

1 **Characteristics of Interactions at Protein Segments External to 2 Globular Domains in the Protein Data Bank**

3 Kota Kasahara^{1,*}, Shintaro Minami², Yasunori Aizawa³

4 ¹ College of Life Sciences, Ritsumeikan University, 1-1-1, Noji-higashi, Kusatsu, Shiga 525-8577,
5 Japan

6 ² Exploratory Research Center on Life and Living Systems, National Institutes for Natural Sciences,
7 38 Nishigo-Naka, Myodaiji, Okazaki, Aichi 444-8585, Japan

8 ³ School of Life Science and Technology, Tokyo Institute of Technology, 4259, Nagatsuda-cho,
9 Midori-ku, Yokohama, Kanagawa 226-8501, Japan

10

11 * To whom correspondence should be addressed. Tel: +81-77-556-1111; Fax: +81-77-561-3729; Email:
12 tkshr@fc.ritsumei.ac.jp

13

14 Running title:

15 Floating segments in protein structures

16

17

18

1

2 **ABSTRACT**

3 The principle of three-dimensional protein structure formation is a long-standing conundrum in
4 structural biology. A globular domain of a soluble protein is formed by a network of atomic contacts
5 among amino acid residues, but regions external to globular domains, like loop and linker, often do
6 not have intramolecular contacts with globular domains. Although these regions can play key roles for
7 protein function as interfaces for intermolecular interactions, their nature remains unclear. Here, we
8 termed protein segments external to globular domains as *floating* segments and sought for them in
9 tens of thousands of entries in the Protein Data Bank. As a result, we found that 0.72 % of residues
10 are in floating segments. Regarding secondary structural elements, coil structures are enriched in
11 floating segments, especially for long segments. Interactions with polypeptides and polynucleotides,
12 but not small compounds, are enriched in floating segments. The amino acid preferences of floating
13 segments are similar to those of surface residues, with exceptions; the small side chain amino acids,
14 Gly and Ala, are preferred, and some charged side chains, Arg and His, are disfavored for floating
15 segments compared to surface residues. Our comprehensive characterization of floating segments
16 may provide insights into understanding protein sequence-structure-function relationships.

17

18 **KEYWORDS**

19 amino acid propensity / loop region / protein data bank / secondary structure / structural bioinformatics

20

21 **INTRODUCTION**

22 Elucidating the principles of the three-dimensional (3D) structure formation of proteins is a long-
23 standing conundrum in the field of structural biology [1]. How a sequence of 20 types of amino acid
24 residues in a polypeptide determines its 3D structure remains largely unclear. Toward illumination of
25 this issue, extensive efforts in structural biology have accumulated a massive amount of structural
26 data for proteins in the Protein Data Bank (PDB) [2]. By taking advantage of the wealth of structural
27 data, the field of so-called “structural bioinformatics”, which tackles extracting biological knowledge
28 from structural databases by using techniques of information science, has arisen [3-6]. Statistical
29 analyses of structural elements in the PDB have provided a bird’s eye view on the characterization of
30 the 3D structures of proteins. For example, statistical analyses revealed amino acid propensities
31 associated with several features, e.g., formation of secondary structural elements [7-10] and loop
32 regions [11]. In addition to secondary structure formation, another key feature to establish protein

1 folding is intramolecular contact between amino acid residues distant in the primary structure; in this
2 paper, we refer to this type of intramolecular contact as *non-local*: contacts between neighboring
3 residues are termed *local contacts*. The statistical analyses of the vicinity of amino acid residues in 3D
4 space have been extensively performed to predict and recognize protein folds [12,13]. For example,
5 the residue-wise contact order, defined as the summation over the distance along the sequence
6 between contacting residues, contains significant information regarding 3D structures (Kinjo and
7 Nishikawa, 2005; Kurt et al., 2008). Propensities of non-local contacts play pivotal roles in
8 establishing folds of globular domains.

9 On the other hand, proteins also have regions without non-local contacts. A typical example is
10 the linker region, which is a flexible segment linking two globular domains. The structural element
11 termed the loop region also tends to have no or only a few non-local contacts. They are also key
12 elements in the 3D structures of proteins. However, in spite of their importance, the nature of the
13 regions without non-local contacts is not well understood.

14 Here, we performed statistical analyses on PDB entries to investigate the nature of regions
15 without non-local contacts. Tens of thousands of PDB entries were processed, and two types of
16 regions consisting of consecutive amino acid residues were defined: (i) regions with non-local
17 contacts, and (ii) those without non-local contacts. We refer to these regions as (i) *supported*
18 segments and (ii) *floating* segments, respectively (Fig 1). We aim to characterize the floating
19 segments in proteins. On the basis of non-redundant PDB entries, the frequency of these floating
20 segments was analyzed with many features of the segments, e.g., length, secondary structures,
21 accessible surface area (ASA), and intermolecular interactions. We present that a considerable
22 number of residues are floating in protein structures deposited in the PDB. While amino acid
23 preferences of floating segments are similar to those of exposed residues, some amino acids
24 exhibited a unique propensity for floating segments.

25

26 MATERIALS AND METHODS

27 Dataset Construction

28 In this study, we constructed two kinds of datasets, named *primary dataset* and *non-redundant*
29 *datasets*. The latter are subsets of the first. The primary dataset was constructed by extracting entries
30 from a snapshot of the PDB on June 14, 2017, with the following criteria: (i) the entry contains at least
31 one polypeptide, (ii) the number of atoms is less than one million, and (iii) the structure was solved
32 with X-ray crystallography with resolution better than, or equal to, 3.0 Å. Extracting information from
33 the PDB was performed by parsing PDBML [14] with in-house scripts (S1 Data).

1 Non-redundant datasets were constructed by picking non-redundant entries from the primary
2 dataset. Single-linkage clustering with 40 % sequence identity was performed using the CD-HIT
3 program [15]. In cases where an entry had more than one chain, the sequence identity of the most
4 similar pair of chains was considered. A non-redundant dataset was constructed by random picking of
5 one entry from each cluster. We constructed 100 non-redundant datasets with different random seeds,
6 and statistics involving them were analyzed.

7 **Detecting Atomic Contacts and Definition of Structural Units**

8 In this study, we assessed interatomic contacts between heavy atoms with the threshold that the
9 interatomic distance is less than 5 Å. When a pair of amino acid residues has at least one interatomic
10 contact, we considered that this pair of residues has an *inter-residue contact*. An inter-residue contact
11 formed between residues of the same molecule is termed an *intramolecular contact*, which can be
12 grouped into two classes: (i) non-local contact for the cases in which the two contacting residues are
13 distant more than five residues in the sequence order, and (ii) local contact for other cases. On the
14 other hand, an *intermolecular contact* is defined as a contact between two different molecules. On the
15 basis of inter-residue contacts, we defined a structural unit named *segment*. A unit consisting of more
16 than three successive amino acid residues with non-local contacts is defined as a *supported* segment,
17 and that without non-local contacts is defined as a *floating* segment (Fig 1).

18

19 **Analyses**

20 We characterized segments in polypeptide chains by focusing on the following points: (i) segment
21 type, defined as floating or supported, (ii) segment length, (iii) secondary structural elements (SSEs),
22 (iv) accessible surface area (ASA), and (v) types of inter-molecular contact partners. (i) The segment
23 type is signified as $T_{seg} \subset \{flo, sup\}$; *flo* and *sup* mean floating and supported segments, respectively.
24 (ii) Segment length L_{seg} is the number of consecutive amino acid residues composing the segment.
25 For simplicity, segment lengths fell into three classes: $T_{len} \subset \{short, middle, long\}$, that were defined as
26 $2 < L_{seg} \leq 4$, $4 < L_{seg} \leq 9$, and $9 < L_{seg}$, respectively. (iii) The type of SSE (T_{SSE}) was assessed by using
27 the DSSP program [16]. In this manuscript, we applied three categories of SSE; $T_{SSE} \subset$
28 $\{helix, beta, coil\}$; *helix* was *G*, *H* or *I* in the DSSP classification, *beta* was *B*, *E*, *T*, or *S*, and *coil* was
29 the others. The representative T_{SSE} of each segment was the most frequent SSE in the residues
30 composing the segment. (iv) The solvent accessibility of each segment was defined as the two levels:
31 surface and buried, $T_{ASA} \subset \{sur, bur\}$. For each amino acid residue, when the ratio of ASA of the
32 residue to that of the same amino acid in Gly-X-Gly motifs is greater than 0.2, the residue is assumed
33 to be surface exposed. Otherwise, it is assumed to be buried. ASA was calculated with the DSSP
34 program [16]. (v) Partners of intermolecular contacts fell into three classes: polypeptides,
35 polynucleotides, and small compounds. These were defined by the entity type as described in the
36 PDB annotation. We considered non-polymer entities ≥ 300 Da as small compounds. The type of

1 interaction partner of a segment is denoted as $T_{int} \subset \{pep, nuc, sc\}$ for polypeptide, polynucleotide, and
2 small compound, respectively.

3 Relative frequencies of various types of segments were analyzed. $F(x)$ indicates the relative
4 frequency of segments with the condition x ,

5
$$F(x) = \frac{N(x)}{N_{total}}, \quad (1)$$

6 where N_{total} denotes the total number of segments in a dataset, and $N(x)$ denotes the number of
7 segments with the condition x . For example, the relative frequency of short segments and that of helix
8 segments are represented as $F(T_{len}=\text{short})$ and $F(T_{SSE}=\text{helix})$, respectively. For simplicity, they can
9 also be denoted as $F(\text{short})$ and $F(\text{helix})$. The conditional relative frequency, the ratio of the number
10 of segments with types x and y to that with type y , is denoted as $F(x/y)$.

11 In addition, we also analyzed characteristics of residues. The relative frequency of residues
12 with the condition x is presented as $F^{res}(x)$. The amino acid type of each residue is denoted as $T_{AA}^{res} \subset$
13 $\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$. The propensity score of each amino acid was assessed with
14 the log-odds score.

15
$$S^{res}(T_{AA}^{res} | x) = \log \left(\frac{F^{res}(T_{AA}^{res} | x) / (1 - F^{res}(T_{AA}^{res} | x))}{F^{res}(T_{AA}^{res}) / (1 - F^{res}(T_{AA}^{res}))} \right), \quad (2)$$

16
$$S^{res}(T_{AA}^{res} | x; y) = \log \left(\frac{F^{res}(T_{AA}^{res} | x, y) / (1 - F^{res}(T_{AA}^{res} | x, y))}{F^{res}(T_{AA}^{res} | y) / (1 - F^{res}(T_{AA}^{res} | y))} \right), \quad (3)$$

17 where x and y indicate conditions. We evaluated the amino acid propensities for segment types S^{res}
18 ($T_{AA}^{res} | T_{seg}$), those for interaction partners in each segment type $S^{res}(T_{AA}^{res} | T_{int}; T_{seg})$, and those for
19 surface or buried residues $S^{res}(T_{AA}^{res} | T_{ASA})$.

20

21 RESULTS AND DISCUSSION

22 Database Statistics

23 The primary dataset consisted of 89,038 PDB entries, 115,287 entities, 225,366 chains, and
24 53,497,598 residues. Single-linkage clustering with 40 % sequence identity resulted in 15,351 clusters.
25 Among these, 12,513 clusters (81.51 %) had less than, or equal to, five members, and 6,846 clusters
26 (44.60 %) of these were singleton clusters (S1 Figure). A non-redundant dataset was constructed by
27 randomly picking one entry from each cluster, and statistical analyses were performed on 100 non-

1 redundant datasets. On average [with standard deviation (SD)], a non-redundant dataset consisted of
2 17,259.01 [37.44] entities, 38,601.25 [140.30] chains, and 8,967,882 [34,917.46] residues.

3 For the PDB entries, we defined floating segments as having at least three consecutive
4 residues without non-local contacts, and defined supported segments as those with non-local
5 contacts, respectively. The primary dataset included 359,501 floating segments and 4,419,603
6 supported ones, with approximately half of residues belonging to a segment. The average numbers
7 [and SD] of floating and supported segments in the non-redundant datasets were 64,285.16 [387.25]
8 and 725,750.9 [2,960.58], respectively. 0.72 % of residues were in floating segments.

9

10 Segment Length and Secondary Structure Elements

11 Distributions of the segment length L_{seg} for each segment type are shown in Fig 2. The number of
12 segments decreases exponentially, along with an increase in the segment length. In particular, there
13 is a steep decrease in the frequency for floating segments in $L_{seg} < 9$. A majority of floating segments
14 consist of only three or four residues ($F(short/flo) = 0.850$), and the ratio of floating segments that are
15 longer than nine residues is only $F(long/flo) = 0.014$. On the other hand, many supported segments
16 have more than nine segments ($F(long/sup) = 0.103$; $F(short/sup) = 0.568$). Floating segments tend to
17 be shorter than supported ones. This may reflect the fact that longer protein regions without molecular
18 contacts should have higher flexibility, and thus it is difficult to identify atomic coordinates of such
19 regions by crystallography. The long flexible regions, such as intrinsically disordered regions, are
20 abundant in nature although their atomic coordinates are not recorded in the PDB. Note that in this
21 study we did not consider missing regions in the PDB entries, which means regions without atomic
22 coordinates.

23 Segment length relates to the secondary structure formation of segments. Longer segments
24 tend to favor formation of helical structures, and only a small fraction of β -structures is observed in
25 long floating segments (Fig 3). Namely, $F(helix/long, flo) = 0.664$, and $F(beta/long, flo) = 0.0410$. This
26 result reflects the fact that a floating, single α -helix is more stable than a floating single β -strand. A
27 majority of long floating β -strands in the dataset have intermolecular contacts. A typical case is
28 formation of an intermolecular β -sheet (S2A Figure; PDB ID: 2O8M). On the other hand, floating
29 helical segments are mainly observed at a solvated terminus of the chain (S2B Figure; PDB ID:
30 2WN9), in a coiled-coil (S2C Figure; PDB ID: 2GL2), and other types of interfaces (S2D Figure; PDB
31 ID: 3RK0). For cases of unstructured (coil), long floating segments, typical cases are at the termini of
32 proteins (S2E Figure; PDB ID: 3TER). The segments in the linker region are also observed (S2F
33 Figure; PDB ID: 2B58). In contrast to the floating segments, the β -structures are not disfavored in
34 long-supported segments (Fig 3). Since β -stands are usually formed as a part of a β -sheet, broad
35 lengths of supported β -stands can be found in protein structures. Since the usual size of β -sheets
36 accords to the category of $T_{len} = middle$, a higher ratio of beta structures in the middle-supported

1 segments are observed, compared with other lengths. Some examples of supported segments in
2 each length are shown in S3A, B, and C Figures. Regarding coil structures, in contrast to the case of
3 floating segments, only a small ratio of longer-supported segments adopt coiled structures. Typical
4 long-supported segments with coil structures penetrate inside the domain (S3D Figure; PDB ID:
5 2BSL), or locate around the surface of the domain (S3E Figure; PDB ID: 3TEH).

6

7 **Segments as interfaces of intermolecular interactions**

8 The relative frequencies of segments for intermolecular interactions are summarized in Fig 4. The
9 interaction partners were categorized into one of the three types; that is, polypeptides ($T_{int} = pep$),
10 polynucleotides ($T_{int} = nuc$), and small compounds ($T_{int} = sc$). For interactions with polypeptides and
11 polynucleotides, floating segments exhibit a higher frequency of appearance in intermolecular
12 interfaces compared with supported segments; the ratio of interacting floating segments are $F(pep/flo) = 0.424$,
13 $F(nuc/flo) = 0.0121$, and those values for supported segments are $F(pep/sup) = 0.282$ and
14 $F(nuc/sup) = 0.00659$, respectively. This is because intermolecular interactions require high solvent
15 accessibility, and almost all floating segments are on the surface ($F(sur/flo) = 0.953$) and this ratio in
16 supported segments is $F(sur/sup) = 0.397$. However, although floating segments are enriched at the
17 surface, binding sites for small compounds prefer supported segments rather than floating ones; the
18 ratios of interacting-floating and interacting-supported segments are $F(sc/flo) = 0.0190$ and $F(sc/sup) = 0.0710$,
19 respectively. Since the binding sites are usually formed as a concave surface (called a
20 cavity or pocket) with a certain size and depth [17], they should be formed by supported segments
21 rather than by floating ones. As an example of binding sites with floating segments, 3-chlorocatechol
22 1,2-dioxygenase binds its ligand with a floating helix-turn-helix conformation (S4A Figure; PDB ID:
23 2BOY). In addition, many entries in this category do not have biologically relevant ligand-binding sites
24 but have contacts with other non-specific small molecules such as lipids. For instance, a light
25 harvesting complex is surrounded by chlorophyll molecules (S4B Figure; PDB ID: 3PL9).

26 Regarding the segment length, longer segments show higher frequencies for interactions
27 since longer segments have larger surface areas, which simply elevates the probability of interactions.
28 For floating segments, typical interactions of long helix and long beta segments are formation of
29 coiled-coil structures and inter-molecular β -sheets, respectively, as shown above (S2C Figure and
30 S2A Figure, respectively). As an exception to the tendency of the segment length, longer supported
31 segments are not enriched in polynucleotide binding sites. Many typical double-stranded DNA binding
32 sites include floating segments with positively charged amino acid residues. For example, bZIP
33 heterodimeric complexes recognize DNA with two floating helices (S5A Figure; PDB ID: 2WT7). A
34 floating-helix segment consisting of 18 residues in the NC2–TBP–DNA ternary complex structure
35 recognizes the DNA with their six acidic residues (S5B Figure; PDB ID: 1JFI). A linker loop with two
36 acidic residues in a replication terminator protein is buried into the major groove of DNA (S5C Figure;
37 PDB ID: 1ECR). In contrast, it is sterically difficult to attach grooves of a double-stranded DNA to

1 supported segments. A majority of supported segments at the polynucleotide binding sites touch the
2 DNA backbone rather than burying into the grooves (for example PDB ID: 3E3Y; S5D Figure). In
3 addition, many single-stranded DNA binding sites are composed of supported segments (for example
4 S5E and F Figures; PDB ID: 2KFN and 3CMW, respectively).

5

6 **Propensities of Amino Acids**

7 Propensity of each amino acid for floating and supported segments was assessed based on the log-
8 odds score (Fig 5B; Eq. 2). In general, bulky non-polar amino acids, e.g., Cys, Phe, Ile, Leu, Met, Val,
9 Trp, and Tyr, are disfavored for floating segments. This tendency is similar for surface residues (Fig 6;
10 the Pearson correlation coefficient (PCC) of propensity scores between floating segments and surface
11 residues is 0.929). This is due to the fact that a majority of floating segments are at the surface;
12 $F(sur/flo) = 0.953$. However, there are some unique features in the amino acid propensity for the
13 floating segments compared to surface residues. (i) Arg and His are disfavored in floating segments,
14 although they are highly enriched as surface residues due to their high polarity. In many cases, they
15 are involved in interfaces of intermolecular contacts. Arg is enriched for contacts with polynucleotides.
16 For example, a 20-residue segment in the Fos-Jun complex has seven Arg residues (S6A Figure;
17 PDB ID: 1A02). At an interface with polypeptides, Arg can stabilize the interactions through formation
18 of salt bridges (S6B Figure; PDB ID: 2E7S). Many His residues in floating segments are observed in
19 His-tag sequences. (ii) Gly is preferred for floating segments, although it is not so favored in surface
20 residues. This is due to its high flexibility, which makes it possible to form unstructured regions,
21 including loops and linkers. For example, a 12-residue segment in a loop region of MHC molecules
22 has five Gly residues (S6C Figure; PDB ID: 1LNU). (iii) Ala is not so disfavored for the floating
23 segments, in spite of its negative score in surface residues due to its hydrophobic side chain. One
24 possible explanation is that α -helices favor Ala residues [7,18]. Floating segments show higher ratios
25 of helix conformation than supported segments (Fig 3). An example of a floating helix with many Ala
26 residues is shown in S6D Figure (PDB ID: 4KE2).

27 We also assessed amino acid propensities for intermolecular interactions with the three
28 categories of molecules; polypeptides, polynucleotides, and small molecules (Figs 5C, D, and E,
29 respectively). The propensity score of floating segments for polypeptide interactions (black in Fig 5C)
30 shows the opposite trend from that of propensity to form floating segments (Fig 6B; their PCC is -
31 0.872). Although hydrophobic amino acids are disfavored for floating segments, they are favored for
32 intermolecular interaction interfaces with other polypeptides. This implies that when disfavored amino
33 acids exist in a floating segment, it is expected that they conduct some functions to recognize another
34 proteins. The exception is Gln, which has a positive propensity score for both conditions; $S^{res}(Q; flo)$
35 and $S^{res}(Q; pep/flo)$ shown in Figs 5B and C, respectively. Gln is often observed at terminal or kinking
36 regions of a helix (S6E and F Figures for examples). The propensity for floating segments in small-
37 molecule binding sites showed a weak correlation to that for polypeptides (Fig 6C; the PCC is 0.614).

1 The major differences are as follows: Gly and Pro are favored, and Gln is disfavored for floating small-
2 molecule binding sites. Since Gly and Pro are enriched in flexible regions, they are observed in loop
3 regions composing a binding site (examples are shown in S6G Figure and H). For the propensity for
4 floating segments in interfaces to polynucleotide, there is no clear correlation with other propensities.
5 In a comparison with the trend in supported segments, floating segments disfavor some hydrophobic
6 amino acids, e.g., Cys, Leu, Pro, and Trp. In addition, while Asp and Glu are disfavored, Asn and Gln
7 are not. They sometimes have direct contacts with a base of polynucleotides (S6I Figure).

8

9 **CONCLUSIONS**

10 In this study, we defined floating and supported segments involved in the 3D structure of proteins (Fig
11 1), and characterized them on the basis of statistical analyses of the PDB. We found considerable
12 numbers of floating segments in known protein structures (0.72 % of residues are in floating
13 segments). The frequency distribution of segment length shows exponential decay along with an
14 increase in the segment length, in both floating and supported segments. The length distribution of
15 floating segments is more biased toward shorter regions than that of supported segments, and most
16 of the floating segments are composed of three or four residues (Fig 2). Three is the minimum length
17 of a segment in the definition; the segment length largely impacts its characteristics. Shorter floating
18 segments tend to form secondary structures (Fig 3). Longer floating segments are enriched in
19 intermolecular interaction interfaces. In particular, beta structures are favored for long-floating
20 segments at the interfaces (Fig 4). Although floating segments are enriched at interfaces for
21 polypeptides and polynucleotides, they are disfavored at interfaces for small compounds (Fig 4).
22 Regarding the amino acid composition, while floating segments are basically similar to surface
23 exposed residues, they have some unique features; higher preferences for small side chains (Gly and
24 Ala) and disfavoring some charged side chains (Arg and His) compared to surface residues (Fig 5A).
25 Interestingly, the propensity scores for polypeptide interactions of floating segments are in an
26 opposite trend from that for all floating segments (Figs 5A and B). Residues disfavored for floating
27 residues tend to be interfaces for protein–protein interactions at floating segments, except for Gln
28 residues.

29

30 **ACKNOWLEDGEMENT**

31 Supercomputer resources were provided by the National Institute of Genetics, Research Organization
32 of Information and Systems, Japan.

33

1 FUNDING

2 This work was supported by the Japan Society for the Promotion of Science KAKENHI, Grant-in-Aid
3 for Young Scientists (Grant Number: JP16K18526).

4

5 REFERENCES

- 6 1. Richardson JS. The anatomy and taxonomy of protein structure. *Adv Protein Chem.* 1981;34: 7 167–339.
- 8 2. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): 9 ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* 2007;35: D301–3. 10 doi:10.1093/nar/gkl971
- 11 3. Thornton JM, Todd AE, Milburn D, Borkakoti N, Orengo CA. From structure to function: 12 approaches and limitations. *Nat Struct Biol.* 2000;7 Suppl: 991–994. doi:10.1038/80784
- 13 4. Kinoshita K, Nakamura H. Protein informatics towards function identification. *Curr Opin Struct 14 Biol.* 2003;13: 396–400.
- 15 5. Sudha G, Nussinov R, Srinivasan N. An overview of recent advances in structural 16 bioinformatics of protein–protein interactions and a guide to their principles. *Progress in 17 Biophysics and Molecular Biology.* Pergamon; 2014;116: 141–150.
- 18 6. Lobb B, Doxey AC. Novel function discovery through sequence and structural data mining. 19 *Curr Opin Struct Biol.* Elsevier Current Trends; 2016;38: 53–61. doi:10.1016/j.sbi.2016.05.017
- 20 7. Wang J, Feng JA. Exploring the sequence patterns in the α -helices of proteins. *Protein Eng Des Sel.* Oxford University Press; 2003;16: 799–807. doi:10.1093/protein/gzg101
- 22 8. Pal L, Chakrabarti P, Basu G. Sequence and Structure Patterns in Proteins from an Analysis 23 of the Shortest Helices: Implications for Helix Nucleation. *J Mol Biol.* 2003;326: 273–291. 24 doi:10.1016/S0022-2836(02)01338-4
- 25 9. Costantini S, Colonna G, Facchiano AM. Amino acid propensities for secondary structures are 26 influenced by the protein structural class. *Biochemical and Biophysical Research 27 Communications.* Academic Press; 2006;342: 441–451. doi:10.1016/j.bbrc.2006.01.159
- 28 10. Rani P, Baruah A, Biswas P. Does Lack of Secondary Structure Imply Intrinsic Disorder in 29 Proteins? A Sequence Analysis. *Biochim Biophys Acta Proteins Proteomics.* Elsevier; 30 2014;1844: 1827–1834. doi:10.1016/j.bbapap.2014.07.020
- 31 11. Crasto CJ, Feng JA. Sequence codes for extended conformation: A neighbor-dependent 32 sequence analysis of loops in proteins. *Proteins.* Wiley-Blackwell; 2001;42: 399–413. 33 doi:10.1002/1097-0134(20010215)42:3<399::AID-PROT100>3.0.CO;2-E
- 34 12. Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids 35 Res.* Oxford University Press; 1997;25: 231–234. doi:10.1093/nar/25.1.231
- 36 13. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis 37 of residue coevolution captures native contacts across many protein families. *Proc Natl Acad 38 Sci USA.* National Academy of Sciences; 2011;108: E1293–E1301. 39 doi:10.1073/pnas.1111471108

- 1 14. Westbrook J, Ito N, Nakamura H, Henrick K, Berman HM. PDBML: the representation of
2 archival macromolecular structure data in XML. *Bioinformatics*. 2005;21: 988–992.
3 doi:10.1093/bioinformatics/bti082
- 4 15. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
5 sequencing data. *Bioinformatics*. Oxford University Press; 2012;28: 3150–3152.
6 doi:10.1093/bioinformatics/bts565
- 7 16. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of
8 hydrogen-bonded and geometrical features. *Biopolymers*. Wiley Subscription Services, Inc., A
9 Wiley Company; 1983;22: 2577–2637. doi:10.1002/bip.360221211
- 10 17. Kawabata T, Go N. Detection of pockets on protein surfaces using small and large probe
11 spheres to find putative ligand binding sites. *Proteins*. Wiley Subscription Services, Inc., A
12 Wiley Company; 2007;68: 516–529. doi:10.1002/prot.21283
- 13 18. Spek EJ, Olson CA, Shi Z, Kallenbach NR. Alanine Is an Intrinsic α -Helix Stabilizing Amino
14 Acid. *J Am Chem Soc*. 1999;121: 5571–5572. doi:10.1021/ja990056x

15

16 **FIGURE LEGENDS**

17 **Figure 1.** Schematic illustration of *floating* and *supported* segments. Open and shaded arrows
18 indicate intermolecular and intramolecular contacts, respectively. The shaded regions in polypeptides
19 are floating segments.

20 **Figure 2.** Histogram of floating (black) and supported (grey) segments regarding segment lengths.
21 Error bars are the standard deviations in 100 non-redundant datasets.

22 **Figure 3.** Ratio of the SSE for each combination of classes: T_{len} and T_{seg} . Black, dark grey, and light
23 grey denote the ratios of *beta*, *helix*, and *coil* structures, respectively. Error bars are the standard
24 deviations in 100 non-redundant datasets.

25 **Figure 4.** The ratios of interacting segments for each combination of classes: T_{SSE} , T_{len} and T_{seg} . (A)
26 Ratios for interactions with polypeptides. (B) Ratios for interactions with polynucleotides. (C) Ratios
27 for interactions with small compounds. Black, dark grey, and light grey denote the ratios of *beta*, *helix*,
28 and *coil* structures, respectively. Error bars are the standard deviations in 100 non-redundant
29 datasets.

30 **Figure 5.** Amino acid propensity scores. (A) The propensities for surface or buried residues; S^{res}
31 ($T_{AA}^{\text{res}} | T_{ASA}$). Black and grey bars denote surface and buried residues for the panel, respectively. (B)
32 The propensities to form floating or supported segments; $S^{\text{res}}(T_{AA}^{\text{res}} | T_{\text{seg}})$. (C) The propensities to
33 interact with polypeptides; $S^{\text{res}}(T_{AA}^{\text{res}} | T_{\text{int}} = \text{pep}; T_{\text{seg}})$. (D) The propensities to interact with
34 polynucleotides; $S^{\text{res}}(T_{AA}^{\text{res}} | T_{\text{int}} = \text{nuc}; T_{\text{seg}})$. (E) The propensities to interact with small compounds; S^{res}
35 ($T_{AA}^{\text{res}} | T_{\text{int}} = \text{sc}; T_{\text{seg}}$). Black and grey bars indicate floating and supported segments for the panels (B),
36 (C), (D), and (E). Error bars are the standard deviations in 100 non-redundant datasets.

1 **Figure 6.** Comparisons of the amino acid propensity scores shown in Fig 5. (A) Comparison between
2 floating and surface segments. (B) Comparison between floating segments and those with peptide
3 interactions. (C) Comparison between floating segments with small compound interactions and those
4 with peptide interactions.

5 **S1 Figure.** The number of PDB entries in each cluster.

6 **S2 Figure.** Examples of long floating segments. The target segment, the chain including the segment,
7 and other chains are shown in red, green, and gray. (A) A floating-beta segment composing an
8 intermolecular β -sheet (a virus serine protease; PDB ID: 2O8M). (B) A floating-helix segment at a
9 solvated terminus (an acetylcholine receptor; PDB ID: 2WN9). (C) A floating-helix segment in a coiled-
10 coil (adhesin FadA; PDB ID: 2GL2). (D) A floating-helix segment at an intermolecular interface (PDB
11 ID: 3RKO). (E) A floating-coil segment at a terminus (the membrane domain of respiratory complex I;
12 PDB ID: 3TER). (F) A floating-coil segment at a linker region (diamine acetyltransferase 1; PDB ID:
13 2B58).

14 **S3 Figure.** Examples of supported segments. The target segment, the chain including the segment,
15 and other chains are shown in red, green, and gray. (A) A supported-short segment in a β -sheet (a
16 putative nucleotide-diphospho-sugar transferase; PDB ID: 3CGX). (B) A supported-short segment in a
17 β -sheet (kinase PhoQ catalytic domain; PDB ID: 3CGZ). (C) A supported-long segment in a β -sheet
18 (a tRNA synthetase; PDB ID: 3TEG). (D) A supported-coil segment penetrating a globular domain (a
19 dihydroorotate dehydrogenase A; PDB ID: 2BSL). (E) A supported-coil segment surrounding a
20 globular domain (a tRNA synthetase; PDB ID: 3TEH).

21 **S4 Figure.** Examples of small-compound binding sites with floating segments. The target segment,
22 the chain including the segment, and other chains are shown in red, green, and gray. (A) A floating
23 segment at a small-compound binding site. (3-Chlorocatechol 1,2-Dioxygenase; PDB ID: 2BOY, (B) A
24 floating segment interacting lipid molecules (a chlorophyll binding protein; PDB ID: 3PL9).

25 **S5 Figure.** Examples of nucleic acid binding sites with floating segments. The target segment, the
26 chain including the segment, and other chains are shown in red, green, and gray. (A) A floating-helix
27 segment in a bZIP heterodimeric complex (PDB ID: 2WT7). (B) A floating-helix segment in the NC2–
28 TBP–DNA ternary complex (PDB ID: 1JFI). (C) A coil-floating segment in a replication terminator
29 protein (PDB ID: 1ECR). (D) A supported segment in the restriction enzyme HindII (PDB ID: 3E3Y). (E)
30 A supported segment in the Klenow fragment of a DNA polymerase (PDB ID: 2KFN). (F) A supported
31 segment in RecA (PDB ID: 3CMW).

32 **S6 Figure.** Examples of floating segments with specific amino acids. The target amino acid residues,
33 the segment including the residues, the chain including the segment are shown in cyan, red, and
34 green, respectively. The binding partners are shown in yellow. (A) A floating segment with many Arg
35 residues in Jun-Fos heterodimer (PDB ID: 1A02). This segment recognizes the double-stranded DNA.
36 (B) A floating segment including Arg in the yeast Sec2p GEF domain (PDB ID: 2E7S). Arg forms the
37 salt-bridge with Asp in the other chain. (C) A floating segment with Gly residues in a MHC molecule

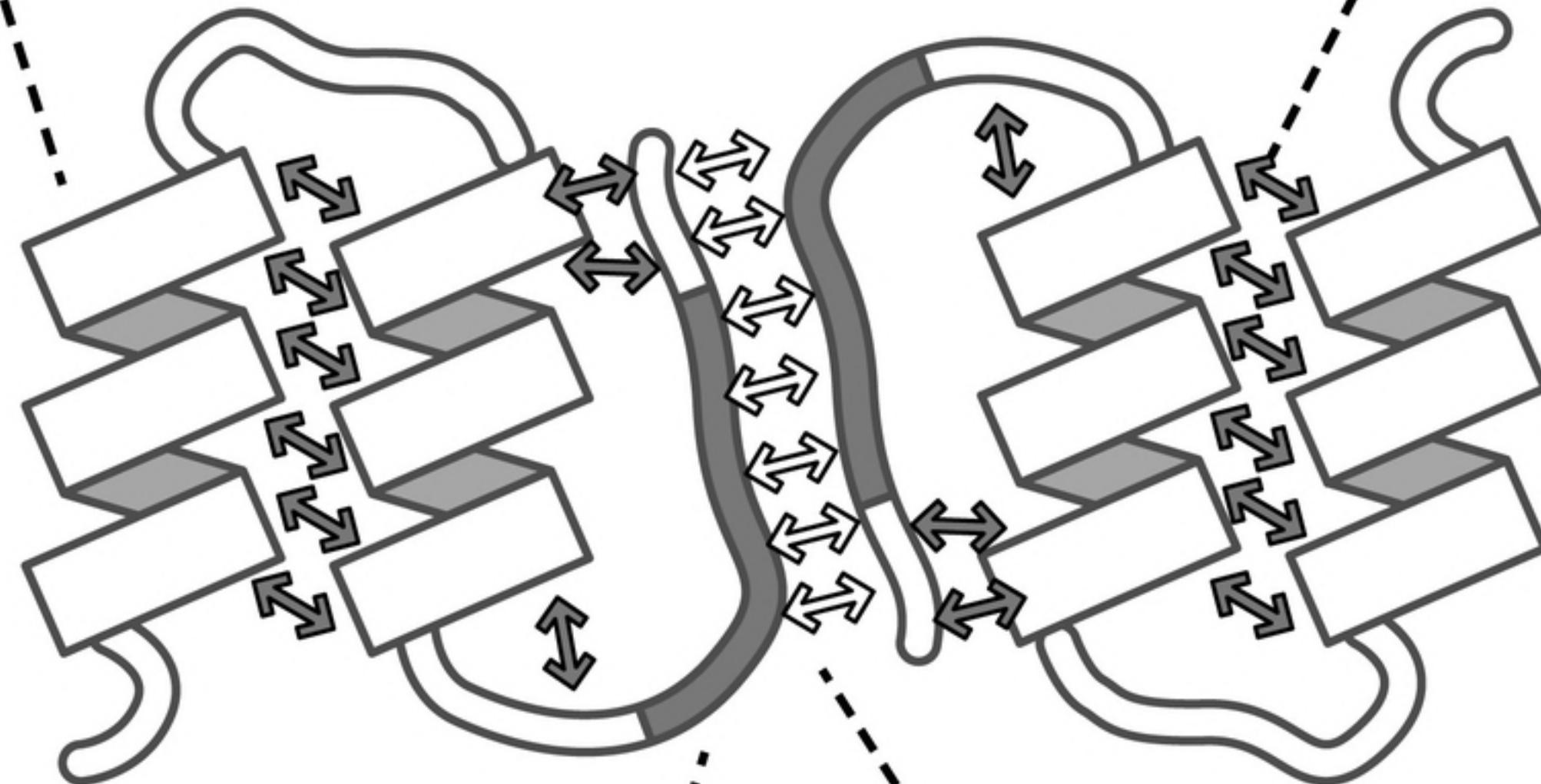
1 (PDB ID: 1LNU). (D) A floating segment with Ala residues in Type I hyperactive antifreeze protein
2 (PDB ID: 4KE2). (E, F) Floating fragments with Gln residues interacting with the other polypeptide: (E)
3 Huntingtin (PDB ID: 4FE8) and (F) enoyl reductase InhA (PDB ID: 4R9R). (G, H) Floating segments
4 including Gly residues at the binding interface to the small molecule: (G) HIV-1 protease (PDB ID:
5 1SH9), and (H) ecdysone receptor (PDB ID: 2R40). (I) A floating segment interacting with the RNA by
6 Pro residues in virus capsid (PDB ID: 1DDL). (J) A floating segment interacting with the siRNA duplex
7 by Asn residues in Piwi protein (PDB ID: 2GBB) .

8 **S1 Data.** Segments are recorded as a .csv file. Each column indicate PDB ID, the type of segment
9 (supported or floating), the length of segment, the type of length (short, medium, or long), the
10 secondary structural elements (helix, beta, or coil), the number of interacting molecules for
11 polypeptide, polynucleotide, and small compound, and the sequence of the segment.

12

Supported segment

Intramolecular contact



Floating segment

Intermolecular contact

