23173051]
49. Hu C, Koehl P. Helix-sheet packing in proteins. *Proteins*. 2010; 78(7):1736–1747. [PubMed: 20186972]
50. Hu C, Koehl P, Max N. PackHelix: a tool for helix-sheet packing during protein structure prediction. *Proteins*. 2011; 79(10):2828–2843. [PubMed: 21905109]
51. Chou KC, Nemethy G, Scheraga HA. Energetics of Interactions of Regular Structural Elements in Proteins. *Accounts Chem Res*. 1990; 23(5):134–141.
52. Cohen FE, Sternberg MJE, Taylor WR. Analysis and prediction of the packing of α -helices against a β -sheet in the tertiary structure of globular proteins. *Journal of molecular biology*. 1982; 156(4): 821–862. [PubMed: 7120396]
53. Fraga K, Joo H, Tsai J. An Amino Acid Code to Define a Protein's Tertiary Packing Surface. *Proteins: Structure, Function, & Bioinformatics*. 2015 Submitted.
54. Di Paola L, De Ruvo M, Paci P, Santoni D, Giuliani A. Protein Contact Networks: An Emerging Paradigm in Chemistry. *Chemical Reviews*. 2013; 113(3):1598–1613. [PubMed: 23186336]
55. Doncheva NT, Klein K, Domingues FS, Albrecht M. Analyzing and visualizing residue networks of protein structures. *Trends Biochem Sci*. 2011; 36(4):179–182. [PubMed: 21345680]
56. Holmes JB, Tsai J. Characterizing Conserved Structural Contacts by Pair-wise Relative Contacts and Relative Packing Groups. *J Mol Biol*. 2005; 354(3):706–721. [PubMed: 16269154]
57. 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]
58. 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]
59. 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]
60. Touw WG, Baakman C, Black J, te Beek TA, Krieger E, Joosten RP, Vriend G. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res*. 2015; 43(Database issue):D364–368. [PubMed: 25352545]
61. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry*. 2004; 25(13):1605–1612. [PubMed: 15264254]
62. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.
63. Harpaz Y, Gerstein M, Chothia C. Volume changes on protein folding. *Structure*. 1994; 2(7):641–649. [PubMed: 7922041]
64. Gerstein M, Tsai J, Levitt M. The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *J Mol Biol*. 1995; 249(5):955–966. [PubMed: 7540695]
65. Voronoi GF. Nouvelles applications des paramètres continus à la théorie des formes quad-ratiques. *J Reine Angew Math*. 1908; 134:198–287.
66. Bron C, Kerbosch J. Finding All Cliques of an Undirected Graph [H]. *Communications of the ACM*. 1973; 16:575–577.
67. 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]
68. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*. 2000; 28(1):254–256. [PubMed: 10592239]
69. SCOP: Structural Classification of Proteins.
70. 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]

71. Bansal M, Kumar S, Velavan R. HELANAL: a program to characterize helix geometry in proteins. *J Biomol Struct Dyn*. 2000; 17(5):811–819. [PubMed: 10798526]
72. Kumar P, Bansal M. HELANAL-Plus: a web server for analysis of helix geometry in protein structures. *J Biomol Struct Dyn*. 2012; 30(6):773–783. [PubMed: 22734588]
73. Itzkovitz S, Alon U. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome research*. 2007; 17(4):405–412. [PubMed: 17293451]
74. Fujiwara K, Toda H, Ikeguchi M. Dependence of alpha-helical and beta-sheet amino acid propensities on the overall protein fold type. *BMC structural biology*. 2012; 12(18)
75. Minor DL Jr, Kim PS. Measurement of the β -sheet-forming propensities of amino acids. *Nature*. 1994; 367(6464):660–663. [PubMed: 8107853]
76. Siepen JA, Radford SE, Westhead DR. Beta edge strands in protein structure prediction and aggregation. *Protein Sci*. 2003; 12(10):2348–2359. [PubMed: 14500893]
77. Fraga K, Joo H, Tsai J. Characterizing Tertiary Packing Surfaces with the Knob-Socket Model. *Proteins: Structure, Function, & Bioinformatics*. 2015 submitted.
78. Christendat D, Edwards A, Joachimiak A, Korolev S. Crystal structure of MT0777. Volume 2013: Midwest Center for Structural Genomics. 2002
79. Ishii R, Nureki O, Yokoyama S. Crystal structure of the tRNA processing enzyme RNase PH from *Aquifex aeolicus*. *The Journal of biological chemistry*. 2003; 278(34):32397–32404. [PubMed: 12746447]
80. Mikol V, Baumann G, Zurini MG, Hommel U. Crystal structure of the SH2 domain from the adaptor protein SHC: a model for peptide binding based on X-ray and NMR data. *Journal of molecular biology*. 1995; 254(1):86–95. [PubMed: 7473762]
81. Fushinobu S, Hidaka M, Honda Y, Wakagi T, Shoun H, Kitaoka M. Structural basis for the specificity of the reducing end xylose-releasing exo-oligoxylanase from *Bacillus halodurans* C-125. *The Journal of biological chemistry*. 2005; 280(17):17180–17186. [PubMed: 15718242]
82. Kaminski MJ, MacKenzie CR, Mooibroek MJ, Dahms TE, Hiramata T, Houghton AN, Chapman PB, Evans SV. The role of homophilic binding in anti-tumor antibody R24 recognition of molecular surfaces. Demonstration of an intermolecular beta-sheet interaction between vh domains. *The Journal of biological chemistry*. 1999; 274(9):5597–5604. [PubMed: 10026176]
83. Banfield MJ, King DJ, Mountain A, Brady RL. VL:VH domain rotations in engineered antibodies: crystal structures of the Fab fragments from two murine antitumor antibodies and their engineered human constructs. *Proteins*. 1997; 29(2):161–171. [PubMed: 9329081]

Figure 1. Knob-socket motifs

(a) In the in the four columns, knob-socket motifs for α -helix, β -sheet, coil and turn are shown. The first row depicts example of 3 residue sockets with the $XY:H$ and XYZ residue notation. As cliques, residues contact each other's side chains. Black solid lines represent the peptide bonds also connecting residues. A broken red line indicates hydrogen-bonded residues. A broken black line indicates only packing contacts. Note, the lengths of the lines are not shown to scale. The second row schematically represents these sockets. The third row shows examples of the 4 residue cliques that define the knob-socket motifs. The knob-socket notations are again overlaid. A knob **B** residue that is a long range 3° contact is added. In the bottom row, the 3D configuration of the knob-socket motifs is shown for each of the 42° structures. (b) In the fifth column, the 3D configurations are projected into 2D representations. The top two show the $XY:H$ and XYZ sockets, respectively and the bottom are the same sockets packed with knob **B**.

Figure 2. Packing clique distributions

(a) Coil packing cliques total 384,165. Of these, the 138,814 (36%) 3 and 212,867 (55%) 4 residue cliques are dominant. Less than 10% of the population comes from the 16,617 (4%) 2 and 15,633 (4%) 5 residue cliques. The singletons and 6 residue cliques are negligible. Among 3 residue cliques, the 58% majority **3** all local are free sockets. Among the 4 residue cliques, the 42% majority **3+1**, 3 local and 1 non-local, are the knob-socket packing motifs.

(b) Turn packing cliques total 40,935. The 26,164 (64%) majority are 3 residue cliques. Only 7,980 (20%) are 4 residue cliques, followed by 2 residue cliques at 6,589 (16%). The 1 and 5 residue cliques constitute less than 1% with no 6 residue cliques. Of the 3 residue

cliques, the 63% local **3** packing sockets almost double the 35% **2+1** cliques. The remaining 2% are **1+1+1** cliques. Of the 4 residue cliques, the major types of cliques are **4** (32%), **3+1** (29%), **2+1+1** (16%) and **2+2** (20%). (c). The mixed 2° structure packing cliques total 4,418,402. The 3,398,102 (77%) majority are 4 residue cliques and the 708,635 (16%) 3 residue cliques are second. These cliques classify into **3+1** (40%), **2+1+1** (34%), and **2+2** (16%).

Figure 3. Positional amino acid distributions of coil and turn sockets

Top panel shows closed (**K**), open (**O**), and strained (**S**) sockets in coil and turn. Solid lines indicate covalent bonded residues. Dashed lines indicate van der Waals contacts only. No line indicates no side-chain contact. Bottom panel shows the percent distributions of the 20 amino acid at each of the **X**, **Y**, and **Z** positions for the 3 **K**, **O**, and **S** sockets. Each amino acid's percent frequency with respect to the total number is presented in the grey scale, where darker indicates higher frequency and lighter indicates lower frequency. The highest value is 21% and lowest is 1%. In coils, the totals are 498,069 **K**, 340,324 **O**, and 114,249 **S** sockets. In turns, the totals are 107,544 **K**, 161,034 **O**, and 136,480 **S** sockets.

Figure 4. Amino acid distribution of coil and turn knobs

Heat maps present amino acid percentages of the total population for coil and turn knobs packing into the 42° structure sockets of α -helix, β -sheet, coil and turn sockets. Each knob-socket population total is given on the right side of each row. On the far right, the grey scale key is shown.

Figure 5. Amino acid composition of knobs and sockets between α -helices and β -sheets

The distribution of the 20 amino acids as a knob **B** and at each position of the **XY:H** socket in α -helices and β -sheets is plotted as a percent frequency with respect to total number in each class. The x-axis lists the amino acid residues, while y-axis labels indicate the socket type and position. For the grey scale ramp on the right, amino acids with frequencies $<0.5\%$ are colored white in the heat map. **(a)** Knob percent frequencies in helix-sheet interfaces. The labels on the y-axis indicate the sockets' 2° structure type that pack knobs labeled on the x-axis. **(b)** The 3 horizontal panels show the 20 amino acids' percent frequency of at each **X**, **Y** and **H** position in the 3 types of sockets at helix-sheet interface. The first panel from the top shows frequencies in α -helix sockets, the second panel and third panels shows frequencies for side-chain (SC) and main-chain (MC) β -sheet sockets, respectively.

Figure 6. Unique packing motifs

Examples from structures are shown (top) along with the 2D knob-socket projections (bottom). **(a)** Helix/sheet junction (1wu4⁸¹). The edge of a β -sheet packs into groove between two ± 1 ridges of an α -helix, and the α -helix packs on both sides of the β -sheet. **(b)** β -sheet edge termination (1bz7⁸²) occurs at the end strand of a β -sheet with a **3+1** clique between the β -sheet edge strand and another 2° structural element. The **3** is a consecutive **XYZ** β -strand socket at a β -sheet's edge, and the **1** stands for the non-local knob **B** residue. Prevents β -sheet growth by blocking further the carbonyl and amino group hydrogen bonding. **(c)** β -strand edge capping (1bz7⁸²) is similar to the previous except that one of the residues in the **XYZ** socket (either **X** or **Z**) comes from another type of 2° structure, which is usually coil. Therefore, this **3+1** clique could potentially inhibit β -sheet growth as well as stabilize β -sheet termini.

Figure 7. Canonical packing patterns in β -sheets and α -helix/ β -sheet packing interfaces

Examples from real structures (top), the knob **B** residue patterns (middle), and the socket patterns (bottom) of the respective packing interface are given. For clarity, the knob **B** residues are color coded, while the sockets are shown in shades of grey. (a) The β -sheet crossing angles at -30° for parallel and 150° for anti-parallel produces the pattern shown based on the structure of 1ad0⁸³. The pattern is a repeating unit of a knob packed into a three stranded diamond pocket. This creates a ridge of knobs along each β -strand that packs into a groove created between two **Y**---**H** lines. (b) The α -helix/ β -sheet cross parallel at -160° and anti-parallel at 155° to produce the canonical packing pattern derived from the 1kjn⁷⁸. Because of the 100° α -helix residue periodicity, helical knobs along they ± 4 residue ridge pack at an angle into three stranded diamond β -sheet pockets that form a groove along **Y**---**H** lines (like the β -sheet packing in (a)). The β -sheet knobs along each strand pack into successive α -helical diamond pockets. (c) The 1uds⁷⁹ structural example of canonical packing at -35° parallel and 140° anti-parallel. Again, the α -helical residues' 100° periodicity cause an α -helix's ± 4 residue ridge to pack into a β -sheet side-chain pockets and 2 stranded diamonds that form a groove bordered by the strands of the β -sheet. Conversely, the β -strand ridges of knobs pack into diamond pockets on the helix along the groove between the ± 4 residues.

Figure 8. Protein Surface Topology Map

The anatomy of protein packing of the SH2 domain 1mil⁸⁰ is illustrated by the motifs of the knob-socket model. The 2° structure elements are color-coded: red α -helices, blue β -strands, and green coil and turns. At the top, a space-filling CPK representation (left), a ribbon diagram (middle), and a topology diagram⁶ (right) are compared. At the bottom, the protein surface topology map is constructed of the motifs from the knob-socket model. Lattices indicate regular 2° structure. Free sockets are in yellow, although the free helical 2:1 sockets are not shaded for clarity. Filled sockets in grey are packed with knobs of filled circles colored according to their 2° structure type. Because the 3 stranded β -sheet and the irregular coil at N-terminus exhibit distinct packing sides, lattices are drawn for both that are separated by a green dashed line. The red dotted lines connect all the 2° structural units in sequential order.

Table 1

Three and four residue packing clique (packing clique) classification depending on the secondary structures (SS: α -helix(H), β -sheet(E), coil(C), and turn(T)) of the residues in the cliques.

packing clique	SS	mixed ^a	single ^b	Sum ^c
Three residue	H	487,789	569,679	1,057,468
	E	156,298	61,936	218,234
	C	576,866	138,814	715,680
	T	249,982	26,164	276,146
	Total ^d	708,615	796,593	1,505,208
Four residue	H	2,178,507	504,529	2,683,036
	E	2,007,809	630,273	2,638,082
	C	2,622,039	212,867	2,834,906
	T	876,721	7,980	884,701
	Total ^d	3,397,992	1,355,649	4,753,641

^a cliques involving 2 types of secondary structures

^b cliques with residues from 1 type of secondary structures

^c sum of cliques formed by mixed and single secondary structures.

^d total number of unique cliques

Table 2

Secondary Structure Classification of Knob-Socket Combinations

	Sockets			
	Helix	Sheet	Coil	Turn
H	295,866	148,431	64,383	6,804
E	98,021	68,740	39,936	3,442
C	118,689	98,527	104,386	11,394
T	35,001	20,649	18,009	4,749