PB93-210979

# Quantitative Characteristics of Primary Amino Acid Sequences Predict 'Fractal' Measures on Tertiary Structures of Proteins

Institut des Hautes Etudes Scientifiques, Bures-sur-Yvette (France)

Prepared for:

Office of Naval Research, Arlington, VA

May 93

BIBLIOGRAPHIC INFORMATION

PB93-210979

Report Nos: IHES/M-93/21

Title: Quantitative Characteristics of Primary Amino Acid Sequences Predict 'Fractal' Measures on Tertiary Structures of Proteins.

Date: May 93

Authors: A. J. Mandell, and K. A. Selz.

Performing Organization: Institut des Hautes Etudes Scientifiques, Bures-sur-Yvette (France).**Florida Atlantic Univ., Boca Raton.

Sponsoring Organization: *Office of Naval Research, Arlington, VA.

Supplementary Notes: Prepared in cooperation with Florida Atlantic Univ., Boca Raton. Sponsored by Office of Naval Research, Arlington, VA.

NTIS Field/Group Codes: 57F, 57B, 99F

Price: PC A02/MF A01

Availability: Available from the National Technical Information Service, Springfield, VA. 22161

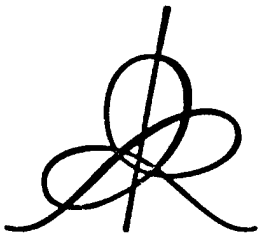
Number of Pages: 6p

Keywords: *Proteins, *Enzymes, *Tertiary protein structure, *Amino acid sequence, Thermodynamics, Protein conformation, Calorimetry, Analysis of variance, Hydrophobicity.

Abstract:

# QUANTITATIVE CHARACTERISTICS OF PRIMARY AMINO ACID SEQUENCES PREDICT "FRACTAL" MEASURES ON TERTIARY STRUCTURES OF PROTEINS

A. J. MANDELL K.A. SELZ



Institut des Hautes Etudes Scientifiques
35, route de Chartres
91440 - Bures-sur-Yvette (France)

Mai 1993

IHES/M/93/21

PB93-210979

# QUANTITATIVE CHARACTERISTICS OF PRIMARY AMINO ACID SEQUENCES PREDICT "FRACTAL" MEASURES ON TERTIARY STRUCTURES OF PROTEINS

Arnold J. Mandell and Karen A. Selz
Departments of Mathematics, Physics, Psychology
and the Clinical Psychology Center
Florida Atlantic University, Boca Raton, FL 33431

(Current Address: Institut des Hautes Etudes Scientifiques)

May, 1993

The quantitative prediction of the tertiary structure of proteins as defined by their $x$-ray crystallographic coordinates, using statistical physical and/or symbolic characteristics of the primary amino acid sequence is a long standing problem in biopolymer physics. An erstwhile missing feature of protein structural data has been a measure for the mapping of a set of $x$- ray observables onto a single real number as a continuously distributed descriptor which could then serve as the object of quantitative prediction. Such a global measure using the $x$-ray coordinates of protein crystals has been developed by Stapleton and associates.[1-3] Computation of the range of inter−$\alpha$-carbon distances indicated that there were protein specific, statistically reliable, fractional power laws (we call them "Stapleton protein fractal dimensions", $d_S$) relating the amino acid monomeric mass density to the $\alpha$-carbon distances with values ranging from 1.26 to 1.87. Although insufficient in orders of magnitude of length to qualify for the definition of fractals, intuitively, the values appear related to the space filling aspects of tertiary structure. For examples, the "curled up," barrel dominated proteins such as equine hemoglobin A and $B$ and sperm whale myoglobin manifested $ad_S \cong 1.65$, whereas in the "stretched out" more "random" chains such as protease A and $B$ from *s. griseus*, $d_S = 1.31$ and 1.32 respectively. From similar computations yielding the "fractal dimension" of a polymer represented by the orbit of a self avoiding random walk in three dimensions, it is speculated that the upper bound on $d_S$ is in the vicinity of 5/3.[1]

Chothia's studies relating a protein's conformation deduced from $x$-ray crystallographic data[4] to its amino acid side chain hydrophobicity values, $hp_i$ (in $cal\,K^{-1}\,mol^{-1}$), derived from the studies of Nozaki and Tanford[5], suggested a measure map to the reals for the amino acid sequence as quantitative predictors. Each protein can be transformed into a hydrophobic sequence, $\Sigma hp_i$, from which a statistical model to predict the proteins' values for $d_S$ could be developed. A representative set of amino acid-hydrophobic transformations[6] in $cal\,K^{-1}\,mol^{-1}$ yield: $q = 0.00, s = 0.07, t = 0.07, n = 0.09, g = 0.10, d = 0.66, e = 0.67, r = 0.85, a = 0.87, h = 0.87, c = 1.52, k = 1.64, m = 1.67, v = 1.87, l = 2.17, y = 2.76, p = 2.77, f = 2.87, i = 3.15$, and $w = 3.77$.

That relationships between measures on $\Sigma hp_i$ and $d_S$ have physical meaning is suggested by two groups of research findings: (1) The related studies by the Stapleton group yielding solvent ionic strength-sensitive, densities of low frequency ($< 300 cm^{-1}$), vibrational state fractional power laws, when probed by temperature dependent, Raman spin-lattice relaxation techniques in heme and iron-sulfer proteins, which were very similar to the proteins' values for $d_S$;[1-3] (2) Calorometric studies of the specific heats of proteins consistent with the presence of internal "soft" modes with low fundamental frequencies ($< 500 cm^{-1}$), easily excitable and subject to the influence of hydrophobic factors in folding, ligand binding, and ionic environment.[7-9]

One might relate these two ideas to the space-filling dependence of $d_S$ with the idea that protein relaxation dynamics might vary in temporal complexity when comparing the potential motions along the spatially one-dimensional amino acid peptide backbone with the more complex, hierarchical multimodal, higher dimensional case involving hydrophobic interactions of the amino acid side chains (off backbone connectivity) in addition to one dimensional pathways. The distributions of modes, $\rho(\gamma)$, between $\gamma$ and $\gamma + \delta\gamma$, would go like $\gamma^{-d_S}, 1 < d_S < 2$. These considerations motivated the development of quantities on $\Sigma hp_i$ which might describe the potential for hydrophobic "mode" structure predictive of Stapleton's measure, $d_S$.

In 35 proteins representing the range of reported values for $d_S$, we studied the relationships between $d_S$ obtained in two series of studies by the Stapleton group[10] and four statistical transformations on $\Sigma hp_i$.

The transformations included (1) As a statistical modulus, the average hydrophobicity per amino acid residue, $\overline{hp} = \frac{\sum hp_i}{n}$ ($cal\ K^{-1}\ mol^{-1}$), (2) As a statistical wave length, the average inter-high hydrophobic run interval in number of amino acid residues, $\omega_{hh}$, in which the values for $hp$ for the amino acid sequences were partitioned into 0 for $hp \leq$ leucine ($i.e.\ \leq 2.17$) and 1 for values $\geq$ leucine; (3) As an estimate of longer range order in the hydrophobic sequence partitioned into four bins (letters) of five amino acids each ($\leq 1.0,\ 1.66 - 1.87,\ 1.52 - 2.17,\ 2.76 - 3.77$), the longest "word", $\lambda_{word}(\Sigma hp_i)$, in number of amino acid residues (a word is defined as a sequence of amino acid residue transformations that appears at least twice along the length of the protein). Whereas $\omega_{hh}$ yields values sensitive to small structure (for example, the hemoglobins and myoglobin with high densities of $\alpha$-helixes manifest the expected values $\cong 3.5$ and $\alpha\pi$ helix-like value of 3.3 was seen in carboxypeptidase as "average turn lengths"), $\lambda_{word}$ varied up to 15 residues. $\omega_{hp}$ is similar to the rotation number used in studies of two dimensional reductions of three dimensional dynamical systems[11] and $\lambda_{word}$ is derived from symbolic dynamics and lexical compression algorithms.[12] (4) As a correction term, we used the percent of the sequence length that was proline, %(PRO), its role as a "structure breaker" in native protein conformations being well known.

We remind ourselves of the Erdos-Renyi "new law of large numbers" which says that the longest expected repetition length in a random sequence is asymptotically $= \log_{base(p)} n$.[13] $\lambda_{word}$ exceeded this value for all proteins studied. For examples, for the four letter code ($p = 0.25$) in the $n = 141$ residue hemoglobin, a longest word length, $\lambda_{word}$ of 3.56 was expected and two distinct $\lambda'_{word}$s of 6 residues were observed; the expectation for protease A was 3.75 and an 11 residue word was observed; for elastase it was 3.95 versus 13.

The proteins studied were: protease A and $B$(*S. griseus*), myoglobin (*sperm whale*), rhodanese (*bovine*), staphylococcal nuclease (*S. aureus*), glyceraldehyde dehydrogenase (*lobster*), thermolysin (*B. amyloliquefaciens*), thioredoxin (*E. coli*), adenylate kinase (*porcine*), alcohol dehydrogenase (*equine*), algal ferredoxin (*S. platensis*), carbonic anhydrase $B$ and $C$(*human*), carboxypeptidase A and $B$(*bovine*), concanavalin A (*jack bean*), cytochromes: $C$(*albacore*), $B5$(*bovine*), $C2$(*R. rubrum*), $C551$(*P. aeruginosa*), $B562$(*E. coli*), dihydrofolate reductase (*L. casei*), elastase (*porcine*), flavodoxin (*clostridium*), hemerythrin $B$(*P. gouldii*), hemoglobin A and $B$(*equine*), lactate dehydrogenase (*dogfish*), lysozyme (*chicken*), subtilisin inhibitor (*S. alborgriseolus*), superoxide dismutase (*bovine*), trypsin inhibitor (*bovine*), chymotrypsin $\alpha$(*bovine*), papain (*papaya*), and subtilisin (*B. amyloliquefaciens*).

Treating the continuous, transformed measures as predictors showed negligible linear intercorrelations, with the exception of a strong negative relationship between between $\overline{hh}$ and $\omega_{hh}$ ($r = -0.805$)(the more dense the $\geq 2.17$, hydrophobic bursts, the higher the average hydrophobicity). Since these measures were redundant with respect to $d_S$, two alternative regression models predicting $d_S$ were constructed incorporating $\omega_{hh}$ in one and $\overline{hh}$ in the other. Using standardized regression coefficients (*i.e.* $\beta$'s), $d_S$-predictive Model I is $[-.246\omega_{hh} - .420\lambda_{word} - .111\%(PRO)]$ and Model II is $[+.221\overline{hh} - .403\lambda_{word} - .429\%(PRO)]$. Model I resulted in a squared multiple $R = 0.340$(adjusted $R^2 = 0.274$) and a highly significant ANOVA $[F(3,31) = 5.661, p = 0.003]$. Similarly, for Model II, $R^2 = 0.354$(adjusted $R^2 = 0.291$) and an ANOVA of $[F(3,31) = 5.332, p = 0.004]$. These findings are consistent with our hypothesis that the values for $d_S$ computed upon $x$-day crystallographic data from protein tertiary structure can be predicted from suitable transformations of the primary amino acid hydrophobicity sequences of the proteins. That $\lambda_{word}$ has a strong negative weighting with respect to $d_S$ suggests that the simple "fractal" interpretation[1-3] of $d_S$ is insufficient.

## REFERENCES


1. Stapleton, H.J., Allen, J.P., Flynn, C.P., Stinson, D.G., and Kurtz, S.R. *Phys. Rev. Lett.* **45**, 1456 – 1459(1980).

2. Allen, J.P., Colvin, J.T., Stinson, D.G., Flynn, C.P. and Stapleton, H.J. *Biophys. J.* **38**, 299 – 310(1982).

3. Wagner, G.C., Colvin, J.T., Allen, J.P., and Stapleton, H.J. *J. Am. Chem. Soc.* **107**, 5589 – 5594(1985).

4. Chothia, C. *Nature*, **254**, 304 – 308(1975).

5. Nozaki, Y. and Tanford, C. *J. Biol. Chem.* **246**, 2211 – 2217.

6. Manavalan, P. and Ponnuswamy, P.K. *Nature* **275**, 673 – 674(1978).

7. Franconi, R. and Finegold, I. *Science* **130**, 158 – 160(1975).

8. Brown, K.G., Erfurth, S.C., Small, E.W. and Peticolas, W.L. *Proc. Natl. Acad. Sci. USA* **69**,1467 – 1469(1972).

9. Sturtevant, J.M. *Proc. Natal. Acad. Sci. USA* **74**,2236 – 2210(1977).

10. Supplemental data for reference # 3 above obtained from Professor Harvey J. Stapleton, Department of Physics, University of Illinois, 1110 W. Green Street, Urbana, *IL*61801.

11. Newhouse, S., Palis, J., and Takens, F. *Publ. I.H.E.S***57**, 5 – 72(1983).

12. Jimenez-Montano, M.A. *Bull Math. Biol.* **46**,611 – 659(1984).

13. Erdos, P. and Renyi, A. *J. Anal. Math.* **22**,103 – 111(1970).

ACKNOWLEDGEMENTS. This work was partially supported by the (USA) Office of Naval Research (Divisions of Theoretical Biophysics,and Biological Intelligence). Appreciation is expressed to the Institut des Hautes Etudes Scientifiques, 91440 Bures-sur-Yvette, France, for its hospitality during the data development and writing of this paper.