# The Nest/Egg−Motif in Proteins

Jessica Meinhardt,[a] Debnath Pal,[a,b] Jürgen Sühnel,[a] Manfred S. Weiss [c]
dpal@mbi.ucla.edu, jsuehnel@imb−jena.de, msweiss@embl−hamburg.de

a) Institute of Molecular Biotechnology, Beutenbergstr. 11, D−07745 Jena / Germany
b) New address: UCLA−DOE Laboratory for Structural Biology & Molecular Medicine, 105 Molecular Biology Institute, BOX 951570, University of California, Los Angeles, CA 90095−1570, USA
c) EMBL Hamburg Outstation, c/o DESY, Notkestr. 85, D−22603 Hamburg / Germany

## Introduction

The main−chain conformations of protein chains can be conveniently analyzed by means of the Ramachandran map displaying the distribution of amino acid backbone torsional angles ϕ and ψ [1]. The interpretation of Ramachandran maps usually focuses on clusters for specific secondary structure elements such as β−sheets and different helix types. In these cases the main−chain conformations of successive amino acid pairs are identical or at least similar. Recently, Watson and Milner−White have discovered that many anion and cation binding sites (where anions and cations can be any atoms exhibiting a full or a partial negative and positive charge, respectively) in proteins are made up by a sequence of three amino acids of which two exhibit "enantiomeric" main−chain conformations [2,3]. The term "enantiomeric" refers to the fact that, contrary to the conformations in β−sheets and helices, the main−chain torsional angles (ϕ,ψ) of the two adjacent amino acids are approximately inverted about the center of the Ramachandran map. The authors have called this motif a nest because three successive residues obeying this torsional angle criterion form a concave depression which can serve as a binding site for an atom or a group of atoms with a full or partial negative charge. In the majority of cases, the nests do bind to an atom or a group of atoms, which we suggest may, as a binding partner of a "nest", be descriptively and conveniently called "egg" [4]. It is intriguing that many structural motifs described previously, such as Schellman loops, the oxyanion holes of serine proteases, P−loops in ATP− or GTP−binding proteins can be subsumed under this nest/egg−concept. If dipeptides with different enantiomeric main−chain torsional angle combinations are considered, the nest/egg−concept can even be extended to cation binding sites although the experimental evidence for this is not as ample as for the anion−binding nests [2,3].Watson and Milner−White conducted their analysis on a limited data base of 67 protein structures. We have performed an extended analysis of their concept in a database about 20 times larger [4] (the same database as the one used in reference [5]).

## Method

The database used contained 1,154 protein chains, 285,794 amino acids and 280,563 dipeptides (terminal residues were excluded from the list of dipeptides, because only one main chain torsional angle can be defined for a terminal amino acid residue). The R and L conformations of amino acids are defined by R: $-140^\circ < \phi < -20^\circ$; $-90^\circ < \psi < +40^\circ$; L: $+20^\circ < \phi < +140^\circ$; $-40^\circ < \psi < +90^\circ$. RL and LR nests are formed if two successive amino acids adopt an R and L or L and R conformation. The peptide bond in between the two amino acids is in both RL and LR nests always in the *trans*−conformation. The propensities $P^{RL}_{ij,rs}$ and $P^{LR}_{ij,rs}$ measure the bias of finding a dipeptide *ij* with a combination of secondary states *rs* in an RL or LR nest cluster as compared to the average occurrence in the total database. They are computed from the ratio of the relative frequency of occurrence $F_{ij,rs}$ in the RL or LR data sets and in the total database. Only dipeptides that contribute more than 1% to the total number of RL or LR nests are taken into account in the results.

## Results

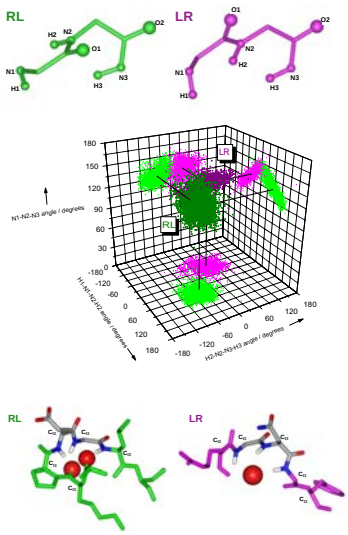


**Figure 1.** Three−dimensional plot of all amino acid pairs obeying the nest criterion. The three axes are: (i) the torsional angle between the four atoms H1−N1−N2−H2, (ii) the torsional angle between the four atoms H2−N2−N3−H3 and (iii) the angle between the three nitrogen atoms N1−N2−N3. The RL−nests are shown in green, the LR−nests in pink. The respective projections are drawn in the lighter colors. The mean values of the distributions of the respective parameters are indicated by solid lines originating from the islands and projecting onto the walls of the cube. Two representative structures corresponding approximately to the centers of the distributions of the geometric parameters for RL and LR are also shown.




**Figure 2.** An example of an RL− and an LR−nest/egg motif as observed in the structure of lysozyme (PDB code: 153L). Depicted are the two nest amino acids and the NH−group of the third amino acid in atom−specific coloring and the flanking amino acids in green (RL) or pink (LR). The egg oxygen atoms are shown as red balls. The RL−nest is comprised by the amino acids Glu24, Gly25 and Leu27 and the respective egg(s) are the carbonyl oxygen atoms of Ala21 and Lys22. The LR−nest consists of Gly88, Asn89 and Gly90 and the respective egg is a water molecule (water 216).

**Table 1.** Occurrences and average geometric parameters of the observed nests.

| | Total | $\phi_1$ | $\psi_1$ | $\phi_2$ | $\psi_2$ | ∢ H1N1N2H2 | ∢ H2N2N3H3 | ∢ N1N2N3 |
|---|---|---|---|---|---|---|---|---|
| RL | 5773 | −91° | −5° | +74° | +23° | +56° | −41° | 123° |
| LR | 2492 | +72° | +21° | −85° | −22° | −44° | +38° | 128° |

**Table 2.** Categories of observed nest/egg−structures.

| | RL | LR |
|---|---|---|
| *Nests* | | |
| Total | 5773 (100%) | 2492 (100%) |
| Occupied | 4419 (76.5%) | 1290 (51.8%) |
| Empty | 1354 (23.5%) | 1202 (48.2%) |
| *Eggs* | | |
| Total[a] | 4838 (100%) | 1386 (100%) |
| Main chain carbonyl−oxygen | 3009 (62.2%) | 560 (40.4%) |
| Side chain carboxylate−oxygen | 630 (13.0%) | 234 (16.9%) |
| Side chain amide−oxygen | 200 (4.1%) | 44 (3.2%) |
| Side chain hydroxyl−oxygen | 795 (16.4%) | 49 (3.5%) |
| Ions [b] | 48 (1.0%) | 55 (4.0%) |
| Water molecules | 156 (3.2%) | 444 (32.0%) |

[a] In accord with Watson and Milner−White, all ligands at a distance of less than 3.8 Å to both N1 and N3 were considered. The total number of eggs is larger than the number of occupied nests, because some nests contain more than one egg.
[b] Possible ions were: phosphate, sulfate, chloride, iron−sulfur−clusters, phosphate−containing cofactors.

**Table 3.** Dipeptide propensities P*ij* and frequencies of occurrence F*ij* for RL− and LR−nests.

| RL | | | LR | | |
|---|---|---|---|---|---|
| ***ij*** | $P_{ij}$ | $F_{ij}$ | ***ij*** | $P_{ij}$ | $F_{ij}$ |
| DG | 14.043 | 382 | GK | 11.521 | 136 |
| TG | 13.251 | 364 | GL | 11.045 | 171 |
| NG | 11.773 | 250 | GR | 10.393 | 96 |
| LG | 10.891 | 398 | GT | 9.558 | 113 |
| SG | 10.615 | 313 | GQ | 9.322 | 65 |
| AG | 10.162 | 377 | GH | 8.696 | 38 |
| MG | 9.591 | 89 | GY | 8.273 | 64 |

**Table 4.** Secondary structure dipeptide propensities P*rs*, frequencies of occurrence F*ij* and propensity ratios for RL− and LR−nests.

| | RL | | LR | | | |
|---|---|---|---|---|---|---|
| ***rs*** | $P_{rs}$ | $F_{rs}$ | $P_{rs}$ | $F_{rs}$ | $pRL_{rs}/pLR_{rs}$ | **preference** |
| Hh | 6.844 | 1146 | 0.055 | 4 | 124.44 | RL |
| hT | 20.951 | 1486 | 0.359 | 11 | 58.36 | RL |
| Gg | 5.441 | 263 | 1.438 | 30 | 3.78 | RL |
| CS | 1.288 | 126 | 1.066 | 45 | 1.21 | |
| CC | 0.601 | 176 | 0.562 | 71 | 1.07 | |
| Tt | 7.102 | 1194 | 8.198 | 595 | 0.87 | |
| TT | 1.621 | 374 | 2.460 | 245 | 0.66 | |
| SS | 1.145 | 129 | 2.386 | 116 | 0.48 | LR |
| Te | 3.695 | 119 | 7.913 | 110 | 0.47 | LR |
| tC | 0.901 | 67 | 2.212 | 71 | 0.41 | LR |
| SC | 0.701 | 68 | 1.887 | 79 | 0.37 | LR |
| ht | 1.029 | 33 | 3.827 | 53 | 0.27 | LR |
| Th | 1.046 | 30 | 4.765 | 59 | 0.22 | LR |
| tT | 0.405 | 49 | 2.337 | 122 | 0.17 | LR |
| tS | 0.775 | 24 | 5.536 | 74 | 0.14 | LR |
| te | 0.122 | 3 | 3.592 | 38 | 0.03 | LR |
| gG | 0.147 | 7 | 6.038 | 124 | 0.02 | LR |
| hH | 0 | 0 | 5.445 | 383 | 0.00 | LR |

| | |
|---|---|
| H | α−helix |
| G | $3_{10}$−helix |
| I | π−helix |
| E | β−strand |
| B | β−bridge |
| S | bend |
| T | turn |
| C | coil |

lower case letters indicate end residues

## Conclusions

- The nest concept unifies a variety of known motifs but also unveils a few novel motifs. It takes ligands, cofactors, water molecules, etc. into account and therefore sets the stage to a general approach to binding sites in proteins.
- 77% RL−nests and 52% LR−nests are occupied by eggs. The preferred egg atoms are main−chain carbonyl oxygens. However, side−chain atoms, ions and water molecules do also occur.
- Glycine is the predominant first amino acid in LR−nests (55%) and second amino acid in RL−nests (61%).
- The nest secondary structure dipeptide propensities cover a range between 0 and 21 and thus indicate a dramatic difference to the average dipeptide and secondary structure distributions.
- A strong preference for RL−nests is found at C−terminal helix ends and at helix/turn interfaces, whereas at N−terminal ends a LR prevalence is observed. Within turn regions both LR and RL nests occur frequently.
- One may also speculate whether nests within the protein context may constitute stable structures very early along the pathway of folding since they are inherently local structures and thus would not require a large entropy reduction upon formation.
- We hope that these new data, when incorporated into protein structure prediction tools, can significantly improve their performance.

## References

[1] G. N. Ramachandran, V. Sasisekharan, *Adv. Prot. Chem*. **1968**, 23, 293−437.
[2] J. D. Watson, E. J. Milner−White, *J. Mol. Biol.* **2002**, 315, 187−198.
[3] J. D. Watson, E. J. Milner−White, *J. Mol. Biol.* **2002**, 315, 199−207.
[4] D. Pal, J. Sühnel, M. S. Weiss, *Angew. Chem. Int. Ed.* **2002,** 41, 4663−4665.
[5] M. Brandl, M. S. Weiss, A. Jabs, J. Sühnel, R. Hilgenfeld, *J. Mol. Biol.* **2001**, 307, 357−377.