ethyl-*N*′-(3-dimethylaminopropyl)carbodiimide hydrochloride; Fmoc: 9-fluorenylmethoxycarbonyl; HATU: O-(7-azabenzotriazol-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate (IUPAC: 1-[bis-(dimethylamino)methyliumyl]-1*H*-1,2,3-triazolo[4,5-*b*]pyridin-3-oxide hexafluorophosphate); HOAt: 1-hydroxy-7-azabenzotriazole; MeBmt: (4*R*)-4[(*E*)-2-butenyl-4,*N*-dimethyl-L-threonine; Nva: norvaline; tBu: *tert*-butyl; TFFH: tetramethylfluoroformamidinium hexafluorophosphate; Trt: triphenylmethyl (trityl).

[1] B. Thern, J. Rudolph, G. Jung, *Tetrahedron Lett.* **2002**, *43*, 5013 – 5016.
[2] B. Thern, J. Rudolph, G. Jung, *Angew. Chem.* **2002**, *114*, 2401 – 2403; *Angew. Chem. Int. Ed.* **2002**, *41*, 2307 – 2309.
[3] N. Sewald, H.-D. Jakubke, *Peptides: Chemistry and Biology*, Wiley-VCH, Weinheim, **2002**.
[4] E. Falb, Y. Yechezkel, Y. Salitra, C. Gilon, *J. Pept. Res.* **1999**, *53*, 507 – 517.
[5] H. Eckert, B. Forster, *Angew. Chem.* **1987**, *99*, 922 – 923; *Angew. Chem. Int. Ed. Engl.* **1987**, *26*, 894 – 895.
[6] L. A. Carpino, M. Beyermann, H. Wenschuh, M. Bienert, *Acc. Chem. Res.* **1996**, *29*, 268 – 274.
[7] L. A. Carpino, A. El-Faham, *Tetrahedron* **1999**, *55*, 6813 – 6830.

New Principles of Protein Structure: Nests, Eggs—and What Next?**

Debnath Pal, Jürgen Sühnel,* and Manfred S. Weiss*

Based on results from refolding experiments on urea-denatured ribonuclease conducted more than 40 years ago, the chemistry nobel laureate Anfinsen formulated the still to a large extent valid paradigm of protein folding: "… *it may be concluded that the information … for the assumption of the native secondary and tertiary structures* (of proteins) *is contained in the amino acid sequence itself.*"[1] As a direct consequence of this, one has to postulate that it should be possible to predict the native structure of a protein from the protein's amino acid sequence alone. However, despite much work of many excellent scientists and a database of experimentally determined protein structures that is increasing frighteningly fast on a daily basis,[2] successes in protein structure prediction are scarce, and the current situation is rather disappointing. The reasons for this are not entirely clear. A large body of experimental information has become available with the boost structural biology has experienced in the last decade and much effort has been put into the thorough analysis of these data over the years,[3–5] but, with the predictive power largely lacking, current knowledge of the basic principles of protein structure is still mainly descriptive. One explanation may be that currently known structural principles do not disclose the complete picture and that new concepts and new ideas are necessary to propel the field from the mainly descriptive into a more predictive mode.

[*] Dr. J. Sühnel, Dr. D. Pal
Institut für Molekulare Biotechnologie
Beutenbergstrasse 11, 07745 Jena (Germany)
Fax: (+49) 3641-656210
E-mail: jsuehnel@imb-jena.de
Dr. M. S. Weiss
EMBL Hamburg Outstation
c/o DESY, Notkestrasse 85, 22603 Hamburg (Germany)
Fax: (+49) 40-89902-149
E-mail: msweiss@embl-hamburg.de

[**] The authors are grateful to E. James Milner-White for introducing them to his nest concept and for stimulating discussions.

In this respect, two interesting papers have been published recently in the *Journal of Molecular Biology*.[6, 7] Based on the analysis of main-chain torsion angles of adjacent amino acid residues, Watson and Milner-White 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. The term "enantiomeric" refers to the fact that the main-chain torsion angles (ϕ,ψ) of the two adjacent amino acids are approximately inverted about the center of the Ramachandran plot.[3] Whereas successive residues with identical or nearly identical main-chain conformations form α helices, β strands, or polyproline type II helices, adjacent residues with enantiomeric main-chain conformations form so-called "nests" when their (ϕ,ψ)-values are close to ($-90°$,$0°$) and ($+90°$,$0°$) or the other way round.

The term "nest" is derived from the fact that the NH groups of three successive residues obeying this torsion 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. Depending on which combination of the two torsion angle pairs is observed, the nests can be divided into RL nests ($\phi_1,\psi_1 = -90°,0°$; $\phi_2,\psi_2 = +90°,0°$) and LR nests ($\phi_1,\psi_1 = +90°,0°$; $\phi_2,\psi_2 = -90°,0°$). Two or more of these nests can also constitute a compound nest, a tandem nest, or a combination of both with up to eight successive residues involved. In the majority of cases the nests 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 an "egg". It is intriguing that many structural motifs described previously, such as Schellman loops, the oxyanion holes of serine proteases, and P loops in ATP- or GTP-binding proteins can be subsumed under this nest/egg concept. If dipeptides with different enantiomeric main-chain torsion angle combinations are considered, the nest/egg concept can

  1433-7851/02/4124-4663 $ 20.00+.50/0

even be extended to cation binding sites,[7] although the experimental evidence for this is not as ample as for the anion-binding nests.[6]

Watson and Milner-White[6, 7] conducted their analysis on a limited data base of 67 protein structures. An extended analysis of their concept in a database about 20 times larger, which had been assembled previously for a different purpose,[8] confirms most of their conclusions and also adds new aspects. As introduced above, nests can be described by the main-chain torsion angles of the two successive amino acids. However, a more convenient description utilizes the two torsion angles involving the NH groups of two successive amino acids and the angle formed by the three nitrogen atoms (Figure 1). Unlike in the Ramachandran plot, where the R and L regions of the RL and LR nests naturally overlap, the two nest classes (RL and LR) form islands clearly separated from each other in the three-dimensional plot shown in Figure 1. The average values of the geometrical parameters are given in Table 1. Two representative structural models of an RL and an LR nest are drawn in Figure 1 and clearly show the concave arrangement of the three NH groups in both cases. Two actual cases are shown in Figure 2.

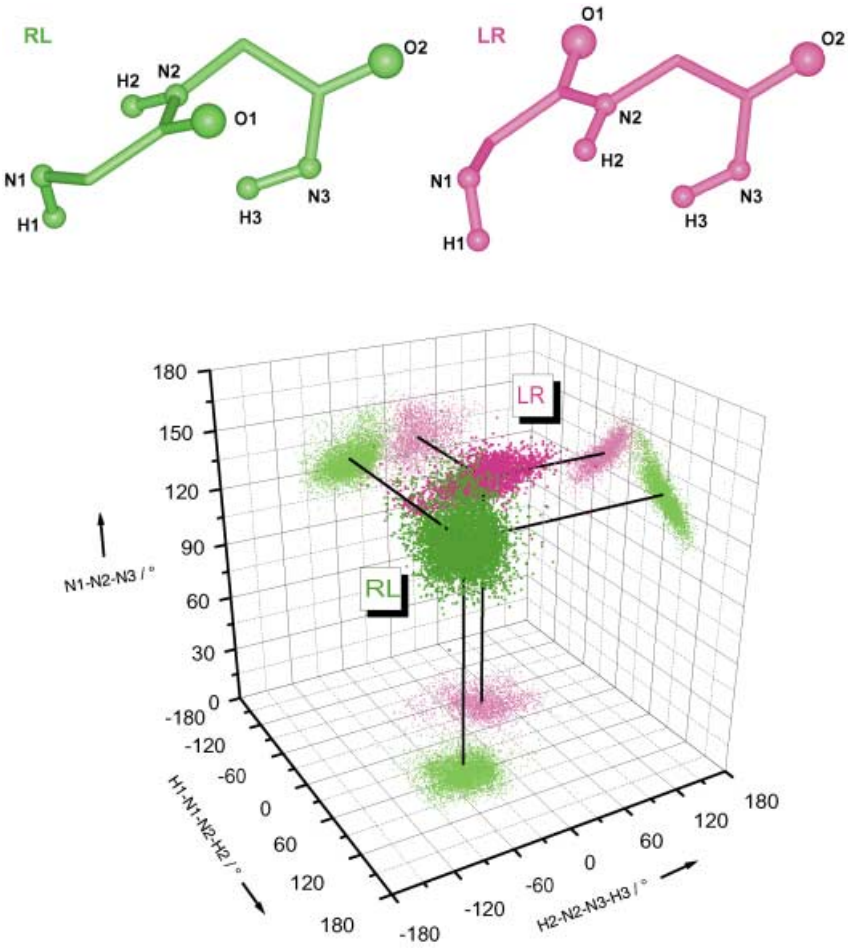

Figure 1. Three-dimensional plot of all amino acid pairs obeying the nest criterion. The three axes are: a) the torsion angle between the four atoms H1-N1-N2-H2; b) the torsion angle between the four atoms H2-N2-N3-H3; and c) the angle between the three nitrogen atoms N1-N2-N3. The RL nests are shown in green and the LR nests in pink. The respective projections are drawn in lighter colors. The mean values of the distributions of the respective parameters above are indicated by solid black lines originating from the islands and projecting onto the three walls of the coordinate system. Two representative structures corresponding approximately to the centers of the distributions of the geometric parameters for the RL and LR motifs are also shown.

Frequencies of nest occurrence in protein structures are given in Table 1 together with the average geometric parameters. Overall, 2.1% of all dipeptides in the database constitute RL nests and about 1.0% LR nests. This brings the total involvement of amino acid residues in nests up to about 5.4%. The next striking observation is the fact that the number of RL nests observed is about double that of LR nests. This effect has already been noted by Watson and Milner-White,[6] although in their case the observed ratio of 4:1 was almost certainly a consequence of the database being too small. At this point, we do not have a good explanation for this prevalence of RL over LR, but the observation seems to imply that the main-chain conformation of amino acids in a polypeptide chain is to some extent determined by the identity of the neighboring residues as well as by their location with respect to chain direction.

Since the L conformation is energetically unfavorable for all amino acids except glycine, it is clear that the majority of the RL nests have to be, and indeed are, of the type Xaa-Gly (61%) and the majority of the LR nests of the type Gly-Xaa (55%). In total, 21% of all glycine residues in proteins are involved in either an RL or an LR nest. This may provide an explanation for the fact that glycine residues are often better conserved than other amino acids.

Nests occur preferentially at the ends of α helices and reverse turns. It is clear from the torsion angle criterion that RL nests are more frequently found at the C-terminal ends of

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

	Total[a]	ϕ_1	ψ_1	ϕ_2	ψ_2	$\sphericalangle H_1N_1N_2H_2$	$\sphericalangle H_2N_2N_3H_3$	$\sphericalangle N_1N_2N_3$
RL[b]	5773	-91°	-5°	$+74^\circ$	$+23^\circ$	$+56^\circ$	-41°	123°
LR[b]	2492	$+72^\circ$	$+21^\circ$	-85°	-22°	-44°	$+38^\circ$	128°

[a] The data base used contained 1154 protein chains, 285 794 amino acids, and 280 563 dipeptides (terminal residues were excluded from the list of dipeptides, because only one main-chain torsion angle can be defined for a terminal amino acid residue).[8] [b] R: $-140^\circ < \phi < -20^\circ$; $-90^\circ < \psi < +40^\circ$; L: $+20^\circ < \phi < +140^\circ$; $-40^\circ < \psi < +90^\circ$; the peptide bond in between the two amino acids is always in the *trans* conformation in both the RL and LR nests. The average values for ϕ and ψ angles of the R and the L region are very similar to the values determined by Watson and Milner-White[6] from their smaller data base.

  1433-7851/02/4124-4664 $ 20.00+.50/0 *Angew. Chem. Int. Ed.* **2002**, *41*, No. 24

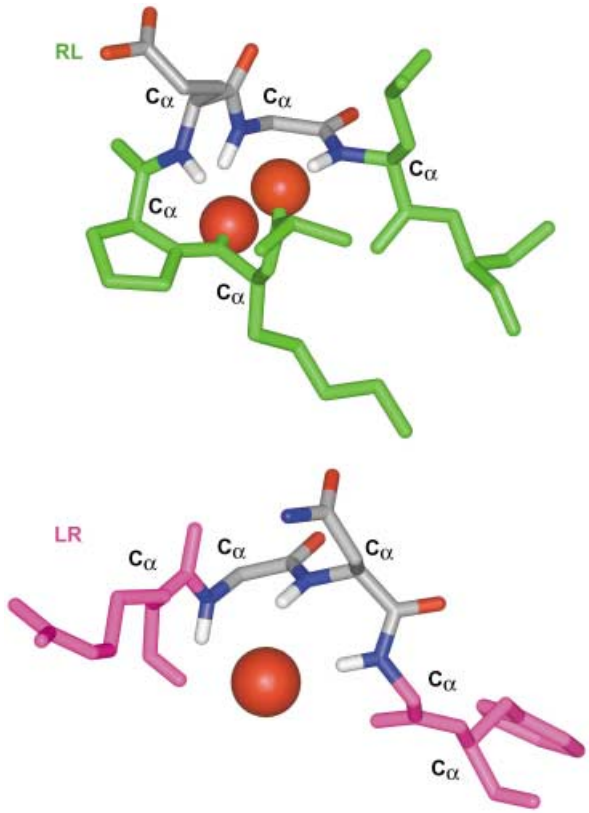

Figure 2. An example of an RL and an LR nest/egg motif as observed in the structure of lysozyme (PDB code: 153L).[10] The two nest amino acids and the NH group of the third amino acid in atom-specific coloring are depicted as well as the flanking amino acids in green (RL) or pink (LR). The oxygen atoms of the egg are shown as red balls. The RL nest is comprised of the amino acids Glu 24, Gly 25, and Leu 27 and the respective egg(s) are the carbonyl oxygen atoms of Ala 21 and Lys 22. The LR nest consists of Gly 88, Asn 89, and Gly 90 and the respective egg is a water molecule (water 216).

α helices, whereas LR nests are more often located at the N-terminal ends. No such preference is observed for typical turns. Also, we have found no occasion where a nest is actually coincident with either a type II or a type II′ turn.[9]

Another important issue is raised by the question: how many nests contain an egg, that is, serve as binding sites, or according to Watson and Milner-White[6] are functional? The answer to this is that 77% of all RL nests and 52% of all LR nests contain a ligand atom bound at hydrogen-bonding distance (Table 2). In most cases, main-chain carbonyl oxygen atoms of amino acids from different parts of the chain constitute the egg. This situation is different when the eggs are formed from side-chain atoms. In the case of Ser-OG and Thr-OG1 atoms as eggs, more than half of them belong to the nest amino acids themselves. Apart from amino acid atoms serving as eggs, many water molecules and ions are also observed. In this respect it is interesting to note that only about 3% of all RL nests bind water, whereas 18% of all LR nests do (Table 2). These observations suggest a general role of nests as binding sites. Also, it may be that the nests together with their respective eggs and not just the nests alone have to be considered as structural motifs.

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:[a]		
total[b]	4838 (100%)	1386 (100%)
main-chain carbonyl O atom	3009 (62.2%)	560 (40.4%)
side-chain carboxylate O atom	630 (13.0%)	234 (16.9%)
side-chain amide O atom	200 (4.1%)	44 (3.2%)
side-chain hydroxy O atom	795 (16.4%)	49 (3.5%)
ions[c]	48 (1.0%)	55 (4.0%)
water molecules	156 (3.2%)	444 (32.0%)

[a] In accordance with Watson and Milner-White,[6] all ligands at a distance of less than 3.8 Å to both N1 and N3 were considered. [b] The total number of eggs is larger than the number of occupied nests, because some nests contain more than one egg. [c] Possible ions were: phosphate, sulfate, chloride, iron-sulfur clusters, phosphate-containing cofactors. For examples, see Table 5 and Figure 5 of ref. [6].

In summary, the intriguingly simple concept presented by Watson and Milner-White appears to be a valuable contribution to the principles of protein structure, since it not only unifies a variety of observed motifs but also unveils a few novel motifs. The concept is not limited to just the protein itself, but also takes ligands, cofactors, water molecules, etc. into account, and therefore sets the stage for a general approach to binding sites in proteins. 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 result in a large reduction in entropy upon formation. It is to be hoped that concepts such as this one, and possibly others still to come, will further our knowledge about protein structures in general and help move the field from structure description to structure prediction.

[1] C. B. Anfinsen, E. Haber, M. Sela, F. H. White, Jr., *Proc. Natl. Acad. Sci. USA* **1961**, *47*, 1309–1314.
[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–242.
[3] G. N. Ramachandran, V. Sasisekharan, *Adv. Protein Chem.* **1968**, *23*, 293–437.
[4] G. E. Schulz, R. H. Schirmer, *Principles of Protein Structure*, Springer, New York, **1979**.
[5] P. Chakrabarti, D. Pal, *Prog. Biophys. Mol. Biol.* **2001**, *76*, 1–102.
[6] J. D. Watson, E. J. Milner-White, *J. Mol. Biol.* **2002**, *315*, 187–198.
[7] J. D. Watson, E. J. Milner-White, *J. Mol. Biol.* **2002**, *315*, 199–207.
[8] M. Brandl, M. S. Weiss, A. Jabs, J. Sühnel, R. Hilgenfeld, *J. Mol. Biol.* **2001**, *307*, 357–377.
[9] T. E. Creighton, *Proteins: Structure and Molecular Properties*, W. H. Freeman, New York, **1993**.
[10] L. H. Weaver, M. G. Grütter, B. W. Matthews, *J. Mol. Biol.* **1995**, *245*, 54–68.

  1433-7851/02/4124-4665 $ 20.00+.50/0