ANNUAL REVIEW OF BIOPHYSICS AND BIOPHYSICAL CHEMISTRY

VOL. 17 1988

ANNUAL REVIEW OF BIOPHYSICS AND BIOPHYSICAL CHEMISTRY

VOLUME 17, 1988

DONALD M. ENGELMAN, *Editor*
Yale University

CHARLES R. CANTOR, *Associate Editor*
Columbia University

THOMAS D. POLLARD, *Associate Editor*
The Johns Hopkins University School of Medicine

ANNUAL REVIEWS INC 4139 EL CAMINO WAY P.O. BOX 10139 PALO ALTO, CALIFORNIA 94303-0897

ANNUAL REVIEWS INC.
Palo Alto, California, USA

International Standard Serial Number: 0883-9182
International Standard Book Number: 0-8243-1817-X
Library of Congress Catalog Card Number: 79-188446

TYPESET BY AUP TYPESETTERS (GLASGOW) LTD., SCOTLAND
PRINTED AND BOUND IN THE UNITED STATES OF AMERICA

SOME RELATED ARTICLES APPEARING IN OTHER *ANNUAL REVIEWS*

From the *Annual Review of Biochemistry*, Volume 57 (1988):

Sequences, Sequences, and Sequences, Frederick Sanger

Molecular and Cellular Biology of Intermediate Filaments, Peter M. Steinert and Dennis R. Roop

Lens Crystallins: The Evolution and Expression of Proteins for a Highly Specialized Tissue, Graeme J. Wistow and Joram Piatigorsky

DNA Repair Enzymes, Aziz Sancar and Gwendolyn B. Sancar

Viral Proteinases, Hans-Georg Kräusslich and Eckard Wimmer

Human Class II Major Histocompatibility Complex Genes and Proteins, Dietmar Kappes and Jack L. Strominger

Cell-Surface Anchoring of Proteins Via Glycosyl-Phosphatidylinositol Structures, Michael A. J. Ferguson and Alan F. Williams

From the *Annual Review of Cell Biology*, Volume 3 (1987):

Laminin and Other Basement Membrane Components, George R. Martin and Rupert Timpl

Molecular Aspects of B-Lymphocyte Activation, Anthony L. DeFranco

Intracellular Transport Using Microtubule-Based Motors, Ronald D. Vale

Myosin Structure and Function in Cell Motility, Hans M. Warrick and James A. Spudich

From the *Annual Review of Genetics*, Volume 21 (1987):

Natural Variation in the Genetic Code, Thomas D. Fox

The Genetics of Active Transport in Bacteria, Howard A. Shuman

From the *Annual Review of Microbiology*, Volume 41 (1987):

Repetitive Proteins and Genes of Malaria, D. J. Kemp, R. L. Coppel, and R. F. Anders

Export of Protein: A Biochemical View, L. L. Randall, S. J. S. Hardy, and Julia R. Thom

An Inquiry into the Mechanisms of Herpes Simplex Virus Latency, Bernard Roizman and Amy E. Sears

High-Resolution NMR Studies of Saccharomyces cerevisiae, S. L. Campbell-Burk and R. G. Shulman

(*continued*)

From the *Annual Review of Physical Chemistry*, Volume 38 (1987):

Fashioning Electron Spin Echoes into Spectroscopic Tools: A Study of Aza-aromatic Molecules in Metastable Triplet States, J. Schmidt and David J. Singel

Molecular Modeling, Peter Kollman

Membrane and Vesicle Fusion, J. H. Prestegard and M. P. O'Brien

Biochemical Applications of Differential Scanning Calorimetry, Julian M. Sturtevant

The Transition Between B-DNA and Z-DNA, Thomas M. Jovin, Dikeos M. Soumpasis, and Lawrence P. McIntosh

Three-Dimensional X-Ray Crystallography of Membrane Proteins: Insights into Electron Transfer, David E. Budil, Peter Gast, Chong-Hwan Chang, Marianne Schiffer, and James R. Norris

From the *Annual Review of Physiology*, Volume 50 (1988):

Voltage Dependence of the Na-K Pump, Paul De Weer, David C. Gadsby, and R. F. Rakowski

Site-Directed Mutagenesis and Ion-Gradient Driven Active Transport: On the Path of the Proton, H. Ronald Kaback

Effects of Lipid Environment on Membrane Transport: The Human Erythrocyte Sugar Transport Protein/Lipid Bilayer System, Anthony Carruthers and Donald L. Melchior

Chloride Transport in the Proximal Renal Tubule, L. Schild, G. Giebisch, and R. Green

Ion Channels in Drosophila, Diane M. Papazian, Thomas L. Schwarz, Bruce L. Tempel, Leslie C. Timpe, and Lily Y. Jan

From the *Annual Review of Plant Physiology and Plant Molecular Biology*, Volume 39 (1988):

Electron Transport in Photosystems I and II, L.-E. Andréasson and T. Vänngård

Enzymatic Regulation of Photosynthetic CO_2 Fixation in C_3 Plants, Ian E. Woodrow and Joseph A. Berry

Coated Vesicles, D. G. Robinson and Hans Depta

ANNUAL REVIEWS INC. is a nonprofit scientific publisher established to promote the advancement of the sciences. Beginning in 1932 with the *Annual Review of Biochemistry*, the Company has pursued as its principal function the publication of high quality, reasonably priced *Annual Review* volumes. The volumes are organized by Editors and Editorial Committees who invite qualified authors to contribute critical articles reviewing significant developments within each major discipline. The Editor-in-Chief invites those interested in serving as future Editorial Committee members to communicate directly with him. Annual Reviews Inc. is administered by a Board of Directors, whose members serve without compensation.

Annual Review of Biophysics and Biophysical Chemistry
Volume 17, 1988

CONTENTS

(*continued*)

Ann. Rev. Biophys. Biophys. Chem. 1988. 17: 1–21

A CRITICAL EVALUATION OF METHODS FOR PREDICTION OF PROTEIN SECONDARY STRUCTURES

Georg E. Schulz

Institut für Organische Chemie und Biochemie der Albert-Ludwigs Universität, Albertstrasse 21, 7800 Freiburg im Breisgau, Federal Republic of Germany

CONTENTS

HISTORICAL DEVELOPMENT OF THE FIELD

Postulation of α-Helix and β-Pleated Sheets

From the early observation that proteins crystallize (55), it could have been derived that they possess defined structures. More than 50 years ago

0883–9182/88/0610–0001$02.00

the first X-ray diffraction studies with protein crystals demonstrated that they are ordered at a scale smaller than the atomic bond distances (12). However, lack of suitable methods (51) prohibited closer inspection for another quarter of a century until the crystal structure of myoglobin was solved (65).

Because of the importance of proteins and thus of protein structures for all processes of life, there were various attempts to elucidate at least some aspects of their structure. X-ray analyses of protein fibers such as wool and silk showed the so-called α- and β-patterns, respectively (9). The patterns revealed local order in these fibers and yielded the spatial repeat distances of atomic groups.

Using these distances, together with current knowledge on dimensions, flexibility, and hydrogen bond (H-bond) formation of peptide bonds, Pauling et al (100) constructed three local ordering schemes: the α-helix, corresponding to the α-type of X-ray pattern, and parallel as well as antiparallel β-pleated sheets, both corresponding to the β-type. Soon afterward, the α-type X-ray pattern was also recognized for crystals of the globular protein hemoglobin (102), indicating strongly that Pauling's constructs are of universal relevance for protein structures.

Local Order in Synthetic Polypeptides

Syntheses of homopolypeptides or polypeptides with random sequences of amino acid residues correspond to the usual polymerizations in materials science. For a number of the chemically well-defined homopolymers, local order corresponding to α-helices and β-sheets could be detected under certain conditions. Furthermore, the structural effect of the random incorporation of a second residue type into an α-helix– or β-sheet–forming homopolymer (guest-host) could be followed.

Based on such experiments (14) the 20 genetically coded amino acids (except glycine and proline) were subdivided into α-helix and random coil–forming groups. These assignments correlated (32) with the spectroscopically derived α-helix and β-sheet contents (17, 21) of globular proteins of known amino acid compositions. Thus the physicochemical data on guest-host relationships in homopolypeptides became applicable to biological proteins.

Levels of Protein Structure

After the elucidation of the first globular protein structure, Linderström-Lang (83) introduced a concept of four levels for the description of biological polypeptide chains: the primary, the secondary, the tertiary, and the quarternary structures. Soon afterward, Anfinsen & Haber (2) demonstrated that the primary structure contains all the structural information

and therefore determines all other stages; thus they converted the four-stage concept to a hierarchical scheme (Figure 1). This scheme has been supplemented by the concepts of supersecondary structures (106) and domains (128).

The hierarchy is such that the lower-level elements determine those of the higher levels. Since the amino acid sequence contains all information, it should be possible to derive the final protein structure step by step from the sequence to the secondary structures, supersecondary structures, domains, globular proteins, and aggregates. However, the information is very intricately encoded, and the depicted lateral segregation (Figure 1) is by no means total. As a consequence, a procedure for working up from sequence to aggregate remains a futuristic aim.

Methods have been developed that could provide some success in the first step, i.e. in deriving secondary structure from amino acid sequence. The secondary structures have only limited value compared to a complete three-dimensional protein structure; however, because the available amino acid sequence data have so proliferated with the advent of DNA cloning and sequencing, the derivation of any structural feature is of interest. Accordingly, it is worthwhile to review the methods for prediction of secondary structure.

DEFINITION OF SECONDARY STRUCTURE

Helices and Sheets as Defined by Hydrogen Bonds

When Pauling et al constructed α-helices and β-sheets (100) they relied on the planarity of peptide bonds and on the presence of linear H-bonds with

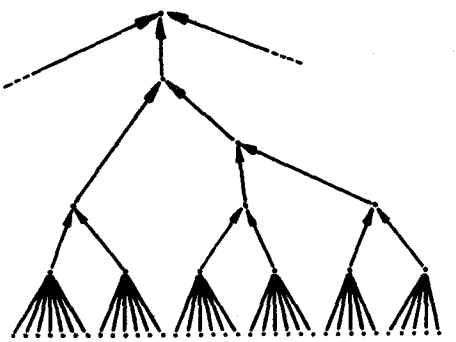

Figure 1 Structural hierarchy in proteins. The conventional primary, secondary, tertiary, and quaternary structures of Linderström-Lang (83) are currently classified as amino acid sequence, secondary structure, globular protein, and aggregate, respectively. The sketch does not account for the spatial interactions between amino acid residues that are far apart along the chain.

lengths of 3 Å (N···O distance) between imide and carbonyl groups. We know now that in the best H-bonds the carbonyl and imide dipoles are not exactly linearly aligned. Rather, a small deviation brings the hydrogen close to one of the sp2-lone electron pairs of the oxygen (118). Neither the linear arrangement nor this optimal arrangement can be achieved in α-helices, where the limitations of the polypeptide backbone geometry force the hydrogen to a more unfavorable position. This fact has been read from a number of protein structures known at a resolution as high as 1.5 Å or better (10). In contrast, the H-bonds in regular parallel and antiparallel β-sheets can assume the optimal arrangement in a much better way. A statistical analysis showed that in accordance with these geometric differences, the H-bonds in β-sheets are generally shorter than those of α-helices (10, 64); the average N···O distances are 2.91 Å in the former and 2.99 Å in the latter. Thus, the contributions of β-sheets to protein stability seem to be higher per H-bond, which is roughly per residue.

Furthermore, α-helices tend to peter out at both ends (64); the values of the conformational parameters are most regular in the α-helix cores, where each residue forms two H-bonds. In many cases, α-helix ends cannot be uniquely defined because the H-bond interactions of residues $(i, i+4)$ become progressively longer and weaker. Therefore, a cutoff criterion has been suggested on the basis of high-resolution structures. An upper H···O distance limit of 2.5 Å was found to be appropriate for α-helices as well as β-sheets (10). For medium-resolution structures with appreciably larger coordinate errors, the 2.5 Å cutoff leads to spurious helix and sheet breaks. Here, a less stringent energy-type criterion seems to be more suitable (60).

In an α-helix there are also $(i, i+3)$ H-bond interactions, which are longer and less well oriented then the regular $(i, i+4)$ ones. Frequently, the $(i, i+3)$ interactions dominate at the α-helix ends, changing it to a 3_{10}-helix. These 3_{10}-helices are difficult to spot when a structure is interpreted only visually and at medium resolution. As a consequence, the residues reported as N- and C-termini of α-helices usually depend on the achieved resolution and vary with the progression of reports. The same applies for β-sheets.

Reverse Turns

The general chain fold of a globular protein resembles a heap of spaghetti, which for the most part run straight through the center and reverse their direction at the surface. These "reverse turns" or "turns" have attracted attention since the first regular turns in proteins with β-sheets were located. Their conformations were originally classified into types I, II, and III (122). Subsequent analyses of high-resolution protein structures, however,

showed that there is a broad range of turns. Only the type-II turns can be clearly separated from the rest. The type-III conformation is practically identical with one winding of a 3_{10}-helix. This similarity necessitated the following suggested convention (64): Type-III turns at the ends of α-helices as well as consecutive type-III turns form a 3_{10}-helix.

To describe all turns irrespective of their types and the presence or absence of an H-bond, a simple distance criterion between Cα-atoms has been proposed (78). Turns are defined by a distance of less than 7 Å between $C\alpha_i$ and $C\alpha_{i+3}$ atoms in all conformations except α-helices. This criterion has been widely adopted for the prediction of turns from the sequence. It demonstrates that a turn is not very stable in itself. Rather, it is a passive kinking point of the chain. Nevertheless, the prediction of this conformation is quite successful, presumably because most of the turns are at the molecular surface and contain polar amino acids (109).

There have been attempts to define further types of reverse turns, but to little avail. The distribution is broad and there is no other clearly distinguished group. In most turns adjacent to α-helices and β-sheet strands, there are residues in α- or β- as well as in turn conformation. Such ambiguities have to be resolved by a convention; usually residues in turns that also participate in α-helices or β-sheets are assigned to the latter conformations.

Coil Conformation

Commonly, all amino acid residues that are not in α-helices, β-sheets, or turns are designated as "random coil." This expression has been adopted from work with synthetic polypeptides in materials science. These polymers either form regular secondary structures or assume a random conformation. In contrast, there are few natural polypeptides or chain segments with random structures; randomness contradicts life. Therefore, "coil" describes a well-defined structure, which is just less regular than secondary structures.

Now that a large amount of protein structural data are at hand, recurring features in coil conformation have emerged. Well known are the β-bulge (107) giving rise to a kink in a β-sheet, a special turnlike H-bonding pattern at the end of α-helices (110), a calcium-binding loop (88, 125), an iron-sulfur cluster binding sequence (1), a general metal-binding feature in zinc fingers (52), a giant anion hole for phosphates (37), the signal peptides (123), and the charge relay system of serine proteases, consisting of a serine backed by an imidazole and a carboxyl (15). It is worthwhile to search proteins for such features in order to obtain structural information prior to a complete structural analysis; thus secondary structure prediction can be supplemented by other structural predictions.

Other Parameters

A parameter popularly followed in many newly established amino acid sequences is the polarity of the amino acid side chains, as derived, for instance, from partition coefficients (97) or vapor pressure differences (129) between water and less polar solvents. Polarity values have also been derived from the known globular protein structures (22, 39, 57, 71, 124). They seem to be of utmost general importance for structural integrity and have therefore been used as a guideline in the residue arrangement of Figure 2, below. In general, the polarity values are averaged over a given number of residues, yielding a function smooth enough to be interpreted in terms of hydrophobic and hydrophilic segments of the polypeptide chain. Such plots are useful for the identification of the membrane-penetrating segments of a membrane-bound protein (6, 39, 42–44, 71) or for the identification of the nonpolar signal peptides (123). They also contain significant information for the localization of reverse turns (108, 109) and β-strands, because these are mostly at the molecular surface and in the molecular interior, respectively.

A technically important property of certain regions at the molecular surface is their antigenicity. On the basis of a number of established antigenic sites in lysozyme and myoglobin, Hopp & Woods (54) designed a general prediction method. Technical interest in this method rose when it became clear that carrier-coupled synthetic peptides more than 10 residues long, containing the amino acid sequence of an antigenic site, could be used to elicit antibodies against the corresponding natural protein (127). Thus, the prediction of an antigenic site from an amino acid sequence allows isolation of the gene product when only the DNA sequence of the gene is known, which for rare proteins is often the case. The best results were obtained with synthetic peptides mimicking the most flexible parts of a protein, presumably because the carrier-coupled peptides are also flexible. Accordingly, Karplus & Schulz (63) designed a method to pick out these parts of the sequence. This method was based on a data base of high-resolution protein structures in which the main-chain mobility had been established. Apart from its technical applications, this method is now also used for the prediction of flexible residues between α-helices and β-sheets, which helps in secondary structure prediction (31).

Another predicted parameter is the location of an amino acid relative to the surface of a globular protein (94). This prediction is based on the frequency distribution of contact numbers for particular residue types, where the contact number corresponds to the number of neighboring residues with $C\alpha_i \cdots C\alpha_j$ distances of less than 8 Å. There has also been a report on the prediction of nonpolar cluster formation (30) based on cluster

analysis inside globular proteins. Moreover, there have been attempts to apply the early observation that the amino acid composition of a protein correlates with its α-helix contents (32) to develop a prediction method. There is a scheme for predicting the amount of α-, β-, and turn (t-) structure from the residue composition (69). Even more daring are the proposals to identify the structural class (α, α/β, α+β, β) from the amino acid composition of a globular protein (66, 92).

EVALUATION OF PREDICTIVE SUCCESS

Before scrutinizing particular methods, I consider how to evaluate predictive success. Besides the α-, β-, and t-conformation predictions described above, attempts to predict the polypeptide conformations in more detail have been rare and generally not very successful (45, 58, 85). Therefore, the discussion is restricted to these three types. The remaining part of the chain is considered a coil, the fourth type.

Predictions of Varying Numbers of Secondary Structure Types

With four conformational types there are $\binom{4}{1}+\binom{4}{2}+\binom{4}{3}+\binom{4}{4} = 15$ possibilities for predictions of single or combined types. These combinations have different levels of correctness for random prediction (Table 1). Clearly, this level is lowest for a simultaneous prediction of all types and highest for the prediction of a single rare type of secondary structure. This should be kept in mind when comparing success rates.

Table 1 Correctness levels for random prediction of the most popular combinations of secondary structures[a]

Secondary structure combinations[b]	Percentage of correctness assigned residues in a random prediction
α, β, t, c	25
α, β, non-αβ	38
α, non-α	61

[a] The presence of a certain percentage of each conformation (α, β, t, c) is assumed. The overall values vary with time as new protein structures emerge and with the interpretation of protein models in terms of α-, β-, and t-conformations (see text). Here, I use α = 27%, β = 23%, t = 25%, and c = 25%. It is assumed that the random prediction honors these percentages.

[b] α = α-helix, β = β-pleated sheet, t = reverse turn, c = coil = non-αβt.

Quality Indices

There are different ways of judging success. Deriving the percentage of correctly predicted residues appears to be an appropriate way to evaluate the achievements (114), but it is not the only way. Correlation coefficients and many quotients have been designed to portray the allegedly most relevant aspects (114). Since the selection of these quotients is subjective, the result of a prediction is better stated in the form of all possible numbers, as shown in Table 2. With these numbers any quality index can then be calculated. The 16 values of Table 2 can be used in a multitude of combinations. A simplistic approach to arrive at a percentage of correctly predicted residues is to add the diagonal and relate it to the sum of all the numbers in Table 2.

Which is the Important Unit to Predict?

Usually, secondary structures are predicted and analyzed on a per-residue basis. However, the assignment of secondary structures is not straightforward; there are ambiguities at α-helix and β-sheet ends (see above). These cannot always be resolved, even for X-ray structures known at high resolution. Furthermore, it is necessary to clarify whether 3_{10} helices are allotted to turn, α-helix, or coil. In addition, owing to the broad range of observed conformations, there is no universally accepted definition for a turn. These assignment difficulties should be solved before the evaluation of predictive success. Therefore, programed procedures have been designed (60, 64, 75) for interpreting known three-dimensional protein structures in terms of secondary structures.

Table 2 A general representation of the results of a secondary structure prediction method[a] evaluated on a per-residue basis[b]

Observed conformation	Predicted conformation				
	α	β	t	c	Total
α	70[c]	21	0	13	104
β	0	17	3	2	22
t	9	3	16	7	35
c	5	4	7	17	33
Total	84	45	26	39	194

[a] The Chou & Fasman method (23, 24), as applied to the test case adenylate kinase (113).

[b] A random prediction using the values of Table 1 would have a correctness level of 26% in this case.

[c] The amount of correctly predicted residues is $(70+17+16+17)/194 = 62\%$, which shows that the success rate with adenylate kinase is high.

The results of different prediction methods (84, 113) suggest that per-residue counting sometimes yields a distorted picture. A particular method may detect all secondary structure segments per se, but may fail in finding the correct ends. The accumulation of these boundary errors, amounting to a few residues at each end, may give a count inferior to that obtained by a method that finds the limits more efficiently but fails to detect all secondary structure segments. At this point one has to reconsider the purpose of these predictions. It will probably not be possible to go through the hierarchy of Figure 1 step by step, because the detection of secondary structures from localized segments of the sequence is not accurate enough, as it does not account for long-range interactions. On the other hand, a multitude of protein structures are now known, and it is clear that they belong to a very limited number of chain folds (112). Thus, it would be a great achievement if the prediction could be used to assign a sequence to a known chain fold, as Crawford et al have done (31). For this purpose, the correct sequence of secondary structure segments is much more important than the correct limits of some of these segments if others are wrong (119).

For the detection of a chain-fold type, it is important that the available information not be reduced too early to binary (yes or no) assignments with respect to a particular secondary structure. For scanning through all chain-fold possibilities, it is more advisable to keep the α-helix, β-sheet, and turn potential curves so that strong and weak assignments can be distinguished. This is particularly true if more than one prediction method is used and discrepancies have to be resolved.

PREDICTION SCHEMES

Probabilistic Methods

SINGLET FREQUENCIES AND PROPENSITIES In the most simple statistical approach to secondary structure prediction, the frequencies of each of the 20 standard residue types are determined in each of the four conformational states (α, β, t, c) in the data base of known protein structures. These frequencies are taken as propensities for a given residue type to occur in the respective conformational state (11). As the observed frequencies can be considered basic to all prediction methods, some published values are depicted in Figure 2.

As a plot of these propensities along the polypeptide chain is usually very erratic, the propensities are locally smoothed by some averaging procedure (11, 108). This procedure gives rise to a potential curve for the respective conformational state (114). At each residue position the predicted conformation is assigned according to the highest potential curve.