Published in final edited form as:

Proteins. 2014 September ; 82(9): 2128–2140. doi:10.1002/prot.24569.

An Amino Acid Code for β -sheet Packing Structure

Hyun Joo and Jerry Tsai[†]

University of the Pacific, Department of Chemistry, Stockton, CA 95212

Abstract

To understand the relationship between protein sequence and structure, this work extends the knob-socket model in an investigation of β -sheet packing. Over a comprehensive set of β -sheet folds, the contacts between residues were used to identify packing cliques: sets of residues that all contact each other. These packing cliques were then classified based on size and contact order. From this analysis, the 2 types of 4 residue packing cliques necessary to describe β -sheet packing were characterized. Both occur between 2 adjacent hydrogen bonded β -strands. First, defining the secondary structure packing within β -sheets, the combined socket or **XY:HG** pocket consists of 4 residues $i, i+2$ on one strand and $j, j+2$ on the other. Second, characterizing the tertiary packing between β -sheets, the knob-socket **XY:H+B** consists of a 3 residue **XY:H** socket ($i, i+2$ on one strand and j on the other) packed against a knob **B** residue (residue k distant in sequence). Depending on the packing depth of the knob **B** residue, 2 types of knob-sockets are found: side-chain and main-chain sockets. The amino acid composition of the pockets and knob-sockets reveal the sequence specificity of β -sheet packing. For β -sheet formation, the **XY:HG** pocket clearly shows sequence specificity of amino acids. For tertiary packing, the **XY:H+B** side-chain and main-chain sockets exhibit distinct amino acid preferences at each position. These relationships define an amino acid code for β -sheet structure and provide an intuitive topological mapping of β -sheet packing.

Keywords

β -sheet packing; tertiary structure; secondary structure packing; packing topology; amino acid code

Introduction

Protein folding has been a challenge for over half a century.^{1,2} This problem has commonly been addressed from three perspectives: (1) the physical determinants of the amino acids that govern the three dimensional (3D) structure, (2) the mechanistic folding pathway from the unfolded polypeptide chain to the native state, and (3) prediction of the 3D structure from its primary sequence. While all three approaches are related, the prediction of a protein's 3D structure from the primary sequence has taken on a special significance with the abundance of genomic sequences. The demand for practical phenotypic interpretation of this data has prompted a substantial effort into automatic

[†]Corresponding Author: University of the Pacific, Department of Chemistry, 3601 Pacific Avenue, Stockton, CA, tel: (209) 946-2298, fax: (209) 946-2607, jtsai@pacific.edu.

methods to solve the 3D structure prediction portion of the protein folding problem.¹⁻³ To produce a true solution, the field would benefit from an improved understanding of the higher orders of protein structure. While the basis of a protein's primary and, to a degree, secondary structure are known,^{4,5} a clear characterization of the packing interactions that dominate tertiary and quaternary structures requires more development. The challenge is to discover a precise description of interactions that are dominated by non-specific van der Waals packing. From an extensive packing clique analysis,⁶ we demonstrated that simple combinations of the tetrahedral packing unit called the knob-socket motif can represent the packing within and between α -helices.⁷ To further prove the construct's ability to intuitively describe packing structure of proteins, the knob-socket analysis is extended in a comprehensive investigation of packing in β -sheets.

Previous studies investigating β -sheet structure have focused primarily on the areas of conformation, stability and topology. The topology of a β -sheet is more complicated with a higher contact order⁸ than that of an α -helix, because the component strand of a β -sheet has two sides and two possible orientations: parallel or anti-parallel. The connectivity of β -strands was one of the first analyses of β -sheet conformation.^{9,10} Further studies detailed the relative orientation and orthogonal packing of β -sheets^{11,12} as well as the flexibility among the parallel, anti-parallel, and mixed β -sheets.¹³⁻¹⁵ Distortions in β -sheet structure, such as “bulging”, “twisting”, and “bending”,¹¹⁻¹³ have been studied extensively.¹⁶⁻¹⁸ Most recently, the close packed β -sheet arrangements in barrel structures were classified into ten different combinations based the number of β -strands and the amount of shear between them.^{19,20} Further studies of β -sheets investigated the stabilizing interactions such as salt bridges and hydrogen bonds.²¹ Utilizing quantum mechanical calculations, the hydrogen bonds within the 14 atom pseudo ring formed by anti-parallel β -sheets were measured to be more stable than the hydrogen bonds within the 12 atom pseudo ring in parallel β -sheets.²² The contributions of aromatic packing interactions on β -sheet stability have also been characterized.²³ As a means to provide a more accurate method for prediction of β -sheet topology, residue pair distances have been used to determine β -strand adjacency,²⁴ register,²⁵ and parallel/anti-parallel orientation.²⁶⁻²⁸ More detail about β -sheet structure was provided by an analysis of inter-strand packing distances.²⁹ These methods as well as machine learning approaches have led to improved prediction of the general topology of strands in a β -sheet.³⁰⁻³³ The precise residue level analysis of β -sheet packing in this work meaningfully complements these previous investigations of conformation, stability, and topology of β -sheets.

By providing a construct that allows a systematic description of residue composition and packing patterns, the knob-socket analysis defines an amino acid code to β -sheet structure. The analysis produces a simpler yet more intuitive picture about how β -sheets form and how they interact with other β -sheets. Similar to the previous knob-socket study of α -helices,⁷ an initial packing clique analysis of β -sheet structures was performed to characterize the nature of knob-socket packing in β -sheets. The added complexity of β -sheet structure in comparison to α -helices required modifications to the knob-socket model of packing. These adjustments and their application to the analysis of β -sheet structure are discussed. For the β -sheet knob-socket model, the amino acid preference in knobs and sockets is explored to define a sequence specific code to β -sheet structure.

Methods

Packing Clique Calculation and Data Set

The packing clique calculation was performed similarly to previous work,^{6,7} which is summarized briefly here. All the contacts among the residues between β -sheets were calculated from Voronoi polyhedron analyses³⁴ of all heavy atoms as implemented before³⁵. In brief, this method finds planes between 2 neighboring atoms. The intersections of these planes form polyhedra built around all the heavy atoms. Those atoms sharing a polyhedral face define a contact, which in effect is a Delaunay tessellation.³⁶ A contact graph between atoms was calculated.³⁷ Based on the atom composition of amino acids, a residue contact map was constructed from the atom contacts. For glycine, contacts included interactions with glycine's backbone atoms. A clique within this graph is a set of residues that all contact each other, which defines a residue packing clique. These packing cliques were identified using the maximal clique detection method.³⁸ The packing clique analysis was performed on all 15,273 domains in the ASTRAL SCOP 1.75 set of structures filtered at 95% sequence identity.³⁹ Protein secondary structure was assigned using DSSP.⁴⁰ The UCSF-Chimera package⁴¹ displayed and output all molecular graphics.

Contact order subgrouping

From the 15,273 domains, a total of 741,834 packing cliques were identified in β -sheet structures. These packing cliques were first grouped according to the number of residues in a clique, and the process produced 5 classes from 2 to 6 residues. These initial groups were further classified into subgroups depending on the residues' relative position in the primary sequence to each other, as explained in the previous packing clique analyses.^{6,7} **ENREF_7** In short, the value represents the number of residues locally in contact with each other, the colon (:) indicates non-local contacts mediated through hydrogen bonding, and a plus (+) separates non-local contacts between residues distant in sequence. For example, a common 4 residue clique notation is **2:1+1**. While all 4 residues contact each other, 2 of the contacts are local in sequence and on the same β -strand. For β -sheets, local or covalently associated residues are within 1 or 2 residues of each other. The remaining 2 interactions are non-local, where the ":" indicates a contact from a residue on an adjacent hydrogen bonded β -strand and the "+" denotes a residue with only non-local contacts to all the other 3 residues. Essentially, this indicates packing between 2 elements of secondary structure. All the packing cliques were annotated following this scheme. Relating directly to packing cliques, the definition of the knob-socket's model for packing is based upon this level of classification between residues. In general, the **XY:H** socket is a local **2:1** packing clique, where **X** and **Y** residues are covalently close in sequence and residue **H** not only packs but is hydrogen bonded to residue **X**. The **XY:HG** pocket combines 2 sockets by adding a fourth residue **G** that is local to and on the same β -strand as residue **H**. This residue **G** packs against all other residues **X**, **Y**, and **H**. The **XY:H+B** knob-socket is the non-local **2:1+1** packing clique. Packing into the just described **XY:H** socket, the knob **B** residue contacts all 3 **X**, **Y**, and **H** residues of a socket.

Conversion of frequency into relative probability

To fairly compare frequencies on the same scale between the various socket types, the frequencies were transformed into relative probabilities, where 1 is equal to an expected observation. The 3 residue **XY:H** sockets have a total of $20^3 = 8,000$ possible amino acid compositions, so each **XY:H** socket has probability of 1 out of 8000. For each 3 residue socket, the probability of the random distribution (average probability: $P(x002DC)r$) and each socket's relative probability (P_i) over the average distribution can be calculated using following equations.

$$\tilde{Pr} = \frac{1}{8000} \times \text{Total number of Sockets} \quad (1)$$

$$P_i = \frac{v_i}{\tilde{Pr}} \quad (2)$$

In equation (2), v_i is the measured frequency of a socket i .

Results and Discussion

Analysis of β -sheet Packing Cliques

To investigate the complexity of β -sheet tertiary structure, packing cliques were calculated from a comprehensive set of interacting β -sheets in the same manner as done previously.⁶ This procedure produced a total of 741,834 packing cliques. These β -sheet packing cliques are first classified based on the number of the residues in the clique. As shown in Figure 1, the clique sizes range from 2 to 6 residues, and these will be described from the least to the most prevalent. Although the 2 and 6 body cliques exhibit only nominal populations that together represent less than 0.1% of the total packing cliques, the 3 and 5 body cliques produce marginally larger populations at 8% and 7%, respectively, of the total packing cliques. At 85%, the 4 residue packing cliques clearly dominate the distribution. This single peak could be thought of as an artifact of the contact calculation.³⁵ However, as described in the methods, the residue contact graph does not enforce 4 body contacts since interactions are defined between atoms that share a Voronoi polyhedral face.³⁴ Also, as discussed below and seen in our previous analyses^{6,7} _ENREF_7, the packing cliques are not restricted to a tetrahedral arrangement and are often planar. Moreover, comparison with the previous packing cliques analysis involving α -helices highlights the authenticity of these results. While the distributions of β -sheet and α -helix packing cliques are similar in the lack of small and larger sizes, the α -helical packing shows the highest population for 3 residue cliques and then 4 residue cliques.⁷ Similarly, the β -sheet packing population's high preference for the 4 residue clique is more of an indication of an underlying regularity in β -sheet packing that is being revealed by this analysis, as explained in detail below.

Continuing with our previously established protocol,⁶ the residues within each packing clique are further classified based on their contact order with one another. This nomenclature is explained briefly. First, by sharing at least one polyhedral face, all residues in the packing clique are in van der Waals contact with each other. Residues belonging to

the same β -strand and therefore close in sequence are summed together. Those belonging to the same β -sheet but not the same β -strand are separated by “:” and are usually in close contact via a hydrogen bond and van der Waals interaction. The remaining residue is non-local in sequence and is denoted with a plus sign “+”. For example, the **2:1+1** packing clique consists of 4 residues where 2 local residues from one β -strand contact 1 residue, which belongs to the same β -sheet on an adjacent strand (indicated by the “:”), that all pack with 1 non-local residue from a different β -sheet (indicated by the “+”). For this contact order classification of packing cliques, the significant contributors are overlaid on the histogram in Figure 1, while the complete breakdown is detailed in Table S1. As expected, more residues produce greater potential for diversity in the types of contact order classes; however, many are not found or consist of insignificant members. In general, the trend across all of the packing cliques is the dominant population of local packing cliques over non-local ones. This result is consistent with what is visually seen on contact maps, where there are far more local interactions than non-local ones. Of the 6 significant classes, 5 are local. The local packing cliques primarily consist of residues all from the same β -sheet. A packing clique spanning 2 hydrogen-bonded β -strands of a β -sheet includes the 3 residue **2:1**, the 4 residue **3:1** and **2:2**, and the 5 residue **3:2**, whereas the **2:1:1** is a 4 residue packing clique spanning 3 hydrogen-bonded β -strands of the same β -sheet. The only significant non-local packing clique is the **2:1+1**, which is described above.

Knob-Socket Model of β -sheet Packing

The knob-socket model provides a practical and intuitive representation of protein packing by abstracting the complicated side-chain interactions into patterns of basic units. From our previous analysis of α -helical packing,⁷ there are 2 types packing units: free sockets and filled sockets. Free sockets define the packing of secondary structure not directly involved in tertiary packing, while filled sockets define the packing of tertiary structure interacting with non-local knob residues. The purpose of this packing clique analysis is to identify β -sheet classes that correspond with the free and filled sockets. While the β -sheet packing clique distribution indicates regularity, a proper model should account for the features unique to β -sheets. Figure 2a depicts a 3 stranded β -sheet as balls and sticks, and Figure 2b shows the same β -sheet in a simplified lattice representation. Unlike the α -helical lattice that can be considered a single continuous surface, the β -sheet lattice exhibits 2 distinct sides. Another feature unique to β -sheet packing is the need to address the prevalent packing of knob residues with the main-chain. One distinct simplification of this β -sheet lattice is the implied rather than the explicit hydrogen bonding between β -strands, which is implemented to enforce lattice regularity on the dissimilar parallel and anti-parallel hydrogen bonding patterns. Based on this lattice in Figure 2b, the free and filled sockets in β -sheet packing are more clearly identified.

In contrast to the 3 residue **2:1** clique found in α -helical packing, the 4 residue **2:2** packing clique best captures the secondary structure representing the free sockets in β -sheet packing. While the local 3 residue **2:1** packing clique is naturally most similar to the α -helical free socket, they are not populated enough in β -sheets at 5% (Table S1) to meaningfully represent the free sockets that describe secondary structure packing of β -sheets. The next candidate is the 4 residue **3:1** packing clique. This class includes several arrangements with

one or more main-chain interactions that are also not suitable candidates to be a socket. As depicted in Figure 2b, a **3:1** packing clique made up of consecutive residues 0, 1, and 2 on the *i* β -strand interacting with residue 5 on the *j* β -strand would not provide enough side-chain specificity to describe β -sheet packing. By forming side-chain specific interactions, the **3:1** packing clique shown in Figure 2c produces an informative socket and consists of 3 consecutive residues on one β -strand designated **X**, **m**, and **Y** packed against residue **H** on an adjacent, hydrogen bonded β -strand. Based on the orientation of residues in β -sheet structure, the **X**, **Y** and **H** side chains face the same side of the β -sheet while the **m** only contributes main-chain atoms by facing the opposite side of the sheet. Due to the lack of side-chain specificity provided by the **m** residue, the designation of this β -sheet socket will be abbreviated to **XY:H** so that the nomenclature is consistent across secondary structure types. While this 3 residue representation describes free sockets in α -helices, the inconsistency of curvature across a β -sheet requires one more adjustment. An α -helix produces a relatively constant surface curvature so that the *i* and *i*+3 residues are almost always contacting, which produces an α -helical lattice with a uniform triangular packing pattern. In contrast, β -sheet surface exhibits variability in curvature such that the **XY:H** socket can occur in 2 mutually exclusive orientations. To account for both free socket states, an open hydrogen bonding box is the basis of the β -sheet lattice (Figure 2b). For example, the hydrogen bonded box formed by residues 0 and 2 on β -strand *i* and residues 4 and 6 on β -strand *j* in the upper left of Figure 2b can form either **XY:H** sockets of 0,2:6 or 2,0:4, but not both. This ambiguity in socket arrangement (highlighted on the left of Figure 2g) determines that the best representation for the free β -sheet socket includes both **XY:H** orientations or the entire 4 residue hydrogen bonded box identified by the **2:2** packing clique. As shown by Figure 2d, the free socket configuration of the 4 residues *i*, *i*+2, *j*, and *j*+2 point to the same side of the β -sheet and can be thought of as a combination of sockets, which for simplicity is abbreviated to a **XY:HG** pocket. Although potential main-chain interactions with the residues facing towards the other side of the sheet are possible, these residues are not included in the pocket designation for clarity, since they do not contribute specificity to packing.

For filled sockets, the model requires a non-local packing clique **XY:H+B**, where a knob **B** residue packs into an **XY:H** socket. As pointed out above, the only significant non-local packing clique is the 4 residue **2:1+1**. While this type of packing clique matches α -helices, the β -sheet filled sockets include a slight complication with the backbone. Analysis of the β -sheet **2:1+1** packing cliques identify 2 types of knob-sockets (side-chain and main-chain sockets) that are different only in how the knob **B** packs with the residue at position **Y**. A side-chain knob-socket packs only the side-chains of the **XY:H** socket with the knob **B** residue (Figure 2e), where residues **X** and **Y** are ± 2 residues apart. In this example, the knob **B** residue positions farther back from the β -sheet of the **XY:H** socket. Schematically, the 2 mutually exclusive side-chain socket orientations are shown in the hydrogen bonding box as part of the β -sheet lattice in Figure 2g, which Figure 2h depicts a schematic of the **XY:H+B** side-chain socket. As the second type of **2:1+1** knob-socket, Figure 2f depicts a main-chain knob-socket, where the knob **B** residue packs deeper into the hydrogen bonding box. In this case, the knob **B** residue interacts not only with the **X** and **H** side-chains but also against the backbone of residue **m** facing the opposite side of the β -sheet. For consistency, the residue

m in the main-chain socket is considered the **Y** residue. As a result, residues **X** and **Y** are next to each other (± 1 residue apart) and the side chains face opposite sides of the β -sheet. A variety of packing arrangements for main-chain sockets are shown on the left of Figure 2g and an individual main-chain socket is schematically shown by Figure 2i.

Similar to the analysis of α -helical structure,⁷ the remaining β -sheet packing cliques are derivative of the free **XY:HG** pocket and 2 types of filled sockets because of the redundant overlap of packing cliques. For example, the residue interactions of a **2:1:1** packing clique are simply a part of 2 free **XY:HG** pockets. In this way, the knob-socket model systematically reduces the complexity of non-specific packing interactions into regular patterns of 3 types of sockets that represents a packing topology. Furthermore, the amino acid composition of these constructs provides a code for protein structure. The free **XY:HG** pocket informs on sequences that prefer non-packed β -sheet structure, and the **XY:H+B** knob-socket defines β -sheet sequences that form tertiary structure.

Amino Acid Code for β -sheet Secondary Structure Packing

The most useful aspect of the knob-socket model is that the basic units of packing relate amino acid composition to structural preferences. For α -helices, a direct comparison of free and filled sockets was performed since both sockets were the same **2:1** type. As developed above, the free socket in β -sheets is a 4 residue **XY:HG** combined socket or pocket. For comparison, filled **XY:HG** pockets were artificially composed by combining the filled **XY:H** sockets within a hydrogen bonded box as shown in Figure 2g for both of the filled side-chain and main-chain sockets. The results are shown in Figure 3 for 50 pairs of **XY** plotted against 50 **HG** pairs. Essentially, the residue composition of these **XY:HG** pockets reveal the amino acids preferences found to form β -sheet structure. The top compares the most prevalent free **XY:HG** pockets with the corresponding filled pockets, and the bottom compares the most prevalent filled **XY:HG** pockets with the corresponding free pockets. In general, the amino acid compositions clearly show preferences to be free or filled: pockets that like to be filled do not like to be free, and vice versa. As a corollary, the white indicates that many sequences are not found to form β -sheet structure. These sequences help to define the negative sequence space of unfavorable combinations. To better interpret the amino acid preferences, the residue pairs are further divided into three groups according to the basic chemistry of the **XY** or **HG** pair.

The free **XY:HG** pocket (Figure 3, top 2 plots) prefers the polar amino acid pairs such as TT, SS, TS, ST, and KT and mixed polar/nonpolar pairs such as WQ, WK, WR, and VS. These pairs generally form pockets with many other amino acid residues, where residues in the **XY** position exhibit preferences to form free pockets with certain **HG** pairs. For example, these pairs are involved in the 5 most frequently observed free pockets of WQ:IK, SS:TD, WK:GE, TT:TT and TS:TT, and conversely are seldom observed as filled pockets. As another example, QY at the **XY** position form good free pockets with **HG** position residue pairs TT, TS KS, and RS. Also, the pockets show orientation specificity. The pocket VS:AS is found only as a free pocket, yet switching the amino acids in the **HG** position produces a VS:SA pocket that is not found as either type. The nonpolar pairs at the bottom of the 2 top plots are shown due to their general prevalence in free **XY:HG** pockets, the

highest being VY:RQ and IV:DR. One possible reason for the existence of these nonpolar pairs in free pockets is that the set of structures were analyzed as monomers, when many are involved in oligomeric quaternary interactions. Potentially, the exposure of the quaternary filled pockets contributes to these being miscounted as free pockets. As support for this possibility, these free pocket compositions also form highly populated filled sockets (Figure 3, bottom plots).

In the filled **XY:HG** pocket comparison shown by the bottom 2 plots of Figure 3, the striking feature is the lack of polar pairs and the prevalence of nonpolar pairs. The most frequent residues observed at **XY** and **HG** positions in the filled pockets are nonpolar residues such as VV, VL, LV, VI, LL, IV, LI, and IL. These nonpolar residue pairs are common in β -sheet structure and form pockets with a variety of amino acid pairs. On the other hand, polar/nonpolar residue pairs include FG, ML, QC, IR, LQ, WQ, CY, and WR. These **XY** pairs display a specific preference for certain **HG** partners to form filled pockets as exhibited by the well populated filled pockets FG:IL, LQ:AC, and LW:QC. Also, the CY residues at the **XY** positions produce filled pockets when specifically paired with WR, WQ, and WK at the **HG** position.

Figure 3 clearly shows the **XY:HG** pockets' ability to reveal an amino acid code for β -sheet secondary structure. Of the 160,000 possible combinations, the composition of the free and filled pockets defines those sequences that favor formation of β -sheet structure.

Furthermore, the sequence differences between the free and filled pockets allows differentiation of those sequences that will also pack at higher orders of protein structure. The finer details of higher order β -sheet packing will be discussed in the following section.

Amino Acid code for β -sheet Tertiary Packing

The knob-socket construct defines tertiary packing in β -sheets as shown by the 2 types of 4 residue **2:1+1** packing cliques in Figure 2g. The difference between them originates from how deeply the knob **B** residue packs to form either a side-chain socket (Figure 2h) or a main-chain socket (Figure 2i). Because they involve non-specific interactions with the β -sheet backbone, main-chain sockets are expected to exhibit less specificity and in many instances can be accounted for by a side-chain socket or filled pocket. However, the main-chain sockets are found independently and provide the fine detail of tertiary packing in β -sheets. For this reason, the main-chain sockets are analyzed along with the side-chain sockets. The total number of **2:1+1** packing cliques split into 31,608 side-chain sockets and 37,078 main-chain sockets. To discover general trends, the positions are analyzed for the distribution over the 20 amino acids, and then amino acid composition of the **2:1+1** packing cliques are investigated.

Figure 4 shows the normalized amino acid percent frequency at each position of both types of knob-socket motifs, and Table S2 presents the numerical data. The only consistency across all knobs and sockets is that Pro is rarely observed. Overall, the class of amino acids contributing to the **XY:H** sockets in β -sheet packing interfaces is mostly hydrophobic aliphatic or aromatic residues, as expected (Figure 4a). For side-chain sockets (Figure 4a, top), the distributions at **X**, **Y**, and **H** show a preference for the longer chain aliphatic Val,

Leu, Ile residues over the aromatic Phe and Tyr. Of the other amino acid residues, only Thr, Met, and Cys exhibit stronger than average populations, where Thr favors the **X** position and Cys disfavors the **Y** position. For the main-chain sockets (Figure 4a, bottom), the general frequency pattern of the amino acids is similar to side-chain sockets favoring aliphatics over the aromatics though more subdued. Also, the residue distribution reveals a higher prevalence of hydrophilic and charged residues. Since the **Y** position in main-chain sockets only interacts through the backbone with non-specific backbone interactions, the side-chain chemistry would not play a role in the packing. Instead, if this β -strand were amphipathic, the **Y** residue's side chain would point away from the hydrophobic core and towards the hydrophilic side of an amphipathic. The amphipathicity would explain the increase in polar/charge groups found in main-chain sockets. The major differences between main-chain sockets and side-chain sockets are the existence of the Ser and Gly residues across the **X** and **Y** positions, the lack of Met at **X** or **Y** positions, and the existence of Arg only at the **Y** position. In Figure 4b, the amino acid frequencies of knob **B** residues packed into either side-chain or main-chain sockets are depicted. Knob **B** residues packing into side-chain sockets exhibit the same trend found in the sockets for preferring long chain aliphatics over aromatics (Figure 4b, top). Surprisingly, neither Trp nor Ala is prevalent as knobs that pack into side-chain sockets. The knobs that pack into main-chain sockets display a similar pattern favoring the aromatics and aliphatics (Figure 3b, bottom) except for the striking prevalence of the bulky Trp residue. While Leu commonly packs as a knob **B** in secondary structure, the Trp knob packing deeply into main-chain sockets is a distinct signature of β -sheet packing with more sequence specificity than side-chain sockets (discussed below). When a Trp packs into a β -sheet (Figure 2f), the Trp knob forms interactions with at least 4 main-chain sockets (Figure 2g, right side). Although Trp may not be prevalent in β -sheets, the Trp as a knob **B** residue contributes significant interactions to packing.

While discussing the individual positions provides basic insight into the preferences of amino acids in β -sheet packing, the knob-socket motif provides a precise method to characterize the amino acid code that produces β -sheet tertiary structure. Figure 5 depicts the composition of the top 100 most prevalent **XY:H** socket motifs packing with 20 knob **B** residues for side-chain sockets and main-chain sockets. As expected, the **XY:H+B** packing cliques reflects the same trends shown at the position level in Figure 4, but this analysis reveals the specific amino acid combinations that favor β -sheet tertiary structure. As with the characterization of pockets in Figure 3, many 4 residue combinations show little to no counts even though they are comprised of the favored amino acids detailed in Figure 4. This feature revealed by the knob-socket model helps to define the negative sequence space of β -sheet tertiary structure.

For side-chain sockets (Figure 5a), only certain combinations of the large aliphatics are favored. The top 10 most populated side-chain sockets VV:V, VL:V, LI:I, LI:F, LV:L, VI:I, VI:V, VV:I, TV:Y, and LV:V exhibit preferences across the knob **B** residues. Of these, the LI:F is unique due to the Phe residue packed at the **H** position. Also, Tyr at the **H** position exhibits selectivity toward TL and TV as **XY** residues. The 2 most populated side-chain sockets VV:V and VL:V favor Tyr, Phe, Val, Ile, and Leu as knob **B** residues. However, specificity of side-chain sockets is clearly evident in Figure 4. The third and fourth most

populated side-chain sockets LI:F and LI:I strongly prefer to pack with a Tyr knob. The side-chain sockets TV:Y, TL:Y, and TL:L favor the long chain aliphatics of Ile and Leu over the aromatic knobs. As another specific interaction, the side-chain sockets with Cys at the **H** position LW:C, VW:C, IW:C, and LQ:C can be thought of as disulfide packing sockets, since they all pack with a disulfide bonded Cys knob **B** residue. In addition, the Tyr knob **B** residue prefers packing with side-chain sockets LI:I, LI:F, LL:L and LM:L over VV:V or VL:V.

As displayed in Figure 5b, the preferences of amino acids in the main-chain sockets result in a distinctly different socket distribution in comparison to side-chain sockets. In general, main-chain sockets show stronger preferences for certain knob **B** residues. The most common main-chain sockets YY:W, YC:W, FT:S, LI:W, FS:L, SG:L, YL:L, and SC:Q are observed 50 times more than average. The main-chain sockets YY:W and YC:W prefer to pack with Gln most and Glu second, but disfavor the aliphatics. The main-chain socket FT:S favors packing with a Trp knob **B** residue, whereas LI:W favors packing with mostly with aliphatic Leu knob **B** residue. Across the main-chain sockets, this increased specificity of knob **B** residues naturally results from increased interaction in main-chain sockets. Because the knob **B** residue packs all the way to the backbone of the β -sheet, shape complementarity between the knob and the socket plays a larger role in these main-chain sockets than with side-chain sockets, where the knob sits shallowly on top of the socket with less interaction specificity.

Comparison of the socket composition between β -sheets and α -helices further supports that the knob-socket model characterizes an amino acid code for protein structure. The differences occur in sequence composition and in structural arrangement. In our previous analysis of α -helix packing,⁷ the filled **XY:H** sockets favor LL:L, LA:L, LL:A, AL:A, and LL:A. The amino acids are distinctly different than either of the filled **XY:H** side-chain or main-chain sockets in β -sheets. In addition, the knob **B** residues that pack into α -helices are similar except for the notable 2 differences. The knob **B** residues in β -sheets disfavor Ala, while the knob **B** residues in α -helices disfavor Trp, which packs well into β -sheet main-chain sockets. Not only are the **XY:H** sockets different in composition, but also in sequence/structure. The β -sheet **XY:H** side-chain socket involves residues $i, i+2$ and j from an adjacent strand, and the main-chain socket involves residues $i, i+1$ and j from an adjacent strand. By contrast, the α -helical **XY:H** socket involves residues $i, i+1$ and $i+4$ or $i, i-1$ and $i-4$. Therefore, even though certain sockets share the same composition, their relative separation in the protein sequence is very different.

Mapping Patterns of β -sheet Packing

The knob-socket model provides a simple and informative representation that identifies the interactions within and between β -sheet structure. By projecting filled sockets and free pockets on a regular lattice, the tertiary packing of a β -sheet structure can be clearly presented and more intuitively understood. Essentially, the knob-socket model provides a two-dimensional topography of packing interactions between secondary structure units. As an example, Figure 6 compares the ribbon diagram with the knob-socket pattern for antitumor antibody 1ad0.⁴² The ribbon diagram (Figure 6a) provides a clear overview of the

classic immunoglobulin fold that contains two β -sheets packing against each other. Because any additional representation of side-chain packing overly complicates the illustration, this representation cannot provide any direct information about tertiary structure. To show the internal tertiary packing, the immunoglobulin fold from Figure 6a is opened up to reveal the 2 β -sheets (Figures 6b and 6c), where the internal packing side of the β -sheets points out of the page. The ribbon diagram is preserved and only relevant side-chains are shown for clarity. While more structural characteristics are shown, the tertiary interactions between the β -sheets produce too much detail in ribbon diagrams. Using the knob-socket model, β -sheet packing can be depicted more clearly on a regular lattice (Figures 6d and 6e). The diagram is a topological map of the internal packing surface. The packing within and between the β -sheets is clearly illustrated with knob **B** residues packing into single as well as combined sockets. M34, C98, L48 and V117 in Figure 6e are examples of knob residues packing into single side-chain sockets. Figure 6d illustrates more complex packing. The knob L4 combines packing with 5 main-chain sockets in an area bounded by residues M34, C98, R100, and Y110 and into the backbone atoms of T99 and W111. In a similar fashion on this same β -sheet, I72 packs into a larger set of 8 main-chain sockets in the area defined by side-chain residues M34, W36, L48, I51, T60, and Y62 and into the backbone atoms of N35, G49, T59, and E61. As noted above, Trp is an amino acid that usually packs into multiple main-chain sockets. In Figure 6e, W36 packs into seven main-chain sockets defined by the side chains of residues S21, E6, L20, C22, I72, L81, and L83 and into the backbone atoms of S21 and Y82. Consistent with our analysis, the prevalence of polar/charge residues in the filled main-chain sockets derives from the non-specific packing into the residues' backbone atoms. Interestingly, Y96 as a knob **B** residue packs into sockets on both β -sheets. Because β -sheets naturally exhibit irregular curvature, residues with long side chains can sometimes contact other residues on the neighboring β -strands within the same β -sheet.

Conclusion

Through a careful analysis of packing cliques, the knob-socket model has been extended to characterize the tertiary structure in β -sheet packing. Comparing β -sheet and α -helix packing, the general themes of the knob-socket model are consistent, although the intrinsic differences between β -sheets and α -helices require adjustment of the free and filled sockets. In particular, the knob-socket model needed to address the structural attributes that a β -sheet has two sides and irregular curvature compared to the consistency of an α -helix's single cylindrical surface. For free sockets, the more general 4 residue **2:2** packing clique was required to account for the variability in β -sheet curvature and defines the **XY:HG** pockets. For filled sockets, the non-local 4 residue **XY:H+B** knob-sockets consisted of 2 types based on the packing depth of the knob **B** residue into **XY:H** side-chain or main-chain sockets. The side-chain socket interacts with the knob **B** residue using only side chains pointing toward the same side of the β -sheet, and the main-chain socket packs the backbone of the residues with the knob **B** residue. This knob-socket model provides an intuitive yet comprehensive description of the residue level packing between β -sheets.

Most importantly, the identification of the appropriate β -sheet packing cliques to represent the knob-socket model provides an insightful tool to investigate and characterize β -sheet tertiary structure, as exemplified by the deconvolution of β -sheet packing structure in Figure

6. In particular, calculating the amino acid composition of the **XY:HG** pocket and **XY:H+B** knob-socket motifs directly relates primary sequence to β -sheet secondary and tertiary structure, respectively. The composition of these constructs provides an amino acid code that defines β -sheet structure. While individual mutational studies have identified certain contributions or residues such as aromatic interactions,²³ the knob-socket relationship provides the next step in broadly understanding how the amino acid primary sequence determines a protein's fold and function. As expected, Val, Ile, and Leu are most commonly observed amino acids in side-chain packing sockets (Figure 3), and yet these residues' arrangement in the **XY:H** side-chain socket is also important for β -sheet structure, as the most frequently observed side-chain sockets are VV:V, VV:I, VV:L, LI:F, LI:I, VL:V, VL:L, LL:V, LL:L, and VI:V (Figure 5). Amino acid composition in main-chain sockets is significantly different. Main-chain sockets favor the residues Tyr, Trp, Phe, Ser, Leu, Ile, and Cys (Figure 4) in the following **XY:H** socket arrangements: YY:W, YC:W, FT:S, LI:W, FS:L, SG:L, YL:L, and SC:Q (Figure 5). The frequency distribution of the knob **B** residue in the knob-socket **XY:H+B** motif identifies residues that prefer to form β -sheet tertiary structure. Of course, the most frequently observed knob **B** residues packing into side-chain sockets are aromatic and hydrophobic amino acids in the following ascending order of frequency: Tyr, Phe, Val, Ile, and Leu. The knob **B** residues packing into main chain sockets include other classes of amino acids with Trp, Ile, Leu, Glu, and Gln. To describe the β -sheet packing that favors only secondary structure, a larger 4 residue pocket motif **XY:HG** is introduced. The amino acid pocket composition is compared between filled and free pockets. This relationship between sequence and structure specifies whether amino acids in a certain pocket combination will form β -sheet structure as well as which residues in that β -sheet will form tertiary structure.

This information contributes to the design and analysis of β -sheet secondary and tertiary packing structure as well as provides a new approach for protein structure prediction. For design, the **XY:HG** pockets can be used to build sequences with high preference to form β -sheet secondary structure and the **XY:H+B** knob-sockets inform how to properly design the tertiary packing of the β -sheet's hydrophobic core. For protein structure prediction, sequences of unknown structure can be searched for patterns of both the **XY:HG** pockets and **XY:H+B** knob-sockets to construct more accurate models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH Grant R01GM104972. We would also like to thank Helen Tsai for the careful reading and editing of this manuscript.

References

1. Dill KA, MacCallum JL. The protein-folding problem, 50 years on. *Science*. 2012; 338(6110): 1042–1046. [PubMed: 23180855]
2. Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. *Annual review of biophysics*. 2008; 37:289–316.

3. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA. The protein folding problem: when will it be solved? Current opinion in structural biology. 2007; 17(3):342–346. [PubMed: 17572080]
4. Levitt M, Chothia C. Structural patterns in globular proteins. Nature. 1976; 261(5561):552–558. [PubMed: 934293]
5. Richardson JS. The anatomy and taxonomy of protein structure. Advances in protein chemistry. 1981; 34:167–339. [PubMed: 7020376]
6. 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]
7. 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]
8. 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]
9. Sternberg MJE, Thornton JM. On the conformation of proteins: The handedness of the connection between parallel β -strands. Journal of molecular biology. 1977; 110(2):269–283. [PubMed: 845952]
10. Sternberg MJE, Thornton JM. On the conformation of proteins: an analysis of β -pleated sheets. Journal of molecular biology. 1977; 110(2):285–296. [PubMed: 845953]
11. Chothia C, Janin J. Relative orientation of close-packed beta-pleated sheets in proteins. Proceedings of the National Academy of Sciences of the United States of America. 1981; 78(7): 4146–4150. [PubMed: 16593054]
12. Chothia C, Janin J. Orthogonal packing of beta-pleated sheets in proteins. Biochemistry. 1982; 21(17):3955–3965. [PubMed: 6751382]
13. Salemme FR. Structural properties of protein β -sheets. Progress in Biophysics and Molecular Biology. 1983; 42(C):95–133. [PubMed: 6359272]
14. Salemme FR, Weatherford DW. Conformational and geometrical properties of β -sheets in proteins. II. Antiparallel and mixed β -sheets. Journal of molecular biology. 1981; 146(1):119–141. [PubMed: 7265226]
15. Salemme FR, Weatherford DW. Conformational and geometrical properties of β -sheets in proteins. I. Parallel β -sheets. Journal of molecular biology. 1981; 146(1):101–117. [PubMed: 7265225]
16. Richardson JS, Getzoff ED, Richardson DC. The beta bulge: a common small unit of nonrepetitive protein structure. Proceedings of the National Academy of Sciences of the United States of America. 1978; 75(6):2574–2578. [PubMed: 275827]
17. Chan AW, Hutchinson EG, Harris D, Thornton JM. Identification, classification, and analysis of beta-bulges in proteins. Protein science : a publication of the Protein Society. 1993; 2(10):1574–1590. [PubMed: 8251933]
18. Chothia C, Murzin AG. New folds for all-beta proteins. Structure. 1993; 1(4):217–222. [PubMed: 8081735]
19. Murzin AG, Lesk AM, Chothia C. Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis. Journal of molecular biology. 1994; 236(5):1369–1381. [PubMed: 8126726]
20. Murzin AG, Lesk AM, Chothia C. Principles determining the structure of beta-sheet barrels in proteins. II. The observed structure. Journal of molecular biology. 1994; 236(5):1382–1400. [PubMed: 8126727]
21. Daffner C, Chelvanayagam G, Argos P. Structural characteristics and stabilizing principles of bent beta-strands in protein tertiary architectures. Protein science : a publication of the Protein Society. 1994; 3(6):876–882. [PubMed: 8069218]
22. Perczel A, Gaspari Z, Csizmadia IG. Structure and stability of beta-pleated sheets. Journal of computational chemistry. 2005; 26(11):1155–1168. [PubMed: 15952205]
23. Budyak IL, Zhuravleva A, Gierasch LM. The Role of Aromatic-Aromatic Interactions in Strand-Strand Stabilization of beta-Sheets. Journal of molecular biology. 2013; 425(18):3522–3535. [PubMed: 23810905]
24. Kikuchi T, Nemethy G, Scheraga HA. Prediction of the packing arrangement of strands in beta-sheets of globular proteins. Journal of protein chemistry. 1988; 7(4):473–490. [PubMed: 3255373]

25. Steward RE, Thornton JM. Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory. *Proteins*. 2002; 48(2):178–191. [PubMed: 12112687]
26. 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]
27. 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]
28. 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]
29. Nagarajaram HA, Reddy BV, Blundell TL. Analysis and prediction of inter-strand packing distances between beta-sheets of globular proteins. *Protein engineering*. 1999; 12(12):1055–1062. [PubMed: 10611399]
30. Brown WM, Martin S, Chabarek JP, Strauss C, Faulon JL. Prediction of beta-strand packing interactions using the signature product. *Journal of molecular modeling*. 2006; 12(3):355–361. [PubMed: 16365772]
31. Jeong J, Berman P, Przytycka TM. Improving strand pairing prediction through exploring folding cooperativity. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*. 2008; 5(4):484–491.
32. Max N, Hu C, Kreylos O, Crivelli S. BuildBeta--a system for automatically constructing beta sheets. *Proteins*. 2010; 78(3):559–574. [PubMed: 19768785]
33. Savojardo C, Fariselli P, Martelli PL, Casadio R. BCov: a method for predicting beta-sheet topology using sparse inverse covariance estimation and integer programming. *Bioinformatics*. 2013; 29(24):3151–3157. [PubMed: 24064422]
34. Voronoi GF. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *J Reine Angew Math*. 1908; 134:198–287.
35. Harpaz Y, Gerstein M, Chothia C. Volume changes on protein folding. *Structure*. 1994; 2(7):641–649. [PubMed: 7922041]
36. Delauney B. Sur la sphère vide. *Bull Acad Sci USSR (VII) Classe Sci Mat Nat*. 1934:783–800.
37. Gerstein M, Tsai J, Levitt M. The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *Journal of molecular biology*. 1995; 249(5):955–966. [PubMed: 7540695]
38. Bron C, Kerbosch J. Finding All Cliques of an Undirected Graph [H]. *Communications of the ACM*. 1973; 16:575–577.
39. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. *Nucleic acids research*. 2004; 32(Database issue):D189–192. [PubMed: 14681391]
40. 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]
41. 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]
42. 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]
43. Brett TJ, Traub LM, Fremont DH. Accessory protein recruitment motifs in clathrin-mediated endocytosis. *Structure*. 2002; 10(6):797–809. [PubMed: 12057195]
44. Monnet C, Bettsworth F, Stura EA, Le Du MH, Menez R, Derrien L, Zinn-Justin S, Gilquin B, Sibai G, Battail-Poirot N, Jolivet M, Menez A, Arnaud M, Ducancel F, Charbonnier JB. Highly specific anti-estradiol antibodies: structural characterisation and binding diversity. *Journal of molecular biology*. 2002; 315(4):699–712. [PubMed: 11812141]
45. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.

Figure 1. Distribution of Packing Cliques in β -sheet packing

A histogram divides the 741,834 β -sheet packing cliques into the five classes based on the number of residues involved in the clique. Values on top of each column indicate the total number of members. The sizes of packing cliques are further sub-divided based on contact order or the number of secondary structural elements contributing to the packing cliques. To produce the best clarity in the figure, only the classes with greater than a 5% population of the total packing cliques are considered significant. The remaining classes are grouped into local (coming from the same β -sheet) and non-local (involve interactions between at least 2 β -sheets). Only the 2 and 6 residue packing cliques are too small with too many diverse classes to show any more detail. Complete data for counts and contact order classes are given in Table S1.

Figure 2. The knob-socket model in β -sheet packing

These β -sheet representations are created using Chimera.⁴¹ (a) Ball and stick representation of the β -sheet from adaptor protein 1kyf.⁴³ (b) The β -sheet lattice of anti-parallel(i,j) and parallel(j,k) orientations. Black solid lines indicate contacts through covalent bonds, and hashed lines indicate van der Waals contacts. Numbers indicate the relative residue positions from the starting residues i , j , and k . The residue numbers in the large spheres represent residues with side chains facing out of the page and the plain residue numbers represent residues with side chains facing into the page. The hydrogen bonding between β -strands is implied with the broken lines and do not follow the structural conventions of anti-parallel or parallel hydrogen bonding patterns for the sake of simplicity. (c) The two-dimensional representation of a local β -sheet 3:1 packing clique or $XmY:H$ socket showing the 4 residues' relative positions. X and Y are the residues belonging to the same strand and connected through peptide bonds, while residue H is the third residue forming a socket aligned with residue X . While X , Y and H face the same side, m faces the opposite side of the β -sheet and provides only interactions with backbone atoms. The two possible orientations occur with the residue number of X higher or lower than Y . In this work, these orientations are treated as the same class. For consistency, this β -sheet 3:1 packing clique is referred to as an $XY:H$ socket. (d) The $XY:HG$ pocket is a local 2:2 packing clique. By combining 2 $XY:H$ sockets, this combined socket or pocket for simplicity describes a hydrogen bonded box between 4 residues. As shown, all 4 residues point out of the page and the 2 residues designated with an m face into the page. To simplify the notation, the m residues are implied in the nomenclature. (e) An example of a knob packing into side-chain sockets from adaptor protein 1kyf.⁴³ The knob B residue Ile packs against two sockets VV:I and LI:V. (f) An example of a knob residue packing into main-chain sockets from anti-estradiol antibody 1jnh.⁴⁴ A Trp knob B residue is packing into 4 main chain sockets, YV:S,

SC:Y, LY:L, and SL:Y. **(g)** Knob-socket packing cliques on the β -sheet lattice. Knobs **B** residues packing into **XY:H** side-chain sockets are shown on the left. On the anti-parallel sheet between β -strands i and j , the two sockets $i, i+2:j+4$, and $j+6, j+4:i$ share a knob **B** residue. Underneath on the parallel sheet between β -strands j and k , a knob **B** residue packs into a single socket $j+6, j+4:k+6$. Example of a **XY:H** main-chain socket packing with a knob **B** residue is shown on the right side. Sockets $i+5, i+4:j+1$, and $j+2, j+1:i+4$ share one knob **B** residue. The residue numbers in the small grey spheres such as $i+5$, $j+1$, and $k+1$ are the residues in the sockets facing the opposite side of the β -sheet. **(h)** The **XY:H+B** side-chain knob-socket. The tetrahedral arrangement of the non-local 4 residue knob-socket packing cliques in 3D space is drawn showing the knob **B** residue contacting all 3 side-chain **XY:H** socket residues. **(i)** The **XY:H+B** main-chain knob-socket. The tetrahedral arrangement of the 4 residue **2:1+1** non-local knob-socket packing clique in 3D space is drawn showing the knob **B** residue contacting the main-chain atoms of residue **Y** and the side-chains residues of **X** and **H**.

Figure 3. Free and Filled XY:HG Pocket Distributions

Distributions of pockets formed by the most frequent 50 **XY** residue pairs and 50 **HG** residue pairs in free and filled pockets are compared. **XY** residue pairs are presented on the y-axis and **HG** pairs are on the x-axis. The figure was generated using the R program package.⁴⁵ The upper two panels compare free socket distributions, where the frequency of free sockets is shown on the left and the filled sockets is on the right. The bottom two panels compare filled sockets, where the frequency of free sockets is shown again on the left and the filled sockets is on the right. So the 2 panels on the left represent frequencies of free **XY:H**G pockets, while the 2 panels on the right represent frequencies of filled **XY:H**G pockets. To improve interpretation, the residue pairs are grouped into “Nonpolar”, “Polar/Nonpolar”, and “Polar” groups. The color ramp on the right side shows the frequency values ranging from high (dark red) to low (blue) and to lowest (grey). The white spaces means the pockets are not observed.

Figure 4. Amino Acid Frequencies in the XY:H+B knob-socket

Heat maps showing the 20 amino acid distributions in 2 types of knob-socket motifs. For each position, percentages of each amino acid were calculated and converted to the grey scale color ramp on the right to indicate amino acid preferences. The figure was generated using the R program package.⁴⁵ The actual values for all amino acid percent distributions are provided in Table S2. **(a)** The positions **X**, **Y**, and **H** are shown for side-chain and main-chain sockets. **(b)** The knob **B** residue distributions that pack into the side-chain or main-chain sockets.

Figure 5. Socket preferences of 20 amino acid knobs

Heat maps showing the preferences for knob **B** residues versus (a) side-chain sockets and (b) main-chain sockets. The figure was generated using the R program package.⁴⁵ The amino acid distribution in knob-sockets at **X**, **Y**, **H**, and **B** positions are calculated and the top 100 **XY:H** sockets are presented in the heat map. The 20 knob **B** amino acids are on x-axis, and **XY:H** is on the y-axis. The color ramp on the right side shows the frequency values ranging from high (dark red) to low (blue) and to lowest (grey). The knob-sockets in white are not observed.

Figure 6. Visualizing β -sheet tertiary packing

(a) Ribbon diagram of antitumor antibody, 1ad0⁴² that consists of two β -sheets packed against each other. (b) The packing side of the bottom β -sheet showing residues with side chains facing against the other sheet. (c) The packing side of the top β -sheet. (d) The knob-socket lattice representation of the bottom β -sheet shown in (b). (e) The lattice representation of the top β -sheet shown in (c). In the β -sheet lattices, a solid grey line represents each β -strand. Representing the packing interface, the residues within white circles are side-chains facing out of the page and form filled **XY:H** sockets. The filled **XY:H** sockets involved in β -sheet tertiary structure are shaded grey, and knob **B** residues are represented by the single letter amino acid code with residue numbers in a sphere. To provide clarity, the **XY:H** sockets sharing a knob **B** residue are surrounded with solid black lines.