# Magic Numbers in Protein Structures

**Lindgård, Per-Anker; Bohr, Henrik**

*Published in:*
Physical Review Letters

*Link to article, DOI:*
10.1103/PhysRevLett.77.779

*Publication date:*
1996

*Document Version*
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

*Citation (APA):*
Lindgård, P-A., & Bohr, H. (1996). Magic Numbers in Protein Structures. Physical Review Letters, 77(4), 779-782. DOI: 10.1103/PhysRevLett.77.779

# Magic Numbers in Protein Structures

Per-Anker Lindgård[1] and Henrik Bohr[2]
[1]*Department of Condensed Matter Physics, Risø National Laboratory, DK-4000 Roskilde, Denmark*
[2]*Center for Biological Sequence Analysis, Department of Physical Chemistry, The Technical University of Denmark, DK-2800 Lyngby, Denmark*

(Received 23 January 1996)

A homology measure for protein fold classes has been constructed by locally projecting consecutive secondary structures onto a lattice. Taking into account hydrophobic forces we have found a mechanism for formation of domains containing magic numbers of secondary structures and multipla of these domains. We have performed a statistical analysis of available protein structures and found agreement with the predicted preferred abundances. Furthermore, a connection between sequence information and fold classes is established in terms of hinge forces between the structural elements. [S0031-9007(96)00734-X]

PACS numbers: 87.10.+e, 05.50.+q, 05.70.Ln

Since the appearance of the first solved protein structures by x-ray diffraction and up to the present time with large databases of high resolution protein data, there has been a scientific endeavor to find a taxonomy that could group protein structures. We here propose a new framework for a structural classification. From an analysis of packing under the influence of hydrophobic forces [1] we find preferred abundance of proteins with "magic numbers" of secondary elements—and we test the paradigm by a statistical analysis of available structural data.

Proteins are interesting polymers that in aqueous solutions form dense globula, which neither dissolve nor phase separate, as emphasized by Dill [2], who derived the thermodynamic theory for these. A main reason for this is the action of the *hydrophobic* or *hydrophillic* force, which is an unspecific interface-tension-like force [1]. Yet a protein with a specific amino acid chain folds, paradoxically [3] in a matter of seconds, to a particular fold, according to information which must be provided via the underlying linear sequence information. A concise review is given by Wolynes [4] in which the folding problem has been related to the spin glass problem, marginal stability, and minimal frustration. Another problem is why proteins seem to have predominant lengths of chains [5] and separate into subunits, secondary structures [6], domains, and finally the functional tertiary or quaternary structures. Berman *et al.* [5] found characteristic peaks in the length distribution of known proteins near multipla of chain lengths of 125 amino acids (residues). The total length may go up to a couple of thousand. A few hundred different structures have so far been determined (in crystalline form). Yet many thousands of proteins have had their sequence determined. It is of great interest to attempt to classify the possible structures.

To introduce a model we will start by looking at the final, known structures. These consist of secondary structures [6] of principally $\alpha$ helices (spiral, stiff subunits of around ten residues) and $\beta$ sheets, which are semiplanar collections of a number of almost straight chain elements ($\beta$ strands of around six residues). The elements are connected with more irregular loops. The structures are twisted and deformed in a characteristic biological way. The first problem which arises is the homology problem: how to define when two protein structures are similar, i.e., whether belonging to the same structural class or not. A strict identity measure such as, for example, that of a minimal root mean square sum for the backbone coordinates is clearly too strict—and even misleading—since similar but differently twisted structures might be judged as unrelated. In crystal structures it is known that most materials—and in particular shape-memory-alloys—assume a high symmetry, simple and open cubic (bcc) phase at high temperatures, just below the melting temperature. This is called the *parent* phase. At a lower temperature, at the *martensitic* transition, they condense into more complicated structures. We wish to describe the protein folds on a similar high symmetry level. To solve the homology problem, we consider the secondary structures as straight sticks and replace the loops by the interconnection lines between the end points of the secondary structures, which are defined by the sequence information. The unit vector $\hat{\mathbf{e}}_\ell$ of the first three elements is found and rectified onto the closest cubic unit vectors. Then the next element is added and the last three elements are locally rectified and joined to the previous, etc. In this way even severely twisted, similar protein structures in the Brookhaven data bank (PDB) can be projected into the same high symmetry fold without being sensitive to details in the angles or the lengths of the elements (which in the rectified structure are all equal). The second problem is to find a systematic and unique descriptor for a fold. This is done by using a Hamiltonian for an interacting spin system. The spins $S_\ell$ are positioned at the junctions of the elements at site $\ell$ and act as "hinge" variables describing the normal to the plane formed by the elements, as well as the sense of allowed bending. The sequence of interaction constants describes in a unique way the directions of the spins and thus the fold. The Hamiltonian for a chain with $\mathcal{N}$ secondary structures (in total $N = 2\mathcal{N} - 1$ of the defined elements) is

$$\mathcal{H} = -\sum_{P=2n+1=1}^{2\mathcal{N}-1} (J_P S_P \cdot S_{P+1} + K_P S_P \times S_{P+1} \cdot \hat{\mathbf{e}}_P) - \sum_{p=2n+2=2}^{2\mathcal{N}-2} (j_p S_p \cdot S_{p+1} + k_p S_p \times S_{p+1} \cdot \hat{\mathbf{e}}_p) . \qquad (1)$$

We have neglected the orientation of the start and end loops and allowed for two sets of interaction constants: The capital letters are for the interactions between the spins at the end of the secondary structures, and the small letters are for those of the loops; $J_\ell, j_\ell$ indicate continuation of the third element in the same plane as the two previous, $K_\ell, k_\ell$ indicate it is perpendicular. The sign controls if the third element must be joined parallel (antiparallel) or perpendicular (antiperpendicular) [right (left) turn]. Many underlying amino acid sequences can be reduced to these basic parameters. This provides the basis for the classification of sequences into fold classes. As a simple example, we show in Fig. 1 the projection of the 4-$\alpha$-helix bundle, which is given by the descriptor $jKj\bar{K}j$, where $\bar{K} = -K$. The descriptor depends on the direction in which the chain is traversed, but it is invariant under rotation. There are five other dense structures with seven elements. They are given by $j\bar{K}jKj$, $jK\bar{k}Kj$, $j\bar{K}k\bar{K}j$, $kJ\bar{k}Jk$, and $\bar{k}JkJ\bar{k}$. The letters $k$ and $K$, which describe chiral turns, change sign when the descriptor identifies a fold traversed in the reverse direction. Similar unique descriptors with $N - 2$ letters can be constructed for any fold with $N$ elements. We find that Haemarythrin, rabbit uteroglobin, and the cytochrome family belong to the mentioned $jKj\bar{K}j$ class. The structure also occurs in T4-lysozym. This is "embellished" by additional structures. This focuses on the question of a definition of families and of a metric. As a measure for closeness between classes we suggest that two proteins, not necessarily of the same length, have the largest similarity if the overlap in their descriptors is maximal.

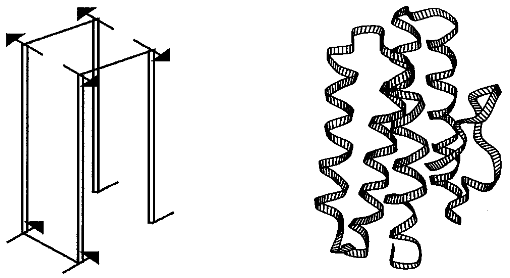

FIG. 1. 4-$\alpha$-helix bundle, Haemarythrin (1HMQ): Right, the actual structure in a ribbon representation. Left, the projected structure. The descriptor is $jKj\bar{K}j$.

Using the fact that the hydrophobic forces tend to confine the proteins and make them contain as little as 3% water [2] in the *native* state, we want to find all folds which are self-avoiding. A scaling and mean field theory [7] of this problem gives the estimate that the number of folds for $N$ elements increases as $(z/e)^N$, where $z$ is the coordination number, in our case $z = 4$, and $e = \ln^{-1}(1)$. For a protein with nine secondary structures and consequently eight interconnecting loop elements we have $N = 17$, and the above theoretical relation gives the number of folds as $(4/e)^{17} \sim 711$. This is already a quite small number. However, the discreteness gives rise to *magic* numbers at which there are particularly few, different folds. Figure 2 shows the exact enumeration [8] of all dense folds on a cubic lattice for elements up to $N = 35$. For $N = 17$ there is a pronounced minimum with only $p(17) = 172$ distinct and predictable folds. The mean field theory overestimates this grossly. Between the magic numbers the number of folds is, on the other hand, much larger. The magic number at $N = 7$, corresponding to the 4-$\alpha$-helix bundle, is a close packing of a $1 \times 1 \times 1$ box. The next closed confinement is the $2 \times 1 \times 1$ box, which we call $B$. The elemental magic numbers at $N = 11, 17, 23, 32$, and 35 can be understood as the optimal packing in closed polyhedra consisting of 1, 2, 3, 5, and 6 $B$ boxes.

In Fig. 3 the statistical distribution of proteins with a specific number of secondary structural elements is shown. We have used the prototypical standard set of 135 proteins with sequence similarity below 25% selected from PDB by Rost and Sander [9]. It is diverse and originally used for secondary structure prediction. The secondary structures are assigned using the renowned DSSP prescription [10] and counted when having at least four identical, consecutive assignments of either beta strands or alpha and $3_{10}$

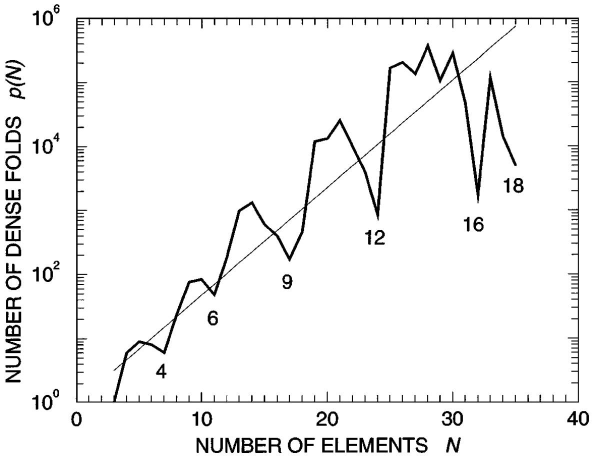

FIG. 2. Full line, number of distinct, dense folds for coordination number $z = 4$, on a cubic lattice as a function of number of elements $N$; thin line, $(z/e)^N$. Notice the deep minima at *elemental magic* numbers at the closed configurations. The added numbers indicate the corresponding number of secondary structures, $\mathcal{N} = (N + 1)/2$ for odd $N$.

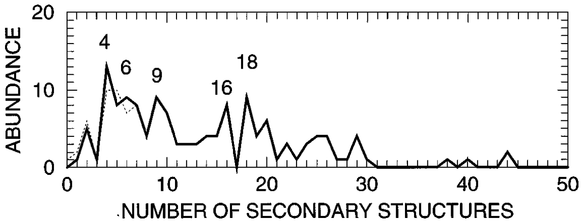

FIG. 3. Statistical abundance of proteins with $\mathcal{N}$ secondary structures. Dotted line, counting when the number of identical, consecutive DSSP assignments ≥4; full line, same but supplemented by 3D structural information for $\mathcal{N} \leq 6$ [11].

helices (not distinguished). The curve clearly shows local maxima in the abundance, which correspond to the optimal packing we find theoretically [11].

The dense packing criterion we have used is a simple count of the neighbors of end points of the elements. This represents the hydrophobic force quite faithfully. First, it is unspecific, i.e., independent of which elements are close to each other. Second, it depends on the "curvature" of the confinement approximately as a surface tension force, i.e., the different sites are rated 3, 4, 5, and 6 for a corner, edge, face, and buried site, respectively. Only the sum counts, in agreement with the nature of the hydrophobic force. One could, in order to introduce a temperature in the problem, assign energy values for the mentioned sites. The found magic numbers are not very sensitive to deviations from a linear weighting which is still consistent with the globular structures. The magic numbers in our model are *universal* in the sense that they do not depend on the specific, chemical interactions between the amino acids, neither between distant parts of the chain nor the interaction along the backbone. They are dictated by the hydrophobic, confining forces. If the weighting is far from linear one can form other families of proteins, for example, those that are dissolved in cell membranes. Clearly, for those the hydrophobic and hydrophillic forces act differently. Families could be imagined with a higher coordination number $z$ or other projected lattices. We have investigated the closed packed folds for the simple cubic lattice case also with $z = 5$, and find again a number of pronounced minima with the same magic numbers as before for the smaller domains.

The folds at the magic numbers are particularly stable and fast folding for the following reasons. They represent closed confinements having minimal surfaces and thus are energetically favorable from the point of view of the hydrophobic forces. Our magic number configurations have a clear energy separation from other folds. This is, according to a hypothesis by Shakhnovich [12,13], a necessary condition for rapid folding. The minimum at $N = 17$ is relatively well pronounced. There is also a well-pronounced minimum at the magic number $N = 35$. The $N = 35$ structure is confined in a $3 \times 2 \times 2$ box. An analysis of the folds shows that a large part is formed of two folds of the $N = 17$ domain interconnected by just a single element, i.e., $2 \times 17 + 1 = 35$. This explains why the domain formation of multipla of $\mathcal{N} = 9$ is a natural consequence of the discrete packing problem. Given the average size of the elements, the magic numbers also rationalize why the size of the domains [5] is as preferred by nature, being in concord with the overall thermodynamic theory [2]. Next we can evaluate how many distinct fold classes exist. If we restrict ourselves to domain structures with $N \leq 17$ we find in total 3906 possible, distinct globular fold classes. This is close to Chothia's estimate of 1000 fold classes, based on a heuristic argument [14].

The exhaustive enumeration in Fig. 2 and the unique description of folds on a cubic lattice are also relevant for the bead model of proteins, which is extensively studied; see a recent discussion, e.g., Ref. [13]. This model is, however, in fundamental principle very different from the present one. It assigns the physics to interactions between two or more different beads or residues distributed along a cubic model protein.

Our approach is more closely related to the models developed by Finkelstein *et al.* [15], where the secondary elements are considered as rigid units. They propose that the predominant occurrence of certain protein fold patterns is due to specific, small thermodynamic advantages and address the paradox of how entropy can play a role in determining the unique, native structure, which has zero entropy. Based on properties of the overall density of states a Boltzmann-like statistics is discussed for the abundance of a native folding pattern with the total number of folds, $M_1$, at, e.g., a given (lower) density. It reads occurrence $\propto \exp(-\tilde{F}_1/k_BT_*)$, where $T_*$ is a universal conformational temperature and $\tilde{F}_1$ is the selective free energy. This contains an entropylike term $-k_BT_* \ln M_1$, which would favor patterns with large $M_1$. We believe the physics is more delicate and involves elements of several phase transitions, in particular, by involving an intermediate phase, as in the martensitic case.

The dense structures we have enumerated do not represent the final native structures, but are somewhat expanded intermediate structures in which the secondary structures are basically developed, although not necessarily exactly in the native shape. There are $p$ different configurations, which are supposed to be degenerate with respect to the hydrophobic forces. We call this the *parent* phase in analogy to the martensitic problem. It is sufficiently free to be able to test the degeneracy $p$ and therefore gain an entropy contribution $-k_BT_* \ln p$, which is a significant part of the free energy at room temperature. We find the presence of such an intermediate phase is supported by experimental investigations [16] and protein engineering studies of Barnase [17]. For a smaller protein (chymotrypsin inhibitor 2) the folding process is found to be more concerted and both secondary and tertiary structures form almost simultaneously [18] in what is called a nucleation-condensation mechanism. However, this is not contrary to our picture because (1) for small $N$ there is no sharp phase transition and (2) the experiment is concerning a dynamic process, whereas we are considering only the statistical properties. Upon lowering the temperature, we assume there is a phase transition to the unique native state. This is driven by the short ranged interactions $f_{ab}(\mathbf{r}_i^a - \mathbf{r}_j^b)$ between the residues $a$ and $b$ at $\mathbf{r}_i^a$ and $\mathbf{r}_j^b$ on the neighboring elements as in the bead model. The energy gain $\Delta E = \sum_{ab} f_{ab}(\mathbf{r}_i^a - \mathbf{r}_j^b)$ is limited because the interactions are highly frustrated. $\Delta E$ depends on the underlying sequence information. A model for the martensitic transformation was recently studied by one of the authors [19]. This exhibits similarly a competition between a $p$ times degenerate parent phase and an energy stabilized low temperature phase. It required a larger $\Delta E$ to produce the (discontinuous, growth by nucleation) transition from or to a more highly degenerate parent phase at a given temperature, because the latter is stabilized by the entropy term $-k_B T_* \ln p$. In the present case a large $\Delta E$ requires a particularly favorable sequence of amino acids, whereas a more random sequence will have a smaller $\Delta E$. In nature the diversity in proteins with different sequences but similar structures is therefore more likely for those for which the parent phase has low degeneracy. Thus we have argued that the structures corresponding to the magic numbers in Fig. 2, as well as proteins with relatively small $N$, are more abundant, as demonstrated by the statistics Fig. 3.

Forces, for example, transmitted through the protein backbone, might weakly act in positioning the incipient structural elements in roughly the right place in space early in the folding process. They could operate already in forming a suitable "partly collapsed" state, as introduced by Itzhaki *et al.* [18] for setting up the folding. In a statistical model they can, however, simply be *represented* by the hinge forces in the Hamiltonian equation (1), defining the actual values of the parameters. In principle these are then determined by the sequence information, selected during the course of the evolution. We propose as a possibility that for a protein with $N$ elements and $p$ times dense packing degeneracy, the hinge forces sum up to give maximum energy gain for the potential native fold. The $p - 1$ other states will then have a higher energy according to how many letters in the descriptor have been violated. The effect is that of a weak symmetry breaking field [20]. Of course, symmetry breaking may also arise from the short range forces and the sequences of $a$ and $b$, etc. However, the result of this effect is known only in the final native phase. The hinge forces are weakly symmetry breaking the intermediate, parent phase, which is conceptually advantageous.

In terms of a simple model for the effect of hydrophobic forces in the protein folding, we have demonstrated the appearance of magic numbers of secondary structural elements. The model has the property that the folds of these number of elements are favorable with respect to being fast folding and the corresponding native ones to being potentially stable, thermodynamically. This allows us to predict a predominant abundance of proteins with such numbers of secondary structures. A statistical analysis of experimental data supports this finding.

It is a pleasure to thank K. Rapacki for assistance with the numerical computations. We also wish to thank the Danish National Research Foundation for financial support.

[1] C. Tandford, *The Hydrophobic Effect: Formation of Micelles and Biological Membranes* (J. Wiley & Sons, New York, 1980).

[2] K. A. Dill, Biochemistry **24**, 1501 (1985).

[3] C. J. Levinthal, Chem. Phys. **65**, 99 (1968).

[4] P. G. Wolynes, *Protein Folds,* edited by H. Bohr and S. Brunak (CRC Press, New York, 1995), pp. 3–17; M. Sasai and P. G. Wolynes, Phys. Rev Lett. **65**, 2740 (1990).

[5] A. L. Berman, E. Kolker, and E. N. Trifonov, Proc. Natl. Acad. Sci. **91**, 4044 (1994).

[6] L. Pauling and R. B. Corvey, Proc. Natl. Acad. Sci. **37**, 729 (1951); L. Pauling, R. B. Corvey, and H. R. Branson, *ibid.* **37**, 205 (1951).

[7] H. Orland, C. Itzykson, and C. de Dominicis, J. Phys. (Paris) Lett. **46**, L353 (1985).

[8] The exhaustive enumeration is done as a test in two ways (for $N \leq 24$) using Eq. (1): (i) by direct selective construction using MATHEMATICA, (ii) by generating all structures and discrimination.

[9] B. Rost and C. Sander, J. Mol. Biol. **232**, 584 (1993).

[10] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).

[11] The database is small but representative. As a test the first and second half of it give the same trend with maximum abundance around $\mathcal{N} = 4$ and 9 and approximate multipla of that. The use of a *native,* twisted structure for the counting introduces an uncertainty because of (1) double counting if an element is broken into two, (2) noncounting if an element is broken beyond recognition. The problem is most serious for small $\mathcal{N}$. Doubt can be resolved by considering the 3D structure.

[12] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).

[13] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, Proc. Natl. Acad. Sci. **92**, 325 (1995).

[14] C. Chothia, Nature (London) **357**, 543 (1992).

[15] A. V. Finkelstein, A. M. Guntun, and A. Y. Badretdinov, FEBS Lett. **325**, 23 (1993); A. V. Finkelstein and O. B. Ptitsyn, Prog. Biophys. Mol. Biol. **50**, 171 (1987).

[16] C. Redfield, R. A. G. Smith, and C. M. Dobson, Nature Struc. Biol. **1**, 23 (1994).

[17] A. R. Fersht, FEBS Lett. **325**, 5 (1993).

[18] L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, J. Mol. Biol. **254**, 260 (1995).

[19] E. Vives, T. Castán, and P.-A. Lindgård, Phys. Rev. B **53**, 8915 (1996).

[20] The proposed hinge force assisted folding is in fact a much more direct and reliable process than the corresponding defect assisted selection of variants, which occurs in "trained" shape memory alloys. Given the code for the hinge forces the descriptor, i.e., the fold class, can be directly predicted from the sequence information.