Supporting Information

Detecting Repetitions and Periodicities in Proteins by Tiling the Structural Space

R. Gonzalo Parra\textsuperscript{1}, Rocío Espada\textsuperscript{1}, Ignacio E. Sánchez\textsuperscript{1}, Manfred J. Sippl\textsuperscript{2} and Diego U. Ferreiro\textsuperscript{1}

\textsuperscript{1}Protein Physiology Lab, Dep de Química Biológica, Facultad de Ciencias Exactas y Naturales, UBA-CONICET-IQUIBICEN, Buenos Aires, Argentina.

\textsuperscript{2}Center of Applied Molecular Engineering, Division of Bioinformatics, Department of Molecular Biology, University of Salzburg, Austria.
Homogeneous Model

We build a toy-protein consisting of \( N \) amino acids arranged on a straight line, schematized on figure S1.

Each tile is uniquely characterized by its length \( L_i \) and it’s center \( Z_i \), and satisfies:

\[
\frac{L_i}{2} \leq Z_i \leq N - \frac{L_i}{2}.
\]

The maximal coverage is gained when the copies arranged continuously. As such, the centers of the copies (\( Z_{ik} \)) is

\[
Z_{ik} = Z_i + n \cdot L_i \quad n \in \mathbb{Z}.
\]

When only full length tile copies are accepted, the center of the copies are restricted,

\[
\frac{L_i}{2} \leq Z_{ik} \leq N - \frac{L_i}{2}.
\]

\[
\frac{L_i}{2} \leq Z_i + n \cdot L_i \leq N - \frac{L_i}{2}
\]

which can be rearranged to get:

\[
\frac{1}{2} - \frac{Z_i}{L_i} \leq n \leq \frac{N}{L_i} - \frac{1}{2} - \frac{Z_i}{L_i}
\]

with boundaries corresponding to

\[
n_{\text{min}} = \left\lfloor \frac{1}{2} - \frac{Z_i}{L_i} \right\rfloor
\]

\[
n_{\text{max}} = \left\lceil \frac{N}{L_i} - \frac{1}{2} - \frac{Z_i}{L_i} \right\rceil
\]
The number of copies that can be place along the protein is \( n_c = n_{\text{max}} - n_{\text{min}} + 1 \) and the coverage \( C_i = n_c \cdot L_i \). The tile score \( \Theta_i \) is:

\[
\Theta_i = \frac{(n_c - 1) L_{ii}}{N - L_{ii}} \tag{8}
\]

The tiling pattern of a toy-protein of \( N = 120 \) residues is shown in figure S2a.

To calculate the coverage when partial copies of a tile are allowed, we first calculate how many residues are uncovered by tessellation of full length tile copies. The most left tile is centered at \( Z_i + n_{\text{min}} \cdot L_i \), and its start is at the aminoacid \( Z_i + n_{\text{min}} \cdot L_i - \frac{L_{ii}}{2} \) which is also the number of amino acids uncovered at the beginning of the protein \( C_{\text{beg}} \). Analogously, the most right tile center is centered at \( Z_i + n_{\text{max}} \cdot L_i \), its end is located at \( Z_i + n_{\text{max}} \cdot L_i + \frac{L_{ii}}{2} \) and the number of amino acids uncovered is

\[
C_{\text{end}} = N - \left[Z_i + n_{\text{max}} \cdot L_i + \frac{L_{ii}}{2}\right].
\]

If \( C_{\text{beg}} \) and \( C_{\text{end}} \) are bigger than \( \alpha L_i \) the partial tile copies contribute to the coverage obtained with the full length copies:

\[
\chi(x) = \begin{cases} 
0 & \text{if } x < 0 \\
1 & \text{if } x \geq 0
\end{cases}
\tag{9}
\]

and their tile score is

\[
\Theta_i = \frac{(n_c - 1) \cdot L_{ii} + C_{\text{beg}} \cdot \chi(C_{\text{beg}} - L_{ii}/2) + C_{\text{end}} \cdot \chi(C_{\text{end}} - L_{ii}/2)}{N - L_{ii}} \tag{10}
\]

The tiling pattern of a toy-protein of \( N = 120 \) residues is shown in figure S2b.
Figure S2: Tiling an homogenous toy-protein. The tiles are ordered according to their size (vertical axis) and their center (horizontal axis) in amino acid units. The tile score $\Theta_i$ is displayed in greyscale, and $\delta\bar{\Theta}$ vs. $L_i$ projected on the left. Panel a) shows the tiling profile obtained when only full length copies of the tile are allowed, and panel b) when partial copies are also accepted.
Figure S3: Zoom in the tiling profile of Porcine Ribonuclease Inhibitor (2bnh,A) (shown in Fig3c). The tiling profile is shown on grayscale, together with the $\delta \Theta_i$ projected on the left. This protein has a characteristic frequency at $L_i=57$. The tiles at this length can be composed with tiles of $L_i=28$ and $L_i=29$ that appear alternated in protein structure. This is reflected in a square-tooth pattern around $L_i \approx 30$ and a second peak in $\delta \Theta_i$. The structures of the native protein and the corresponding tiling at the specified length ($L_i$) and center ($Z_i$) is shown, using the same coloring scheme of Fig.2. a) Tiling with $L_i=28$, $Z_i=388$. c) Tiling with $L_i=29$, $Z_i=153.5$. 
Figure S4: Tiling TIM barrels examples. The tiling profile is shown on grayscale, together with the $\delta \Theta_i$ projected on the left. The structures of the native protein and the corresponding tiling at lengths ($L_i$) and center ($Z_i$) corresponding to 2-fold, 4-fold and 8-fold are shown, using the same coloring scheme of Fig.2

a) Ribulose-phosphate 3-epimerase (pdb:1rpx,A) 2-fold: $L_i = 108$, $Z_i = 63$, 4-fold $L_i = 46$, $Z_i = 92$, 8-fold $L_i = 27$, $Z_i = 105.5$
b) HisF (pdb:1thf,D) 2-fold: $L_i = 121$, $Z_i = 181.5$, 4-fold $L_i = 53$, $Z_i = 58.5$, 8-fold $L_i = 22$, $Z_i = 94$
c) Glycosomal, Triosephosphate isomerase (pdb:5tim,A) 2-fold: $L_i = 129$, $Z_i = 186.5$, 4-fold $L_i = 50$, $Z_i = 29$, 8-fold $L_i = 23$, $Z_i = 115.5$
d) Narbonin (pdb:1nar,A) 2-fold: $L_i = 97$, $Z_i = 116.5$, 4-fold $L_i = 43$, $Z_i = 180.5$, 8-fold $L_i = 25$, $Z_i = 251.5
Figure S5: Tiling β-propeller examples. The tiling profile is shown on grayscale, together with the $\delta \Theta_i$ projected on the left. The structures of the native protein and the corresponding tiling at the specified length ($L_i$) and center ($Z_i$) is shown, using the same coloring scheme of Fig.2. β-propeller proteins that contain different number of 'blades' are shown. a) 5-bladed Tachylectin-2 (pdb:1tl2.A), $L_i = 47$, $Z_i = 213.5$ b) 7-bladed WD repeat-containing protein-5 (pdb:3smr.A), $L_i = 42$, $Z_i = 138$ c) 6-bladed 3-phytase (pdb:3ams.A), $L_i = 45$, $Z_i = 251.5$ d) 4-bladed Interstitial collagenase (pdb:1fbl.A) $L_i = 49$, $Z_i = 399.5$. 
**Table S1:** Survey of the tile and tessellation parameters in protein structures

| PdbID | Architecture | Protein | Selected Tile ($T_i$) | Tessellation |
|-------|--------------|---------|-----------------------|--------------|
|       |              |         | $N^a$ | $\Xi^b$ | $L_i^c$ | $Z_i^d$ | $\Theta_i^e$ | $n_{T_i^f}$ | $C_i^g$ | $I_i^h$ | $NR_i^2$ |
| HM    | toy model    | 120     | 1.0000 | 79*  | 39.5* | 1.00 | 2 | 1.00 | 0.00 | 0.00 |
| 2b9c,A| Coiled-Coil  | 136     | 0.8761 | 35  | 143.5 | 0.93 | 4 | 0.98 | 0.00 | 0.02 |
| 2fo7,A| TPR          | 136     | 0.7897 | 68* | 35*  | 1.00 | 2 | 1.00 | 0.00 | 0.00 |
| 3esk,A| TPR          | 128     | 0.7845 | 41* | 297.5* | 0.94 | 4 | 1.00 | 0.00 | 0.00 |
| 2bnh,A| Leucine      | 456     | 0.7628 | 57* | 139.5* | 0.96 | 8 | 0.99 | 0.00 | 0.01 |
| 2j8k,A| $\beta$-Solenoid | 175 | 0.7569 | 10* | 87* | 0.91 | 17 | 0.93 | 0.03 | 0.04 |
| 4db6,A| Armadillo    | 197     | 0.7367 | 42* | 33* | 0.99 | 5 | 1.00 | 0.00 | 0.00 |
| 3ltm,A| Heat         | 185     | 0.7039 | 31* | 62.5* | 0.97 | 6 | 1.00 | 0.00 | 0.00 |
| 2i13,A| Zn-Finger    | 154     | 0.7030 | 28* | 62*  | 0.87 | 5 | 0.89 | 0.00 | 0.11 |
| 2xtw,A| $\beta$-Solenoid | 210 | 0.7029 | 10  | 32  | 0.85 | 19 | 0.88 | 0.05 | 0.07 |
| 1k1a,A| Ank          | 228     | 0.7022 | 33  | 241.5 | 0.93 | 7 | 0.96 | 0.04 | 0.00 |
| 1awc,B| Ank          | 153     | 0.6918 | 66* | 89*  | 0.97 | 3 | 0.97 | 0.00 | 0.03 |
| 1n11,A| Ank          | 408     | 0.6804 | 33* | 510.5* | 0.90 | 12 | 0.95 | 0.00 | 0.05 |
| 1ihb,A| Ank          | 156     | 0.6794 | 33* | 79.5* | 0.95 | 5 | 0.99 | 0.01 | 0.00 |
| 1mx2,A| Ank          | 156     | 0.6789 | 33* | 79.5* | 0.95 | 5 | 0.99 | 0.01 | 0.00 |
| 3ltj,A| Heat         | 191     | 0.6771 | 93* | 59.5* | 0.73 | 2 | 0.97 | 0.00 | 0.03 |
| 1ixv,A| Ank          | 229     | 0.6765 | 34  | 190  | 0.90 | 7 | 1.00 | 0.00 | 0.00 |
| 3u4t,A| TPR          | 258     | 0.6757 | 34* | 382* | 0.85 | 7 | 0.92 | 0.04 | 0.04 |
| 1nfi,E| Ank          | 213     | 0.6748 | 106*| 123* | 0.79 | 2 | 0.92 | 0.05 | 0.03 |
| 3sla,A| Armadillo    | 166     | 0.6744 | 42  | 204  | 0.89 | 4 | 0.98 | 0.02 | 0.00 |
| 1n0r,A| Ank          | 126     | 0.6725 | 33* | 18.5* | 0.97 | 4 | 0.99 | 0.00 | 0.01 |
| 2rfm,A| Ank          | 183     | 0.6699 | 33  | 117.5 | 0.93 | 6 | 1.00 | 0.00 | 0.00 |
| 3lbx,A| Spectrin     | 140     | 0.6674 | 63* | 50.5* | 0.45 | 2 | 0.79 | 0.21 | 0.00 |
| 1ot8,A| Ank          | 209     | 0.6578 | 33* | 157.5* | 0.89 | 6 | 0.94 | 0.00 | 0.06 |
| 1plq@1| Quaternary   | 774     | 0.6515 | 132*| 190,A*| 0.92 | 6 | 0.95 | 0.05 | 0.00 |
| 1tr4,A| Ank          | 226     | 0.6488 | 33  | 55.5  | 0.90 | 7 | 0.98 | 0.00 | 0.02 |
| 1blx,B| Ank          | 160     | 0.6465 | 65  | 76.5  | 0.86 | 3 | 0.90 | 0.00 | 0.10 |
| 3ow8,A| $\beta$-Propeller | 300 | 0.6317 | 42  | 200  | 0.93 | 7 | 0.99 | 0.01 | 0.00 |
| PDB Code | Type         | Structure | AccessNumber | RMSD | M. | T. | R. | R.| R. |
|----------|--------------|-----------|--------------|------|----|----|----|---|----|
| 1h4a,X   | β-γ-crystallin | Hevein    | 174          | 0.6226 | 43 | 149.5 | 0.87 | 4  | 0.95 | 0.03 | 0.02 |
| 1k7u,A   | Hevein       | Ank       | 171          | 0.6225 | 43 | 150.5 | 0.99 | 4  | 1.00 | 0.00 | 0.00 |
| 1bu9,A   | Hevein       | Ank       | 168          | 0.6200 | 33* | 87.5* | 0.88 | 5  | 0.97 | 0.00 | 0.03 |
| 2vj3,A   | EFG-Like     | Hevein    | 120          | 0.6155 | 38* | 453* | 0.93 | 4  | 1.00 | 0.00 | 0.00 |
| 3smr,A   | β-Propeller  | Hevein    | 304          | 0.5969 | 42 | 138   | 0.92 | 7  | 0.99 | 0.00 | 0.01 |
| 1ybi,A   | Hevein       | Ank       | 284          | 0.5890 | 142 | 223   | 0.86 | 2  | 0.97 | 0.02 | 0.01 |
| 1vpk@1   | Quaternary   | Hevein    | 734          | 0.5856 | 128* | 297   | 0.78 | 6  | 0.93 | 0.07 | 0.00 |
| 1r8p@0   | Quaternary   | Hevein    | 162          | 0.5800 | 43* | 58.5* | 0.78 | 4  | 1.00 | 0.00 | 0.00 |
| 4atg,A   | Hevein       | Ig-Like   | 195          | 0.5788 | 87* | 277.5* | 0.49 | 2  | 0.83 | 0.00 | 0.17 |
| 2rik,A   | Hevein       | Ig-Prop-   | 280          | 0.5760 | 94  | 140   | 0.94 | 3  | 1.00 | 0.00 | 0.00 |
| 2hhb@1   | Quaternary   | Hevein    | 574          | 0.5700 | 21* | 66.5* | 0.76 | 24 | 0.84 | 0.12 | 0.04 |
| 2afg,A   | Hevein       | Ank       | 129          | 0.5586 | 41  | 30.5  | 0.85 | 3  | 0.98 | 0.02 | 0.00 |
| 1d9s,A   | Hevein       | Ank       | 130          | 0.5576 | 66* | 30*   | 0.71 | 2  | 0.98 | 0.00 | 0.02 |
| 1ap7,A   | Hevein       | Ank       | 168          | 0.5469 | 32  | 114   | 0.78 | 5  | 0.93 | 0.02 | 0.05 |
| 1ikn,D   | Hevein       | Ank       | 220          | 0.5540 | 106* | 128*  | 0.59 | 2  | 0.92 | 0.05 | 0.03 |
| 1b3u,A   | Hevein       | Heat      | 588          | 0.5488 | 39  | 530.5 | 0.87 | 15 | 0.98 | 0.02 | 0.00 |
| 1jt4,A   | Hevein       | Trefoil   | 137          | 0.5169 | 41* | 30.5* | 0.77 | 3  | 0.92 | 0.01 | 0.07 |
| 2pnn,A   | Hevein       | Ank       | 248          | 0.5026 | 11  | 135.5 | 0.65 | 17 | 0.75 | 0.23 | 0.02 |
| 1kmn,A   | Hevein       | Trefoil   | 129          | 0.4975 | 40* | 27*   | 0.84 | 3  | 0.91 | 0.05 | 0.04 |
| 3jxi,A   | Hevein       | Ank       | 253          | 0.4927 | 11  | 286.5 | 0.63 | 19 | 0.81 | 0.17 | 0.02 |
| 1fq0,A   | Hevein       | TIM       | 213          | 0.4909 | 88  | 60    | 0.50 | 2  | 0.89 | 0.03 | 0.08 |
| 1hxn,A   | Hevein       | β-Propeller | 210      | 0.4909 | 9*  | 247.5* | 0.66 | 19 | 0.80 | 0.18 | 0.02 |
| 1t12,A   | Hevein       | β-Propeller | 235      | 0.4851 | 47* | 119.5* | 0.99 | 5  | 1.00 | 0.00 | 0.00 |
| 1rpx,A   | Hevein       | TIM       | 230          | 0.4838 | 109 | 171.5 | 0.60 | 2  | 0.94 | 0.00 | 0.06 |
| 1qo2,A   | Hevein       | TIM       | 241          | 0.4810 | 15* | 42.5* | 0.53 | 11 | 0.71 | 0.22 | 0.07 |
| 1s70,B   | Hevein       | Ank       | 291          | 0.4776 | 33* | 86.5* | 0.66 | 7  | 0.77 | 0.10 | 0.13 |
| 1thf,D   | Hevein       | TIM       | 253          | 0.4707 | 121* | 179.5* | 0.64 | 2  | 0.94 | 0.00 | 0.06 |
| 2a0n,A   | Hevein       | TIM       | 251          | 0.4697 | 121* | 181.5* | 0.67 | 2  | 0.96 | 0.00 | 0.04 |
| 1s4u,X   | Hevein       | β-Propeller | 390      | 0.4612 | 11  | 330.5 | 0.60 | 28 | 0.79 | 0.18 | 0.03 |
| 3pg0,A   | Hevein       | Trefoil   | 140          | 0.4605 | 47* | 25.5* | 1.00 | 3  | 1.00 | 0.00 | 0.00 |
| 3ams,A   | Hevein       | β-Propeller | 352      | 0.4582 | 167* | 244.5* | 0.50 | 2  | 0.92 | 0.00 | 0.08 |
| PDB  | Protein  | Length | Len  | Width | Max Fl | Cent  | Cover (%) | Insert (% | Non-R | Max Fl | Coverage (%) |
|------|----------|--------|------|-------|--------|-------|-----------|-----------|--------|--------|--------------|
| 2v70  | Leucine  | 210    | 24   | 628   | 0.65   | 7     | 0.83      | 0.09      | 0.08   |        |              |
| 2a62  | Cadherin | 322    | 109  | 167.5 | 0.77   | 3     | 0.98      | 0.00      | 0.02   |        |              |
| 1s2w  | TIM      | 275    | 124* | 179*  | 0.42   | 2     | 0.81      | 0.00      | 0.19   |        |              |
| 5tim  | TIM      | 249    | 23   | 115.5 | 0.47   | 8     | 0.76      | 0.19      | 0.05   |        |              |
| 3kea  | Ank      | 282    | 99   | 134.5 | 0.34   | 2     | 0.61      | 0.00      | 0.39   |        |              |
| 8tim  | TIM      | 247    | 23   | 116.5 | 0.45   | 7     | 0.67      | 0.21      | 0.12   |        |              |
| 1dx5  | EGF-Like | 118    | 13*  | 442.5*| 0.57   | 7     | 0.77      | 0.23      | 0.00   |        |              |
| 3ehq  | Ank      | 182    | 67*  | 93.5* | 0.42   | 2     | 0.72      | 0.00      | 0.28   |        |              |
| 2aja  | Ank      | 347    | 11   | 226.5 | 0.81   | 28    | 0.88      | 0.11      | 0.01   |        |              |
| 1mbd  | Globin   | 153    | 7    | 12.5  | 0.79   | 19    | 0.87      | 0.11      | 0.02   |        |              |
| 1ycs  | Ank      | 193    | 33*  | 400.5*| 0.51   | 4     | 0.65      | 0.00      | 0.35   |        |              |
| 1gfl  | GFP      | 230    | 7    | 92.5  | 0.72   | 26    | 0.78      | 0.21      | 0.01   |        |              |
| 1sw6  | Ank      | 301    | 7    | 485.5 | 0.58   | 29    | 0.67      | 0.31      | 0.02   |        |              |
| 1nar  | TIM      | 289    | 11   | 153.5 | 0.53   | 20    | 0.75      | 0.22      | 0.03   |        |              |
| 1vju  | TIM      | 325    | 13   | 63.5  | 0.46   | 16    | 0.66      | 0.33      | 0.01   |        |              |
| 2iho  | Trefoil  | 292    | 51*  | 82.5* | 0.39   | 3     | 0.53      | 0.00      | 0.47   |        |              |
| 1dcq  | Ank      | 276    | 7    | 388.5 | 0.64   | 32    | 0.81      | 0.18      | 0.01   |        |              |
| 3phz  | Trefoil  | 285    | 13   | 85.5  | 0.51   | 17    | 0.77      | 0.22      | 0.01   |        |              |
| 1bfl  | β-Propeller | 367 | 13  | 386.5 | 0.49   | 19    | 0.71      | 0.25      | 0.04   |        |              |
| 1hx1  | Globular | 377    | 7    | 216.5 | 0.62   | 44    | 0.81      | 0.19      | 0.00   |        |              |
| 2vg3  | Globular | 284    | 11   | 198.5 | 0.52   | 18    | 0.69      | 0.22      | 0.09   |        |              |
| 4blm  | Globular | 261    | 15   | 185.5 | 0.52   | 13    | 0.74      | 0.25      | 0.01   |        |              |
| 3dtm  | Globular | 263    | 13   | 185.5 | 0.53   | 15    | 0.73      | 0.26      | 0.01   |        |              |
| 2psg  | Globular | 326    | 15   | 270.5 | 0.50   | 17    | 0.77      | 0.22      | 0.01   |        |              |

*a* Protein length. *b* Tileability Score. *c* Tile center. *d* Tile Score. *e* Number of tessellated tile copies. *f* Fraction of the coverage given by insertions between tiles. *g* Fraction of the coverage of the Non-Repeating regions. *h* Fraction of the coverage accounted by tile copies. *i* Proteins for which there is more than one tile that have identical max Θ, at the characteristic frequency. *(†)* Protein Complexes, the letter after Z indicates the chain ID.