**New Facets of Larger Nest Motifs in Proteins**

Debnath Pal,a\* Subhankar Sahub and Raja Banerjeeb,c\*

aDepartment of Computational and Data Sciences,

Indian Institute of Science, Bengaluru 560 012, Karnataka, India

bDepartment of Bioinformatics, cDepartment of Biotechnology

Maulana Abul Kalam Azad University of Technology, Kolkata 700 064, West Bengal, India

\*Correspondence Email: dpal@iisc.ac.in, Tel: +918022932901, FAX: +918023606332

Co-correspondence Email: banraja10@gmail.com

**Table SI** | Top twenty ranked dipeptides by occurrence frequency and their propensity (both given in parenthesis) in Nest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| All | RL | LR | Simple | Larger |
| DG (3166/1.27) | DG (3127/1.13) | GL (1161/0.99) | LG (2401/1.11) | DG (920/**1.59**) |
| LG (2911/1.23) | LG (2891/0.96) | GK (1044/1.18) | DG (2246/1.18) | GL (779/**1.62**) |
| AG (2579/1.4) | AG (2551/1.2) | GA (771/0.97) | AG (2037/1.3) | SG (701/**1.76**) |
| TG (2523/1.37) | TG (2515/1.02) | GT (735/1) | TG (1974/1.23) | GK (683/1.45) |
| SG (2355/1.33) | SG (2342/1.08) | GS (683/0.95) | SG (1654/1.2) | TG (549/**1.79**) |
| NG (2001/0.89) | NG (1934/0.93) | GV (682/1.05) | NG (1497/0.87) | AG (542/**1.62**) |
| EG (1428/1.12) | EG (1408/0.99) | GE (670/1.09) | KG (1088/1.09) | LG (510/1.49) |
| KG (1419/1.06) | KG (1392/0.99) | GR (601/1.06) | EG (1059/1.05) | NG (504/0.96) |
| RG (1314/1.1) | RG (1289/1.01) | GI (517/0.99) | RG (1024/1.1) | GE (426/1.45) |
| GL (1190/**2.35**) | QG (1099/1.03) | GQ (460/0.96) | QG (899/1.17) | GT (425/**1.7**) |
| QG (1119/1.17) | FG (918/0.86) | GY (459/1.25) | FG (804/0.93) | GA (393/1.38) |
| GK (1106/**1.51**) | YG (887/0.87) | GD (447/1.06) | YG (787/0.97) | GS (387/1.38) |
| FG (934/0.99) | VG (717/1.04) | GF (433/1.07) | HG (611/0.89) | GV (387/**1.97**) |
| YG (914/1.04) | HG (715/0.84) | GG (351/0.92) | VG (579/1.24) | GR (383/1.22) |
| GG (870/**0.18**) | IG (671/1.05) | GN (338/0.94) | GG (566/**0.23**) | EG (369/1.31) |
| GA (798/**1.88**) | *LN* (524/1.16) | *NL* (253/1.24) | IG (537/1.28) | KG (331/1.02) |
| GT (744/**2.71**) | GG (519/0.89) | GH (240/0.99) | *LN* (443/1.31) | GD (314/1.06) |
| HG (738/0.91) | MG (516/1.03) | GM (191/1.03) | MG (428/1.16) | GG (304/**0.17**) |
| VG (721/1.36) | CG (412/0.8) | GW (191/1.09) | GK (423/1.41) | GI (298/**1.85**) |
| GS (712/**1.71**) | NN (410/1.31) | NA (155/1.12) | GL (411/**2.81**) | RG (290/1.09) |

Propensity values > 1.5 are in bold, and < 0.5 are in underlined bold. Non-gly peptides are italicized.

**Table SII** | Top twenty ranked secondary structures found in dipeptides by occurrence frequency and their propensity (both given in parenthesis) in Nest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| All | RL | LR | Simple | Larger |
| Tt (13740/**2.72**) | hT (11254/**2.74**) | Tt (4457/**2.11**) | Tt (9236/**2.70**) | Tt (4504/**2.78**) |
| hT (11358/**2.53**) | Tt (9283/**3.08**) | hH (2794/**5.20**) | hT (8239/**2.50**) | hT (3119/**2.60**) |
| Hh (8518/**6.17**) | Hh (8483/**4.70**) | TT (1827/1.46) | Hh (7570/**5.33**) | SS (1457/**4.04**) |
| TT (4680/0.79) | TT (2853/0.64) | SS (1015/**4.69**) | TT (3858/0.84) | hH (1159/**3.87**) |
| hH (2802/**3.90**) | Gg (1727/**1.89**) | tT (981/**2.54**) | Gg (1663/**20.1**) | Hh (948/**9.09**) |
| CC (2200/**5.93**) | CC (1608/**6.50**) | gG (934/**15.2**) | hH (1643/**3.91**) | TT (822/0.59) |
| SS (2114/**6.41**) | CS (1183/**5.33**) | Te (813/1.39) | Te (1456/**1.80**) | CC (778/**3.99**) |
| Gg (1949/**20.9**) | SS (1099/**7.58**) | SC (604/**2.94**) | CC (1422/**7.54**) | tC (694/**2.62**) |
| Te (1895/**1.69**) | Te (1082/**1.86**) | CC (592/**4.96**) | CS (882/**7.59**) | CS (646/**2.32**) |
| CS (1528/**4.31**) | SC (691/**4.28**) | tS (503/**1.97**) | SC (793/**6.25**) | tT (622/1.10) |
| tT (1383/1.29) | tC (571/**5.31**) | tC (480/**1.99**) | tT (761/**1.95**) | tS (616/**1.62**) |
| SC (1295/**3.75**) | eS (406/**7.78**) | Th (415/**1.35**) | SS (657/**8.41**) | SC (502/**1.99**) |
| tC (1051/**3.34**) | tT (402/0.85) | CS (345/**2.74**) | gG (634/**49.8**) | Te (439/1.39) |
| gG (964/**41.7**) | he (315/0.59) | ht (339/0.34) | tC (357/**3.92**) | Th (420/0.60) |
| tS (705/**2.34**) | ht (293/0.11) | te (301/**1.67**) | he (314/0.54) | ht (413/0.30) |
| Th (666/0.19) | Th (251/0.09) | Gg (222/**27.1**) | eC (268/**4.78**) | gG (330/**31.2**) |
| ht (632/0.17) | hh (245/0.09) | Tg (178/0.77) | Th (246/0.09) | te (323/**2.54**) |
| eS (450/**4.90**) | Ct (219/0.31) | Se (134/0.88) | EE (228/**176**) | Gg (286/**22.3**) |
| he (370/0.44) | eC (210/3.61) | eT (118/**1.89**) | Ge (224/**2.16**) | hh (264/0.44) |
| te (350/**1.73**) | tS (202/**2.09**) | Sh (113/1.40) | ht (219/0.09) | eS (257/**4.47**) |

Propensity values > 1.5 are in bold, and < 0.5 are in underlined bold.



**Figure S1** | A color gradient tile showing the propensity values of the amino acids in position 1 and 2 in the Nest with respect to various datasets. All: All dipeptides in the data set including Nests; Nest\_All: Only Nest Dipeptides having either RL or LR enantiomeric conformational states; Nest\_RL: Dipeptides in Nest having only RL conformational state; Nest\_LR: Dipeptides in Nest having only LR conformational state; Simple Nest: Isolated dipeptides having either RL or LR conformational states; Larger Nest: Nests other than simple Nests. Here \_RL or \_LR data set means a subset screened to contain only RL or LR conformation dipeptides in them. Propensity values between 0.5 and 1.5 are uniformly colored by a deep blue color corresponding to 1 in the legend. Other values above 1.5 and below 0.5 are as per the legend.