

COMPUTATIONAL APPROACHES TO UNDERSTANDING THE PROTEIN
STRUCTURE

by
PELİN AKAN

Submitted to Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabanci University

July 2002

© Pelin Akan 2002

ALL RIGHTS RESERVED

ABSTRACT

This thesis is composed of two different parts, aiming to predict and understand the protein structure from their contact maps. In the first part, residue contacts of a protein are predicted using neural networks in order to obtain structural constraints for the three-dimensional structure. Physical and chemical properties of residues and their primary sequence neighbors are used for the prediction. Our predictor can predict 11% of the contacting residues with a false positive ratio of 2% and it performs 7 times better than a random predictor.

In the second part, a new method is developed to model a protein as a network of its interacting residues. Small-world network concept is utilized to interpret the parameters of residue networks. It is concluded that proteins are neither regular nor randomly packed but between these two extremes. Such a structure gives the proteins the ability for fast information relay between their residues. They can undergo necessary conformational changes for their functions on very short time scales. Also, residue networks are shown to obey a truncated power-law degree distribution instead of being scale-free. This shows that proteins have fewer structurally weak points, whose failure would be total damage for the system. This finding conforms to evolutionary plasticity of proteins: Having a low number of weak points makes the mild DNA mutations to be translated into the protein structure as highly tolerable.

ÖZET

Bu tez çalışmasında, proteinlerin temas matrisleri kullanılarak yapıları tahmin edilmeye ve anlaşılmaya çalışılmıştır. İki bölümden oluşan bu tezin ilk bölümünde, sinir ağları kullanılarak, proteinler için yapısal sınırlamalar bulmak amacıyla residü temasları tahmin edilmiştir. Bu tahminler için residülerin fiziksel ve kimyasal özellikleri, ve birincil sekanstaki komşuları kullanılmıştır. Sonuç olarak, birbiriyle temas eden residülerin % 11'i doğru, temas etmeyen residülerin % 2'si yanlış tahmin edilmiştir, ve rastlantısal bir tahminden 7 kat daha iyi sonuçlar elde edilmiştir.

İkinci bölümde, bir proteini, temas eden residülerinden oluşan bir ağ olarak modellemek için yeni bir yöntem geliştirilmiştir. Bu ağların yapısal özelliklerini anlayabilmek için küçük-dünyalar fikri kullanılmıştır. Gösterilmektedir ki, residüler proteinler içinde ne düzgün ne de rastlantısal bir şekilde organize edilmiştir, küçük-dünya ağlarına benzer bir organizasyona sahiptirler. Böyle bir yapı, proteinleri çok kısa zamanlar dahilinde büyük yapısal değişimler geçirebilmesini olanaklı kılmaktadır. Ayrıca, residü ağlarının komşu sayısı dağılımları da kesik ölçeksiz dağılımlar şeklindedir. Bu da proteinlerin çok az sayıda yapısal hassas noktalar içerdiğini göstermektedir. Proteinlerin evrim sürecinde sayısız biyolojik işlevi gerçekleştirebilecek şekildeki değişimleri bu sonucu desteklemektedir. Bunun nedeni,, az sayıda hassas noktanın varlığı küçük DNA mutasyonlarının proteinlerinin yapısına yansımaya olanak sağlamasıdır.

ACKNOWLEDGEMENTS

I owe my deepest and sincere thanks to Assoc. Prof. Canan Baysal. I consider myself lucky to have met such an intelligent and dynamic person at the right time in my education. She contributed a great deal of time and energy to this thesis and has become a constant source of advice, support and love.

Special thanks goes to my dearest, Murat Kaymaz, whose love and friendship have been essential for my success.

TABLE OF CONTENTS

1. INTRODUCTION	1
2. PREDICTION OF CONTACTING RESIDUES IN PROTEINS USING NEURAL NETWORKS	3
2.1 Overview	3
2.2 What Are Artificial Neural Networks?.....	6
2.2.1 Training.....	8
2.2.2 Multilayer Perceptron: A NN architecture.....	8
2.2.3 Learning Algorithm	12
2.2.4 Learning and Generalization.....	12
2.2.5 Complexity of the Network	14
2.3 Description of the Problem and the Solution Model.....	14
2.3.1 Input and Output of the NN.....	15
2.3.1.1 Surface Area.....	16
2.3.1.2 Hydrophobicity	16
2.3.2 Contact Definition.....	19
2.3.3 Datasets	20
2.3.4 NN Architectures	21
2.3.4.1 Network 1 (N1)	21
2.3.4.2 Network 2 (N2)	21
2.3.4.3 Network 3 (N3)	22
2.3.4.4 Network 4 (N4)	23
2.3.5 Evaluation of the Network Performance	24
2.4 Results and Discussions	26
2.4.1 Experiment 1	27
2.4.2 Experiment 2.....	28
2.4.3 Experiment 3.....	29

2.4.4	Experiment 4.....	30
2.4.5	Test Results.....	31
3.	PROTEINS AS NETWORKS OF THEIR INTERACTING RESIDUES	34
3.1	Overview	34
3.2	A Closer Look at Small-World Networks	38
3.2.1	Characteristic Path Length (L).....	41
3.2.2	Clustering Density (C).....	42
3.2.3	Degree Distribution.....	43
3.3	Network Model for Proteins.....	49
3.3.1	Protein Network Generation	49
3.3.2	Random Network Generation	50
3.3.3	Protein Network Generation Using DT	51
3.3.4	Calculation of L	52
3.3.5	Calculation of C	52
3.3.6	Degree Distribution.....	53
3.3.7	Radial Distribution Function	53
3.4	RESULTS AND DISCUSSION.....	54
3.4.1	Radial Distribution Function	54
3.4.2	Scaling of L	55
3.4.3	L in Actual and Random Networks.....	58
3.4.4	Clustering Coefficient in Actual and Random Networks	59
3.4.5	Degree Distribution.....	61
4.	CONCLUSIONS	65
4.1	NN Predictor for Contacting Residues	65
4.2	Characterization of Residue Networks	67
	REFERENCES	70
	APPENDIX.....	74

LIST OF TABLES

Table 2.1. Surface area and hydrophobicity features before re-scaling.....	18
Table 2.2. Residue features after re-scaling.....	19
Table 2.3. Performance of N1 on the validation dataset.....	27
Table 2.4. Performance of N2 on the validation dataset.....	29
Table 2.5. Performance of N3 on the validation dataset.....	30
Table 2.6. Performance of N4 on the validation dataset.....	31
Table 2.7. The performances of the best networks on the test dataset	33
Table 3.1.Examples of small-world behavior; $L \geq L_{random}$ but $C \gg C_{random}$	43
Table 3.2. Parameters of L vs $\log(N)$ plot in Figure 4.3.	56

LIST OF FIGURES

Figure 2.2.1.A Biological Neuron.	7
Figure 2.2.2. One processing unit of an artificial NN (neuron).	9
Figure 2.2.3. Layer of S number of neurons operating in parallel.	10
Figure 2.2.4. Linearly separable patterns.	11
Figure 2.2.5. Multilayer perceptron architecture	11
Figure 2.2.6. Mean squared error in training and validation phases.	13
Figure 2.3.1. Architecture of N1 and N2	22
Figure 2.3.2. N3 architecture	23
Figure 2.3.3. N4 architecture for a pair of residue i and j	26
Figure 3.1.1. A residue network of generated at 7 Å.	37
Figure 3.1.2 Another representation of a residue network at 7 Å.	37
Figure 3.2.1. The transition from regular to random regime in a simple topology 40	40
Figure 3.2.2. Calculation of clustering coefficient of i^{th} vertex in a network.	42
Figure 3.2.3. Degree distribution of random and small-world networks.	45
Figure 3.2.4. Physical constraints on $P(k)$	47
Figure 3.3.1. Construction of DT from a set of points.	51
Figure 3.4.1. Radial distribution function of C_{β} atoms.	55
Figure 3.4.2. L versus protein lengths.	56
Figure 3.4.3. Scaling of L with protein length.	57
Figure 3.4.4. Scaling of L versus N in networks generated by DT	58
Figure 3.4.5. L in actual and random networks.	59
Figure 3.4.6. C in actual and random networks.	60
Figure 3.4.7. Average $P(k)$ of residue networks generated at 7 Å.	62
Figure 3.4.8. Log-log plot of $P(k)$ at three different cutoff radii.	63

LIST OF SYMBOLS

C_{α}	Central carbon atom attached to a hydrogen, an amino group, a carboxyl group and the side chain group in an amino acid
C_{β}	Side chain carbon atom bonded to C_{α} atom of a residue
\AA	Angstrom

LIST OF ABBREVIATIONS

<A>	Average accuracy of a neural network
All pr.	All proteins in the dataset
<i>C</i>	Clustering density
<i>C. elegans</i>	<i>Caenorhabditis elegans</i>
CC	Correctly predicted contacting residues by the neural network
COF	A protein dataset comprising 225 proteins
COLD	Constrained optimization with limited deviations
DT	Delaunay Triangulation
FP	Non-contacting residues predicted as contacts by the neural network
HOT	Highly optimized tolerance
<i>L</i>	Characteristic path length
LRN	A protein dataset comprising 196 proteins
N1	Neural network 1 architecture
N2	Neural network 2 architecture
N3	Neural network 3 architecture
N4	Neural network 4 architecture
NN	Artificial neural network
R	Improvement of the prediction over a random predictor
TS97	A protein dataset comprising 176 proteins
WWW	World Wide Web