

## INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

**The quality of this reproduction is dependent upon the quality of the copy submitted.** Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

# UMI

A Bell & Howell Information Company  
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA  
313/761-4700 800/521-0600



**INVESTIGATION INTO PROTEIN FOLDING PREDICTION OF  
HELICES USING TECHNIQUES IN COMPUTER SCIENCE**

A Dissertation

by

NEAL ANDREW KRAWETZ

Submitted to the Office of Graduate Studies of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 1998

Major Subject: Computer Science

**UMI Number: 9830944**

---

**UMI Microform 9830944  
Copyright 1998, by UMI Company. All rights reserved.**

**This microform edition is protected against unauthorized  
copying under Title 17, United States Code.**

---

**UMI**  
300 North Zeeb Road  
Ann Arbor, MI 48103

**INVESTIGATION INTO PROTEIN FOLDING PREDICTION OF  
HELICES USING TECHNIQUES IN COMPUTER SCIENCE**

A Dissertation

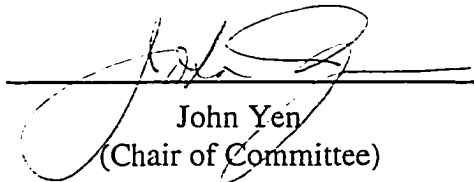
by

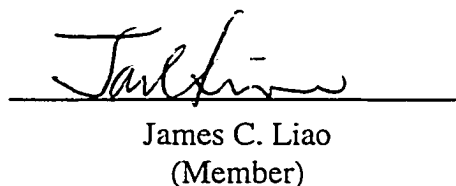
NEAL ANDREW KRAWETZ

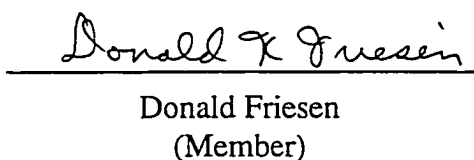
Submitted to the Office of Graduate Studies of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

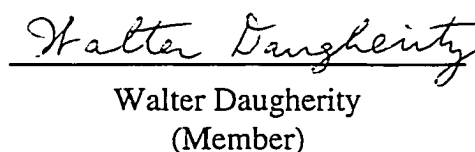
DOCTOR OF PHILOSOPHY

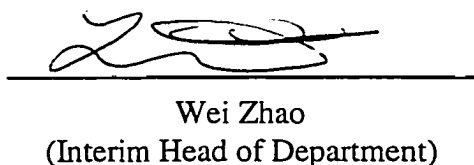
Approved as to style and content by:

  
John Yen  
(Chair of Committee)

  
James C. Liao  
(Member)

  
Donald Friesen  
(Member)

  
Walter Daugherty  
(Member)

  
Wei Zhao  
(Interim Head of Department)

May 1998

Major Subject: Computer Science

## ABSTRACT

Investigation into Protein Folding Prediction of Helices Using Techniques in  
Computer Science. (May 1998)

Neal Andrew Krawetz, B.A., University of California, Santa Cruz.

Chair of Advisory Committee: Dr. John Yen

The research presented in this dissertation focuses on the application of computer science techniques in the field of theoretical biochemistry. This interdisciplinary study analyzes current black-box neural network systems and applies information from the analysis into a novel step-wise (white-box) Bayesian prediction system with heuristic refinement that provides insight into the prediction and performs comparably with existing prediction models.

This research studies the prediction process that determines a protein's helices from the primary amino acid sequence. Existing neural network prediction systems are analyzed and some of the factors that the systems consider important in the prediction process are identified. This information is then applied in a step-wise Bayesian prediction system and refined using heuristics that incorporate high-level knowledge of the helix structure. The result is a white-box prediction system that is at least as accurate as the black-box system and provides insight into the prediction process.

The Bayesian prediction system used in this research focuses on the prediction of helices by using region-specific, position-dependent helical propensities. Because variations in local amino acid sequences determine whether a helix is formed, we infer that each amino acid has explicit preferences toward specific regions (N-, C-

terminals, and middle) in helices. These region-specific, position-dependent propensities appear to correlate with spatial organization along the helix wheel. Furthermore, the statistical analysis indicates that the helix propensities are conditionally independent for structural determination. Using this information, a statistical approach is proposed for determining helix locations from known sequences of amino acids. Incorporating high-level helical patterns with knowledge-based postprocessing provides a novel step-wise approach to helix location identification.

To David McCall, Tom Yohe, Radford Stone, George Newall, and their gang,  
for their influence.



## ACKNOWLEDGMENTS

I would like to thank Dr. Liao and my committee members for their feedback and support, without whom this research would not have been possible. Additionally, Dr. Pooch and Dr. Childs provided support and advice in my times of need. My sincerest thanks go to Jeff Cloyd for introducing me to this field of study and for exposing me to the protein folding problem. Additionally, I thank all the people who happily listened to my babble and nodded their heads at the appropriate times: Ryan Wood, Ellen Mitchell, Willis Marti, Shridhar Muppidi, Chris Cunningham, Badari Devalla, and the rest of the System Support Office staff. I should also mention my parents (Sharon and Howard), my siblings (Julie and Bruce), and my best friend Michelle “Kat” Karako for their support. And of course, Grandma, with her wonderful advice: “Whatever you do, do it good.” And my everlasting gratitude to Michelle Mach, M.L.S., for her words of wisdom, encouragement, and exquisite bibliographic skills.

Finally, I would like to mention the people whose brains I happily picked for information: Andrew Dalke, Barry Isralewitz, Matthew Wander, Rosemary Braun, Jim Phillips, and Dorina Kosztin at the Department of Theoretical Biophysics, University of Illinois, Urbana-Champaign, and Dr. Tom Ioerger and Dr. Nick Pace of Texas A&M University.

## TABLE OF CONTENTS

	Page
ABSTRACT .....	iii
ACKNOWLEDGMENTS .....	vi
TABLE OF CONTENTS .....	vii
LIST OF FIGURES .....	ix
LIST OF TABLES .....	xi
I. INTRODUCTION .....	1
A. The problem and motivation .....	1
B. Objectives and overview of the research .....	3
C. Organization of the dissertation .....	4
II. REVIEW OF RELATED WORK .....	6
A. Probabilistic methodology .....	6
B. Protein structure .....	7
C. Neural networks .....	21
D. Protein folding prediction systems .....	25
III. DATA COLLECTION .....	29
A. Data sets commonly incorporated by prior models .....	29
B. Nonhomologous data set incorporating mutagenesis .....	30
C. Data set implementation .....	32
IV. ACCURACY METRICS .....	34
A. Overall correctness: $Q_1$ .....	34
B. Combined correctness: $Q_3$ .....	35
C. Over-prediction: $Q_{predict}$ .....	35
D. Under-prediction: $Q_{observe}$ .....	36
E. Correlation coefficient: $C_{coef}$ .....	36
V. MODIFICATIONS TO THE NEURAL NETWORK SYSTEM .....	37
A. Null amino acid representation .....	37
B. Weight matrix analysis .....	38
C. Association with amino acid properties .....	53
D. Implicit assumptions and limitations in the neural network system .....	55

	Page
VI. PROBABILISTIC ANALYSIS REGARDING HELIX PROPENSITIES . . . . .	57
A. Definitions . . . . .	57
B. Computing likelihoods . . . . .	59
C. Testing the conditional independence assumption . . . . .	62
D. Analyzing probabilistic information regarding helix formation propensities . . . . .	64
VII. HELIX PATTERN METHODOLOGY . . . . .	71
A. Thresholded helix prediction definition . . . . .	71
B. Knowledge-based refinement schemes . . . . .	72
C. Helix pattern methodology implications . . . . .	78
VIII. IMPLEMENTATION OF BAYESIAN INFERENCE SYSTEM WITH KNOWLEDGE-BASED POSTPROCESSING . . . . .	79
A. Regional probability computation . . . . .	79
B. Implementation of Bayesian inference system with heuristic refinement . . . . .	80
C. Comparison with other prediction models . . . . .	84
IX. CONCLUSION AND FUTURE RESEARCH . . . . .	88
A. Conclusion on basic neural network with sliding window . . . . .	88
B. Significance of the research . . . . .	89
C. Insight into the Bayesian prediction system . . . . .	90
D. Limitations of the Bayesian inference system and areas for future research . . . . .	93
REFERENCES . . . . .	95
APPENDIX . . . . .	100
A. Data sets . . . . .	100
B. Conditional probabilistic propensities . . . . .	114
VITA . . . . .	127

## LIST OF FIGURES

Figure	Page
1. Atomic structure of the amino acid backbone. $C_\alpha$ denotes the $\alpha$ -carbon which connects to the sidechain, R. . . . .	8
2. The 20 common amino acid sidechains. Along with the atomic configuration, the amino acid's common name, abbreviated name, and single letter representation are provided. . . . .	9
3. Histogram of helix lengths indicating the most common sizes of helices. . . . .	14
4. Direction of helix dipole. The N-terminal of the helix is positively charged and the C-terminal is negatively charged. The amount of charge differential varies for each helix. . . . .	15
5. Spatial positioning along the helix wheel. A: the positions of adjacent amino acids in the primary sequence. B: the helix spiral around the wheel. . . . .	16
6. Hydrogen bonding between turns in the helices of protein 1CRN (Teeter, 1984), as displayed by Rasmol (Sayle, 1994). The thin lines denote hydrogen bonds. The amino acid positions, $i \in [0, 11]$ , have been labeled along one of the helices. . . . .	17
7. Example of a sheet from the lysozyme mutant, 1L30 (Alber et al., 1988), as displayed by Rasmol (Sayle, 1994). The sheet is highlighted along with its hydrogen bonds. Strands 3 and 4 form a parallel sheet while strands 1, 2, and 3 form antiparallel sheets. . . . .	19
8. The basic perceptron model with $n$ inputs, $x_0$ through $x_n$ , and one output $y'$ . The inputs are scaled by the weight matrix, $w_0$ through $w_n$ and combined by the perceptron. The desired output, $y$ , is only used during the training of the weight matrix. . . . .	22
9. Perceptron with floating threshold, $t$ . . . . .	23
10. A basic NETtalk-based prediction system with sliding window used to predict helices. After predicting Leu in the context of Asn-Glu-Ala-Leu-Pro-His-Asp, the system will slide the window to predict Pro in the context of Glu-Ala-Leu-Pro-His-Asp-Gly. . . . .	25

11. A trained neural network weight matrix, ordered by window position. The system threshold is 0.04. ....	40
12. Clusters of helix forming and helix breaking positions around the helix wheel. Areas indicate the number of amino acids with weights above the threshold, as determined by the neural network system. The helix wheel appears divided into four distinct regions: two helix forming regions and two helix breaking regions. ....	41
13. The effect of suppressed positions (holes) within the input window, measured with the metric $Q_I$ . Holes representing the suppression of one, two, and three amino acids are presented. ....	43
14. The effect of suppressed positions (holes) within the input window. The results are measured with the $C_{coef}$ metric. Holes representing the suppression of one, two, and three amino acids are presented. ....	44
15. The effect of suppressed positions (holes) within the input window as presented by the $Q_{observe}$ metric. Holes representing the suppression of one, two, and three amino acids are presented. ....	45
16. The effect of suppressed positions (holes) within the input window, using the $Q_{predict}$ metric. Holes representing the suppression of one, two, and three amino acids are presented. ....	46
17. The effects of variable asymmetrical window sizes on the prediction accuracy of the sliding window neural network model. The $Q_I$ accuracy of the system shows the three-dimensional surface from variable window sizes on helix prediction. ....	50
18. The effects of variable asymmetrical window sizes on the prediction accuracy of the sliding window neural network model. A: the effects of variable N-terminal sizes on the $Q_I$ accuracy of the system. B: the effects of different C-terminal sizes on the prediction ....	51
19. Independence analysis comparing computed occurrence rate of amino acid pairs in helical regions with the observed frequency. The high degree of similarity between the observed and computed probabilities suggests that amino acid helical propensities combine independently. ....	63

20. The outliers from the independence analysis are due to infrequency between the computed pairs. The outliers do not support the hypothesis that amino acid helical propensities are a dependent interaction. The threshold values of 0.20, 0.50, and 1.00 represent 25%, 50%, and 90% of the data points. . . . . 64
21. The protein 1AAT as displayed by Rasmol (Sayle, 1994). The helix along residues 312-341 is highlighted and the bend at residues 319 (Asp) and 320 (Asn) is labeled. . . . . 91

## LIST OF TABLES

Table	Page
1. Common amino acid propensity measurements .....	11
2. Sample unique amino acid contexts over three proteins .....	31
3. Distribution of amino acids in data set .....	33
4. Comparison of null amino acid effectiveness .....	38
5. Effects of associated amino acid attributes on helix prediction .....	55
6. $P(r \in [N, M, C] \text{ at } 0 \mid \text{Proline at } i)$ over the window $i \in [-7, +7]$ .....	59
7. High-level knowledge-based rule set .....	75
8. Effectiveness comparison of heuristics .....	81
9. Comparison of prediction models .....	85
10. Example of helix prediction: 1AAT (Torchinsky et al. 1982) .....	92

## I. INTRODUCTION

The research presented in this dissertation focuses on the application of computer science techniques in the field of theoretical biochemistry. This interdisciplinary study analyzes current black-box neural network systems and applies information from the analysis into a novel step-wise (white-box) prediction system which provides insight into the prediction and performs comparably with existing prediction models.

### A. The problem and motivation

A protein's shape and configuration determines how it interacts with structures such as DNA, RNA, and other proteins. The protein's form depends on the sequences of amino acids which make up the protein. The amino acids are chained together in a primary sequence. These amino acids interact with other parts of the primary sequence, folding the protein into a complex bonded structure. A biochemical research problem attempts to determine the final shape of the protein from the primary sequence. To resolve this problem, biochemists and biophysicists use a variety of approaches, including measured observations of the stabilized protein, molecular dynamics simulators to model the chemical interactions, and prediction systems to develop a best-guess with regards to protein structures.

Physical observation of the folded protein is one of the most reliable approaches used to determine protein configuration. X-ray crystallography is used to create a three-dimensional electron density map of the entire protein. Difficulties may arise when large globular proteins partially occlude internal regions, or when proteins do not crystallize easily. Furthermore, the electron density map may be ambiguous or difficult to interpret. The entire process, from crystallizing the protein to the final

---

The journal model followed in this dissertation is *Protein Science*.



interpretation of the electron density map, may take anywhere from a few months to a few years.

Many proteins are derived from a common source, either through evolutionary mutations from a single protein or by common configurations for required interactions. These protein families share common sequences and structures. There are an estimated 1,500 protein families, although the available data contains only about 120 families (Chothia, 1992). Because of the vast number of unknown protein families and the large amount of time required to “observe” single proteins, theoretical techniques are used to provide a best-guess of the protein structure.

Proteins are composed of 20 basic amino acids<sup>1</sup>. These amino acids have known atomic configurations and are readily identifiable. Through the use of known molecular dynamics (MD), simulators can be used to model the forces acting on each atom and determine the final configuration. Although faster than X-ray crystallography, current MD simulators still require weeks or months of computer time.

The three-dimensional structure of the protein is not always essential to determine the protein’s general shape and function. The amino acids are known to form any of three stable secondary structures: helices, sheets, and turns. By identifying likely secondary structure locations, protein folding prediction systems offer a good best-guess of the protein configuration in a few seconds to a few minutes. Although not as accurate as MD simulators or X-ray crystallography, a reliable prediction system can be used to identify likely protein structures, speed up MD simulations, and reduce ambiguity from electron density maps, as well as determine likely structure locations in newly identified amino acid sequences.

---

<sup>1</sup>Although other rare compounds, such as hydroxylproline and sarcosine, may be included in some primary sequences, the vast majority of primary sequences consists of the basic 20 amino acids.

## B. Objectives and overview of the research

The common protein folding prediction approaches are easily classified into white-box and black-box systems. White-box prediction systems allow for a clear view of the factors involved in the prediction and identify how these factors influence the prediction process. White-box protein folding prediction systems are generally designed around *a priori* information that is considered important in the folding process. Chou and Fasman (1974), for example, identified hydrophobicity as a significant factor in determining secondary structures; the Chou-Fasman algorithm is designed to combine hydrophobicity measurements.

Black-box protein folding prediction systems use little or no *a priori* structural information in the prediction process. Although the previous black-box systems generally perform better than the white-box systems, there has been little research done to identify what physical factors are important to the prediction system.

The objectives of this research are summarized as follows:

1. *Design a novel approach for providing more data to the black-box and white-box systems.* Currently, 120 protein families account for nearly 20,000 amino acids, which can be used to train a black-box system without introducing biases from a specific family. We propose a novel approach for incorporating an additional 20,000 amino acids from variances in related proteins without introducing specific family biases.
2. *Analyze and extend a classical black-box (neural network) prediction system.* Analyzing the information stored within the neural network allows the determination of factors that the system considers significant in the prediction process.
3. *Based on the analysis of the black-box system, design a novel step-wise prediction system of comparable, or better, performance.* By using the key information identified by the black-box system, a white-box system can be created with comparable performance to the black-

box systems. In this work, a probabilistic approach is used to combine helical likelihoods. The resulting Bayesian inferences are refined using heuristics based on helical structure.

4. *Correlate results with findings from other disciplines.* Although the probabilistic information used to determine the secondary structure prediction was not determined from *a priori* structural information, the information does correlate with known atomic, amino acid, and helical information.
5. *Provide novel insight into the prediction process.* Through correlation with known structures and a step-wise prediction system, insight determining the essential factors of the folding process, and how they relate, is obtainable.

To limit the complexity of the system and research, this work focuses on the prediction of only one secondary structure: helices. Prediction of other secondary structures and of the tertiary and quaternary configurations are beyond the scope of this work.

### **C. Organization of the dissertation**

Chapter I discusses the overview and motivation behind this work, and its objectives and scope. Chapter II summarizes related studies in the areas of protein science, probability and Bayesian inference, neural network systems, and protein folding prediction models. An explanation of the data set used and data collection method is provided in Chapter III. Chapter IV discusses the metrics used to compare prediction systems.

Chapter V analyzes a black-box neural network protein folding prediction system. In Chapter VI, the probabilistic helical propensities are defined and analyzed. Chapter VII discusses the helix pattern methodology used by the novel prediction system. In Chapter VIII, the results from an implementation of the Bayesian inference

system with knowledge-based postprocessing are presented. Chapter IX summarizes the finding of this research and suggests areas for future work.

## II. REVIEW OF RELATED WORK

This chapter reviews the fundamental methodologies, concepts, and technologies used by this research. Section A reviews the basic principles regarding probabilities and Bayesian inference used in Chapter VI as the basis of the prediction system. Section B discusses basic protein structure, including amino acid composition, secondary structures, the tertiary and quaternary interactions, and folding methodology. Section C provides a brief overview of related knowledge-based systems. Chapter VII uses similar knowledge-based systems to combine high-level information about protein structure to enhance the prediction process. Neural network systems and the classical sliding window approach are reviewed in Section D and analyzed in Chapter V. The final section reviews current protein folding prediction systems, illustrating the current state of the art.

### A. Probabilistic methodology

Bayes' theorem (Laplace, 1812) is used to convert the knowledge of  $P(A|B)$  into a calculation of  $P(B|A)$ :

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}. \quad (1)$$

By assuming that events combine independently, we can substitute  $P(A)$  with

$$\begin{aligned} P(A) &= P(A, B) + P(A, \neg B) \\ &= P(A|B) \times P(B) + P(A|\neg B) \times P(\neg B), \end{aligned} \quad (2)$$

and derive the conditionally independent form of Bayes' theorem:

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A|B) \times P(B) + P(A|\neg B) \times P(\neg B)}. \quad (3)$$

When used with protein folding prediction, the prior probabilities,  $P(B)$  and  $P(\neg B)$ , represent the occurrence rates of the secondary structure being predicted. The conditional probability,  $P(A|B)$ , generally represents the occurrence rate of an amino acid given a known structure; this occurrence rate is readily available. The posterior probability,  $P(B|A)$ , denotes the desired structural prediction from the amino acids. The implementation of Bayes' theorem with respect to the protein folding prediction problem is discussed in Chapter VI.

## **B. Protein structure**

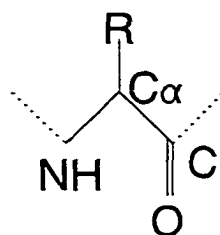
Proteins consist of repeating substructures called amino acids. These amino acids define a known atomic configuration. The folding and final shape of a protein depends on the interactions of the atoms in the amino acids (Branden & Tooze, 1991).

### *1. Amino acid configuration*

The amino acids consist of two components: the backbone and the sidechain. While each amino acid has the same backbone structure<sup>2</sup>, there are different sidechains which define the unique amino acids. The backbone, as shown in Figure 1, is viewed from the nitrogen terminal (N) to the carbon terminal (C). The sidechain, R, branches from the  $\alpha$ -carbon ( $C_\alpha$ ). The backbones of neighboring amino acids connect from N-terminal to C-terminal, forming a single chain referred to as the protein backbone. The protein backbone is not a rigid structure; it can bend into coils or loops, forming complex structures.

---

<sup>2</sup>Proline is the only amino acid with a different backbone structure due to a sidechain that connects directly to the backbone.



**Fig. 1.** Atomic structure of the amino acid backbone.  $C_{\alpha}$  denotes the  $\alpha$ -carbon which connects to the sidechain, R.

The composition of the sidechain determines the amino acid. There are 20 different types of sidechains which are common. Although rare mutations to the sidechain structure exist, analysis of these amino acids is beyond the scope of this research due to their rarity. The 20 different amino acids are shown in Fig. 2. These amino acids are frequently referred to by their common names, abbreviated names, or single-letter representations. For example, alanine may be referred to as “Ala” or “A”.

#### a. Primary sequence

The *primary sequence* of a protein lists the amino acids along the protein backbone, from the N-terminal to the C-terminal, without regards to three-dimensional structure. Proteins are known to fold in consistent patterns; a protein which is stretched into a single long chain will, over time, refold into its lowest-energy state. Because the primary sequence describes the covalent bonds of the atoms in the protein, it is hypothesized that the primary sequence contains all the necessary information for determining the final folded structure.

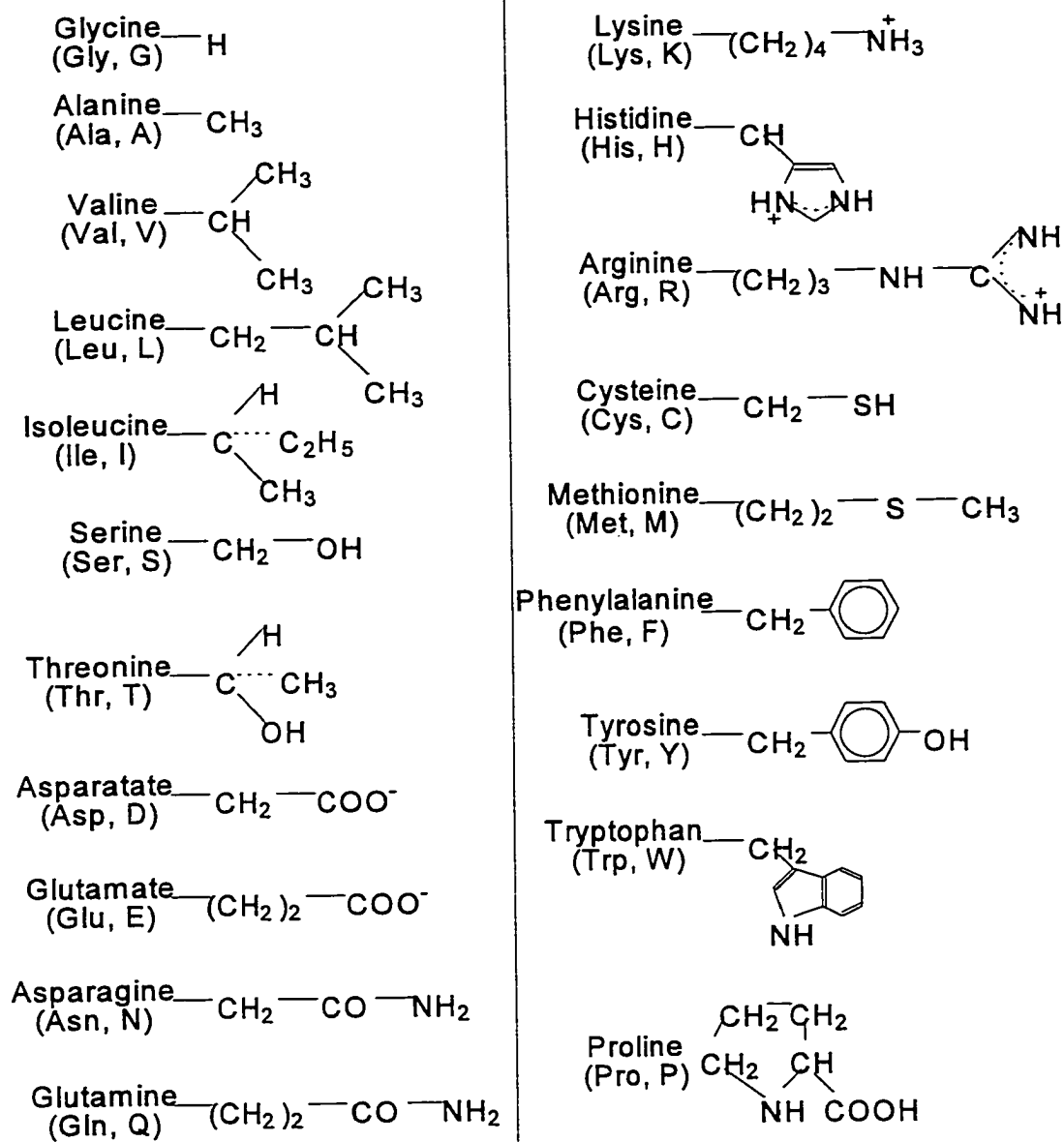


Fig. 2. The 20 common amino acid sidechains. Along with the atomic configuration, the amino acid's common name, abbreviated name, and single letter representation are provided.



## b. Classical propensities

Each sidechain connects to the backbone at the  $\alpha$ -carbon, allowing for the sidechain to rotate freely. The exception to this is proline, where the sidechain connects to the backbone in two places, forming a ring. Although the backbone is positively charged at the N-terminal and negatively charged at the C-terminal, most of the physical attributes of the amino acids are due to interactions with the sidechains rather than with the backbone.

It is commonly hypothesized that the amino acids within the primary structure, without external influences, interact to form protein structures. There are many forces involved in determining the interactions. The measurement of these forces can be directly linked to each amino acid. An amino acid's interaction *propensity* is a measurement of the specific amino acid's contribution in the interaction. Propensity measurements, by definition, are unique to the type of amino acid and are not influenced by adjacent amino acids in the primary sequence. Table 1 lists some of the more common propensity scales.

**Table 1.** *Common amino acid propensity measurements*

Amino Acid	Van der Waals volume ( $\text{\AA}^3$ )	Hydrophobicity (Chou & Fasman, 1974)	Hydrophobicity (Fauchere & Pliska, 1983)	Hydropathy (Kyte & Doolittle, 1982)
Ala (A)	67	1.45	0.42	1.8
Arg (R)	148	0.79	-1.38	-4.5
Asn (N)	96	0.73	-0.82	-3.5
Asp (D)	91	0.98	-1.05	-3.5
Cys (C)	86	0.77	1.34	2.5
Gln (Q)	114	1.17	-0.30	-3.5
Glu (E)	109	1.53	-0.87	-3.5
Gly (G)	48	0.53	0.00	-0.4
His (H)	118	1.24	0.18	-3.2
Ile (I)	124	1.00	2.46	4.5
Leu (L)	124	1.34	2.32	3.8
Lys (K)	135	1.07	-1.35	-3.9
Met (M)	124	1.20	1.68	1.9
Phe (F)	135	1.12	2.44	2.8
Pro (P)	90	0.59	0.98	-1.6
Ser (S)	73	0.79	-0.05	-0.8
Thr (T)	93	0.82	0.35	-0.7
Trp (W)	163	1.14	3.07	-0.9
Tyr (Y)	141	0.61	1.31	-1.3
Val (V)	105	1.14	1.66	4.2

Each atom consists of neutrons and protons, surrounded by a shell of electrons. The larger the atom, the larger the electron shell. When atoms bond, their electron shells merge, allowing the electron shells around the molecule to cover an area around the entire molecule. The Van der Waals volume, which measures the size of each atom's electron shell, describes the size of each amino acid's sidechain.<sup>3</sup> Although this measurement is not commonly used as a propensity, it is frequently used in molecular dynamics simulators when determining atomic interactions.

There are many different hydrophobicity scales used for amino acid propensities. Because amino acids are generally studied while in aqueous solution, the interaction between the sidechain and the ambient medium becomes a powerful force. Amino acids that are hydrophilic tend to move toward the ambient solution, while hydrophobic amino acids are generally repelled from water molecules. Consequently, protein structures commonly have clustered regions of amino acids that are strongly hydrophilic or hydrophobic.

### c. $\Phi$ and $\Psi$ angles

The covalent bonds to the carbons in the backbone are fairly rigid; these bonds are at a fixed angle of  $100^\circ$  from the N-terminal to the  $\alpha$ -carbon to the C-terminal, and form a backbone plane. Although the backbone plane is fairly rigid, the N- $C_\alpha$  and  $C_\alpha$ -C bonds can rotate, allowing the sidechain's position to vary in relation the backbone plane. The angle of rotation along the N- $C_\alpha$  bond is referred to as the  $\phi$  angle; the rotation of the  $C_\alpha$ -C bond determines the  $\psi$  angle. These angles control the amount of rotation along the backbone and physically limit the secondary structure formation. Although every amino acid can form any N- $C_\alpha$ -C angle, some angles are

---

<sup>3</sup>The Van der Waals force describes the repulsive property of an atom. When many atoms are in a close proximity to each other (pressure), they position themselves in a lowest energy configuration. The Van der Waals volume describes the minimum energy distance between atoms.

more common for certain amino acids.

## 2. *Secondary structures*

Localized regions of amino acids interact to form basic secondary structures: helices, sheets, turns, and coils. The helices, sheets, and turns are defined by bonding angles. Coils represent a lack of bonding. A coil is a free-floating structure that is generally unrestricted in configuration, unlike the other secondary structures that are considerably more rigid. It is through these secondary structures that the protein shape is given stability.

### a. Helices

This research focuses on helix identification and information content as determined from the primary sequence. When the protein backbone winds into a spring-like structure and is held by hydrogen bonding between amino acids, helices are formed. Helices are defined by specific formations and have consistent physical characteristics, including the telltale spring-like structure, the directional dipole, and amino acid positioning along the helix wheel. Some factors, such as Ncaps, Ccaps, and helix bundles are known to affect the stability of a helix.

#### (1) Types of helices

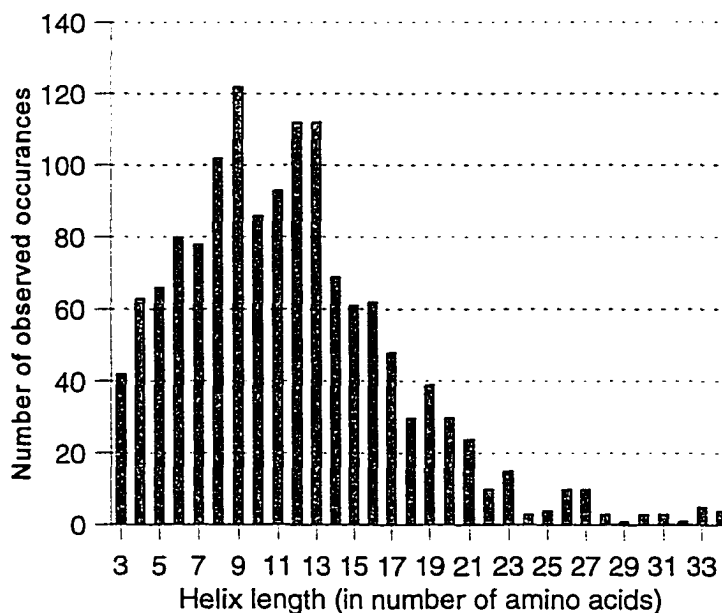
Amino acids are involved in helices about 35% of the time.<sup>4</sup> There are ten different helix formations, although some are only theoretical and have not yet been

---

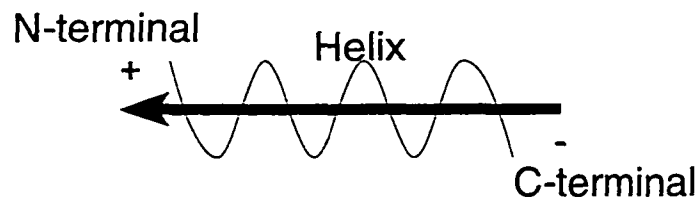
<sup>4</sup>The percentage of amino acids in helices was determined from the data set, defined in Chapter III. The structural occurrence rates from the data set are similar to other published findings (Qain & Sejnowski, 1988; Muskal & Kim, 1992; Rost & Sander, 1993b).

observed. The most common helix is the right-handed  $\alpha$ -helix. This helix has a tight circular loop with 3.6 amino acids per turn. When viewed through the shaft of the helix, the  $\alpha$ -helix appears circular. Other helices include  $3_{10}$ -helices which have 3 amino acids per turn and 10 atoms per hydrogen bond, and  $\pi$ -helices with 4.4 amino acids per turn. Unlike the  $\alpha$ -helix, the narrow shaft in  $3_{10}$ -helices causes atoms to be close together, generating high Van der Waals forces (repulsion between atoms). Thus,  $3_{10}$ -helices are generally not a lowest-energy formation and are only found in extreme circumstances (Creighton, 1993). The  $\pi$ -helices, with their wide shafts, may be flexed and are usually unstable.

For the focus of this research, all helices are considered right-handed  $\alpha$ -helices. These helices account for more than 95% of the currently observed helices. Although this is a weak assumption for helices in general, the techniques used in this research are not dramatically affected by 5% noise in the data. Helices vary in length from three amino acids to more than 30 amino acids; most helices are six to 13 amino acids in length, containing two to four loops, as shown by the histogram in Figure 3.



**Fig. 3.** Histogram of helix lengths indicating the most common sizes of helices.



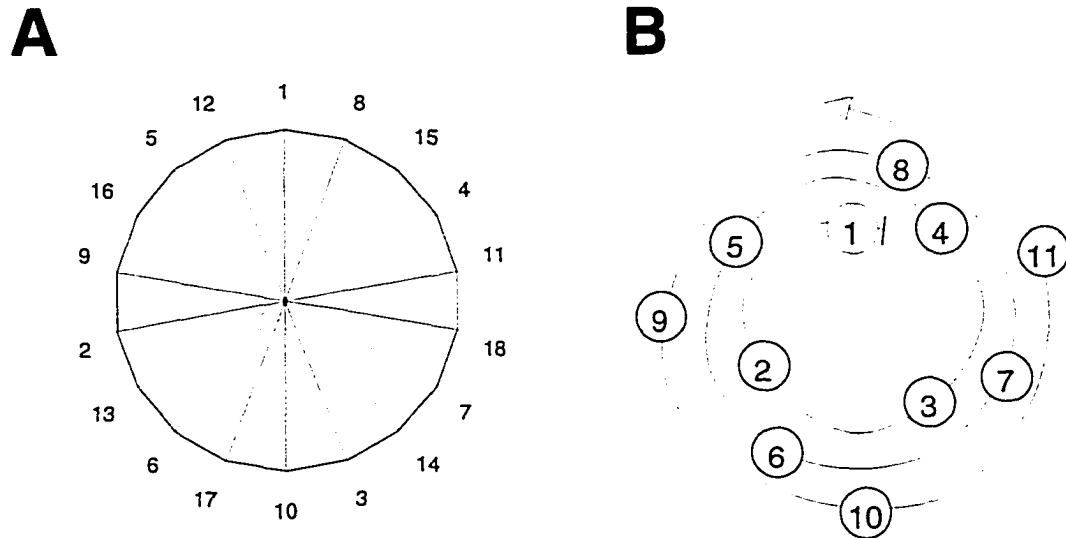
**Fig. 4.** Direction of helix dipole. The N-terminal of the helix is positively charged and the C-terminal is negatively charged. The amount of charge differential varies for each helix.

### (2) Helix dipole

The *dipole moment* is an asymmetrical electrostatic charge with a measurable magnitude and specific direction. Every atom, molecule, and compound has a dipole moment. The dipole of a helix runs along the length of the helix. It is found near the center of the helix, running from the C-terminal to the N-terminal, as illustrated in Figure 4. The N-terminal of the helix has a positive charge, while the C-terminal is negatively charged. The amount of charge variation differs with each helix.

### (3) Helix wheel

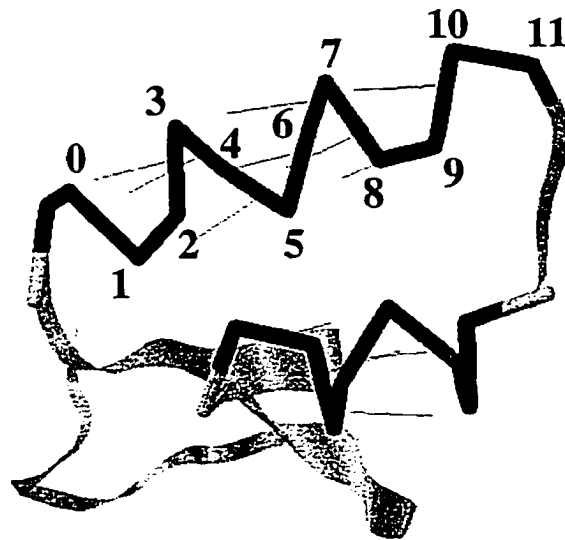
Amino acids are spaced along the length of the helix at regular intervals. There are approximately 3.6 amino acids per turn in the helix and the amino acid backbones are bent at an average of 100 degrees. Each amino acid can interact with its physically adjacent neighbors. As shown in Figure 5, the amino acid at position 5 in the helix can interact vertically, along the length of the helix with amino acids in positions 1, 9, 12, and 16, as well as with the adjacent amino acids along the primary sequence, in positions 3, 4, 6, and 7 (and possibly further). The vertical amino acid interactions along the helix provide stability by forming hydrogen bonds which hold the helix loops together. The rise along the helix is  $1.5\text{\AA}$  per amino acid, or  $5.4\text{\AA}$  per turn.



**Fig. 5.** Spatial positioning along the helix wheel. **A:** the positions of adjacent amino acids in the primary sequence. **B:** the helix spiral around the wheel.

#### (4) Helix stability

Helices are held in stable formations by the hydrogen bonding between vertically aligned amino acids. In  $\alpha$ -helices, hydrogen bonds occur every 13 atoms along the backbone. As shown by the helix wheel, the amino acids at position  $i$  can form hydrogen bonds with the amino acids at positions  $i+3$  or  $i+4$  (Fig. 6). These hydrogen bonds hold the backbone in the spiral configuration.



**Fig. 6.** Hydrogen bonding between turns in the helices of protein 1CRN (Teeter, 1984), as displayed by Rasmol (Sayle, 1994). The thin lines denote hydrogen bonds. The amino acid positions,  $i \in [0, 11]$ , have been labeled along one of the helices.

#### (5) Helix bundles and hydrophobicity

Helix bundles occur when multiple helices interact with each other (Holm & Sander, 1993). The helices become aligned and reinforce each other's stability. Frequently, helix bundles are amphipathic, having one side of the helix wheel hydrophobic and the other side hydrophilic. The similar hydrophobic regions of the helices become adjacent and interact to increase the helices' stability (Kamtekar & Hecht, 1995).

#### (6) Ncaps and Ccaps

The first and last few amino acids in a helix have been identified as having an extreme influence on the formation of helices (Aurora et al., 1994, Aurora & Rose, 1998). These regions, referred to as the Ncap and Ccap, are suspected of being significant in the formation of helix terminals. Helix capping (Aurora & Rose, 1998)



extends the hypothesis that helix conformation is “specified by a stereochemical code, similar to DNA where strand complementary is determined by hydrogen bonds” (Aurora et al., 1994).

#### b. Sheets

Sheets consist of nonsequential regions in the primary sequence which bond into a plane (Fig. 7). There are two types of strands that make up sheets: parallel and antiparallel. In parallel strands, the N-terminus<sup>5</sup> of one sequence is aligned with the N-terminus of the bonded sequence. Antiparallel strands align the N-terminus of one sequence with the C-terminus of the bonded sequence. The stability of the sheet may depend on all amino acids in each strand; changing a single amino acid may break all of the bonds in the sheet.

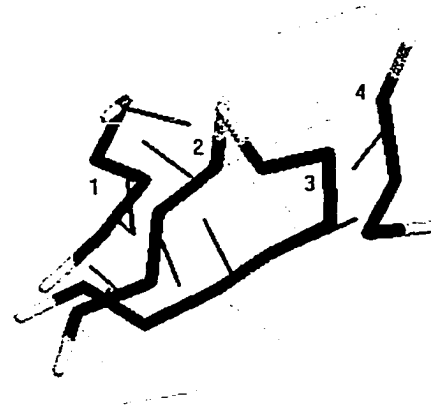
#### c. Turns and coils

The remaining secondary structures are turns and coils. Turns consist of three or four amino acids which form a strong bond, effectively bending the primary sequence into a tight curve. Unlike other secondary structures which are physically independent, turns may overlap with the terminal amino acids in helices and sheets.

Coils, also referred to as “random coils” or “loops,” denote the absence of helices, sheets, and turns over a region of the protein. In general, there is little or no hydrogen bonding along coils, giving the region little stability.

---

<sup>5</sup>A “terminal” is a specific atom, position, or amino acid. A “terminus” is in the direction of the terminal.



**Fig. 7.** Example of a sheet from the lysozyme mutant, 1L30 (Alber et al., 1988), as displayed by Rasmol (Sayle, 1994). The sheet is highlighted along with its hydrogen bonds. Strands 3 and 4 form a parallel sheet while strands 1, 2, and 3 form antiparallel sheets.

### *3. Tertiary and quaternary interactions*

The secondary interactions of a primary sequence are capable of influencing each other. This influence is referred to as the tertiary interaction. These interactions include helix rotations and alignment with other helices and sheets, and the curving of sheets into beta barrels. For example, two amphipathic helices, denoted by having distinct hydrophobic and hydrophilic sides, prefer to align so that their hydrophobic sides are adjacent. Due to stresses formed by tertiary interactions, some secondary structures may be reshaped or broken; tertiary interactions are thought to be the primary cause of  $3_{10}$ -helices (Creighton, 1993).

In addition to the tertiary interactions, an entire protein chain may be influenced by other adjacent chains, forming quaternary interactions. Due to the high-

level complexity of quaternary interactions, few molecular dynamics simulators or structure prediction systems consider these influences.

#### *4. Minimum energy folding states*

The final shape of a protein is a minimal energy state. When energy is added to the system by temperature, pressure, or motion, bonds and structures become unstable. The final shape of a protein minimizes the energy loss in the system. Molecular dynamics modeling attempts to simulate the system until it achieves the final stable configuration. Other approaches attempt to define the folding process into distinct energy states. Structural prediction systems assist these approaches by providing a best-guess as to the initial configuration.

##### a. Molecular dynamics modeling

Molecular dynamics modeling (MD) determines the stable protein structure by modeling all atoms in the system and all atomic interactions. Programs such as XPLOR and NAMD model between 10,000 and 500,000 atom systems, including all bonding interactions. Due to the required small time step (around one picosecond) and the incredible number of interactions to monitor, these systems generally run for weeks or months on high-end workstations and supercomputers.

##### b. Two-stage folding process

Boczko and Brooks (1995) use molecular dynamics modeling to illustrate a two stage folding process. In the first stage, secondary structures rapidly form. In particular, secondary structures over sequential amino acids, such as helices, are constructed. They attribute this to local minimal energy along the primary sequence. Since a helix is a minimal energy structure covering a sequential area of amino acids,

helices seem to appear before sheets or turns. After forming the local minimal energy state (many helices), the tertiary interactions proceed to “unfold” the local minimal energy structures as more optimal energy formations are created. Helices may be modified or broken during this process, altering a local minimum to a higher energy state in order to stabilize the entire systems energy. This second “unfolding” stage takes significantly longer than the initial “folding” stage.

The two-stage folding process directly relates to this research. The prediction system described in Chapter VI and VII overpredicts helices. This overprediction can be explained by the two-stage folding process: the prediction system appears to determine the results of the first folding stage.

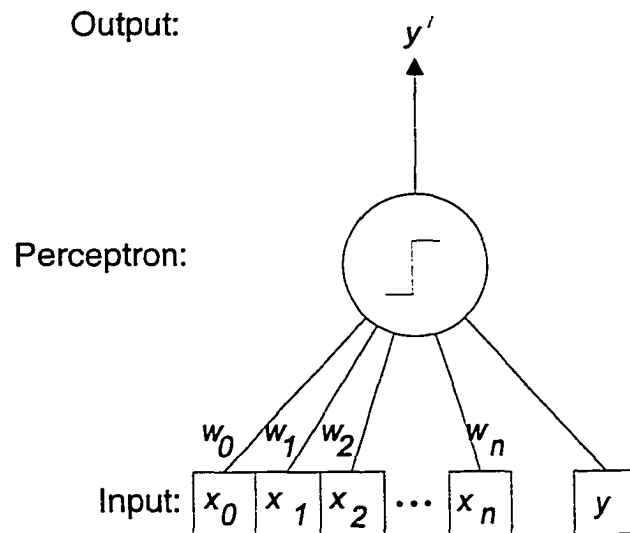
### C. Neural networks

In contrast to Bayesian probabilistic systems, a neural network determines the mapping from the input space to the output space from a learned transformation matrix. Because of the transformation complexity between the primary sequence and secondary structures, neural networks have been used to “learn” the transformation.

#### 1. Perceptron

The basic perceptron (Fig. 8) compares the sum of a set of inputs ( $x_0, x_1, \dots, x_j$ ) weighted by their importance, or weight ( $w_0, w_1, \dots, w_j$ ), to a threshold  $t$  (Hecht-Nielsen, 1989; Hertz et al., 1991). The single output of the system,  $y'$ , represents the results of the comparison:

$$y' = \begin{cases} 1 & \text{if } \sum_{i=0}^n w_i x_i \geq t \\ 0 & \text{if } \sum_{i=0}^n w_i x_i < t. \end{cases} \quad (4)$$



**Fig. 8.** The basic perceptron model with  $n$  inputs,  $x_0$  through  $x_n$ , and one output  $y'$ . The inputs are scaled by the weight matrix,  $w_0$  through  $w_n$  and combined by the perceptron. The desired output,  $y$ , is only used during the training of the weight matrix.

More generalized perceptrons use the raw summation value or apply sigmoidal functions to the system output rather than applying a strict threshold,  $t$ .

During the training phase, the desired output,  $y$ , is supplied to the system. The weights are then adjusted based on the difference between the predicted and desired outputs, and scaled by the learning rate,  $\eta$ :

$$w' = w + \eta \cdot x(y - y'). \quad (5)$$

## 2. Neural network floating threshold

A single perceptron computes a sum of products of the inputs,  $i_x$ , and the weight assigned to the input,  $w_x$ . The resulting value,  $o_x$ , may be scaled using a sigmoid or other limiting function, but is finally compared to a single-value threshold. To best optimize the system, the perceptron uses a floating threshold,  $t$ . Rather than treating the threshold as a fixed constant ( $t^\wedge$ ),

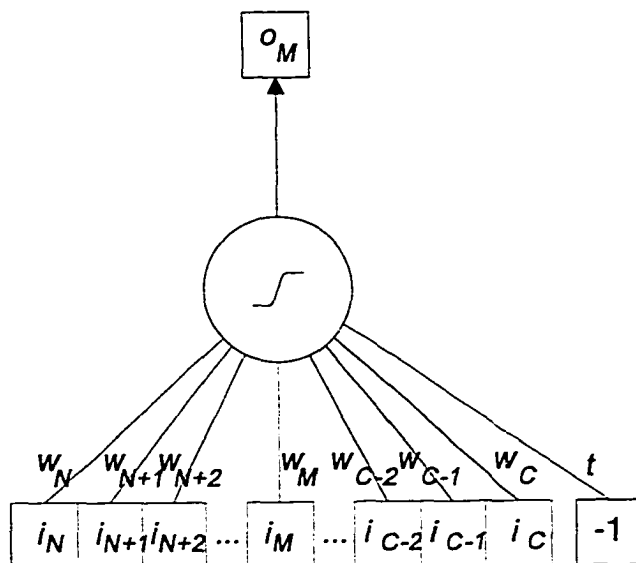


Fig. 9. Perceptron with floating threshold,  $t$ .

$$\sum_{x=n}^c i_x w_x \geq t'; \quad (6)$$

the threshold is handled like another weight in the system, but assigned the constant input value of  $-1$ :

$$\left( \sum_{x=n}^c i_x w_x \right) + (-1) \times t \geq 0. \quad (7)$$

The value of the floating threshold,  $t$ , is modifiable by the neural network during the training stage. The resulting output produced by the neural network is summed with the floating threshold and compared with the constant value zero (Fig. 10).

### 3. NETtalk: basic neural network sliding window model

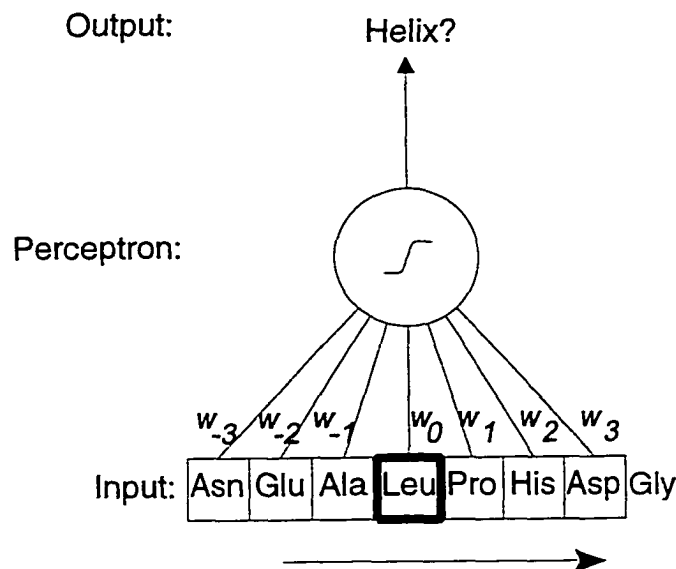
The simplest approach to protein folding prediction with neural networks uses

a NETtalk-based approach (Qain & Sejnowski, 1988; Holley & Karplus, 1989; Muskal & Kim, 1992). The NETtalk system (Sejnowski & Rosenberg, 1986) uses a neural network system to determine the pronunciation of English words. This system uses an input window of sequential letters, spaces, and punctuation, and predicts which of 30 phonetic sounds should be produced. Each perceptron is trained independently and the perceptron with the strongest prediction generates the sound.

A system based on the NETtalk neural network correlates the pronunciation of English letters with the structures of a protein. The linear sequence of English letters in a sentence resembles the primary sequence of amino acids. Similarly, the pronunciation of each letter corresponds with the secondary structures; there are a finite number of pronunciations for each letter, depending on the adjacent letters. The tertiary interaction of proteins is similar to the same series of letters having different sounds. For example, interactions within the semantic context alter the pronunciation of the letters "READ" (rēd or rĕd): "I will *read* it" (rēd) or "I have *read* it" (rĕd). Similarly, the primary sequence Arg-Glu-Ala-Asp-Ile-Asn-Gly, "READING" using the symbols in Table 1, is found predominantly in helices, but also in sheets and coils. The structure of this sequence is dependent on interactions with other secondary structures.

The protein's quaternary interactions, while playing a significant role in some protein formations, are similar to the subtle changes in pronunciation between related sentences of a paragraph. Due to the complexity of the quaternary interactions and the unknown degree of influence they have on a system, secondary structure prediction systems generally disregard the quaternary effects.

In the NETtalk-based systems, a window of the primary sequence is used as an input to a perceptron then predicts the most likely secondary structure (Fig. 10). After each prediction, the window is moved to cover a different section in the same primary sequence. The predicted structure is associated with the amino acid in the middle of



**Fig. 10.** A basic NETtalk-based prediction system with sliding window used to predict helices. After predicting **Leu** in the context of Asn-Glu-Ala-**Leu**-Pro-His-Asp, the system will slide the window to predict **Pro** in the context of Glu-Ala-Leu-**Pro**-His-Asp-Gly.

the window.

#### D. Protein folding prediction systems

The best way to determine a proteins structure is to observe it, usually with an electron microscope or X-ray crystallography. Unfortunately, observing the protein is difficult, time-consuming, and expensive. Protein folding prediction systems attempt to determine the secondary, tertiary, or quaternary structures without actually observing the system.

There are four main approaches to protein folding prediction: statistical models, homology and motif based systems, artificial intelligence, and molecular dynamics simulation.



## *1. Homology and motif based prediction systems*

Homology is defined as the quality of similarity between protein sequences.

As described by Reeck et al. (1987):

In its precise biological meaning, “homology” is a concept of quality. The word asserts a type of relationship between two or more things. Thus, amino acid or nucleotide sequences are either homologous or they are not. They cannot exhibit a particular “level of homology” or “percent homology.” Instead, two sequences possess a certain level of similarity. Similarity is thus a quantitative property. Homologous proteins or nucleic acid segments can range from highly similar to not recognizably similar (where similarity has disappeared through divergent evolution).

Approaches using homology compare a primary sequence with unknown secondary structure with a similar primary sequence that has a known secondary structure. The hypothesis for homology modeling assumes similar primary sequences have similar structures.

Along the same lines as homology, motifs denote small patterns of amino acids with known likely structures. Motifs, such as “AAxAR is a helix” where ‘x’ represents any amino acid, are mostly correct, but a large number of motifs are required for general application.

## *2. Statistical prediction models*

Statistical prediction models use observed occurrences of amino acid interactions to estimate likely structures. These systems include classical propensity models and probabilistic systems.

Statistical systems using classical amino acid propensities are generally designed around the propensity. For example, Chou and Fasman (1974) considered hydrophathy as a key propensity. As a result, the Chou-Fasman algorithm combines

windows of measured hydropathy indices as a means to determine secondary structures. The combinational method is based on a hypothesis concerning the propensity's interactions.

In contrast, probabilistic Bayesian networks (Delcher et al., 1993; Klinger & Brutlag, 1994; Koretke et al., 1996) first determined dependent sequence positions and then apply the observed occurrence rates of amino acids matching the combined pattern. In probabilistic models there is generally no correlation made between the observed occurrences and classical propensities.

### *3. Artificial intelligence models*

Models using artificial intelligence (AI) generally focus on neural networks and extend the statistical approach. Rather than manually specifying the propensities or the combination method, perceptrons (or similar systems) are used to determine the important combination patterns and the optimal propensities. These systems are generally more accurate than statistical, homology, or motif models in predicting general protein structure, but they are frequently criticized for providing no insight into the prediction process. Neural network prediction systems are accurate black-boxes. This is in contrast with the statistical or homology approaches where the factors determining the prediction are readily identifiable.

### *4. Molecular dynamics simulators*

Molecular dynamics (MD) simulators can be used to predict secondary and tertiary formations. Through MD modeling, the simulators can approximate the position of each atom in the protein. This process is slow and time-consuming; modeling a moderate protein with 200,000 atoms can take weeks or months of simulation time. Additionally, simulation factors may bias the result; simulated folding in a vacuum is faster than simulating an aqueous solution, but may not be as

accurate. Other biases may stem from force fields (holding the aqueous solution in a bubble), torus mapping (an atom which moves too far “up” will appear at the “bottom” of the simulated environment), or cumulative floating point errors in the computations.

More often, MD simulators are used to test the stability of a theoretical protein configuration. Through the use of prediction techniques, an approximate initial configuration can be identified. The MD simulators are then used to model the predicted shape and determine whether it is stable or unstable.

### III. DATA COLLECTION

This chapter discusses the two common types of protein data sets used by prediction models. In addition, a novel approach is described that combines the two types of data sets for use with this research.

All proteins used by this research, and nearly all proteins studied by related prediction models, are publicly available from the Brookhaven Protein DataBank, PDB (Abola et al., 1987). Protein entries in the PDB contain publication references, the primary sequence, secondary structure locations, and three-dimensional atomic coordinates for each (non-hydrogen) atom in the proteins. Some of the proteins in the PDB are theoretical or have been determined by simulations, although most have been identified by X-ray crystallography or nuclear magnetic resonance spectroscopy (NMR).

#### A. Data sets commonly incorporated by prior models

Historically, two types of data sets have been used in protein folding prediction models: homologous and nonhomologous protein data sets. A homologous protein data set is used when small changes in the primary sequence are desirable. Models which use homologous data, such as Chou-Fasman (1974) and Kyte-Doolittle (1982), measure the differences between similar proteins. It is hypothesized that the observed changes between similar proteins are due to the differences in the primary sequence. Thus, a nonhomologous data set would provide too many differences, leading to ambiguity in the measured results.

Mutagenesis can aid the homologous-based systems. For example, there are over 100 mutagenic variances to lysozyme (Matthews et al., 1973) available in the PDB. Using these proteins, accurate measures of hydrophobicity and specific amino acid interactions can be made and compared with the specific differences in the primary sequence.

The alternate type of data set, containing nonhomologous proteins, is generally used by statistical and AI-based approaches. Homologous proteins, by definition, are more similar than different. When homologous proteins are used in a statistical model, the repetition of similar structure biases likelihoods. Neural network systems memorize the similar configurations and treat the differences as noise, rather than generalizing the data. Thus, homologous data sets are not very useful for these systems due to the large amount of repeated data. Nonhomologous protein data sets remove redundant sequences, effectively eliminating a source of bias.

The PDB contains far more homologous proteins than nonhomologous. While there is an ample supply of similar proteins for small measurements based on specific differences in the primary sequence, there are only about 250 nonhomologous proteins (Chothia, 1992) containing around 20,000 atoms available for systems which are biased by homology.<sup>6</sup>

### **B. Nonhomologous data set incorporating mutagenesis**

While nonhomologous data sets are applicable to statistical and AI-based prediction systems, the small differences in homologous proteins are explicitly ignored. A novel approach for generating a larger data set suitable for statistical and neural network-based systems incorporates both nonhomologous proteins and nonhomologous regions of similar proteins.

Many of the previous approaches use homologous proteins (Chou & Fasman, 1974; Kyte & Doolittle, 1982). While these proteins are somewhat different, they are in the same family. Unfortunately, related proteins generally have related structures and can bias a probabilistic data set. Other approaches use nonhomologous proteins (Qain & Sejnowski, 1988; Rost & Sander, 1993a) from the PDB (Abola et al., 1987)

---

<sup>6</sup>In October, 1992, the PDB contained 1007 protein structures. By January, 1998, it had grown to contain nearly 7000 protein structures. The data presented in this research only reflects the PDB from October, 1992 (see Appendix A).

consisting of about 20,000 amino acids. In addition to nonhomologous proteins, we include nonhomologous segments from homologous proteins. For example, there are over 100 mutants of lysozyme (Matthews, 1973) in the PDB which differ only by a few amino acids. To avoid biasing the data set, only windows of seven amino acids containing a unique amino acid taken in the context of the primary sequence were included. Therefore, each sequence of seven amino acids in the collected data set is unique.

By incorporating unique amino acid contexts from homologous proteins, the data set size can be doubled without introducing additional bias.<sup>7</sup> In addition, repeated primary sequences from small identical regions found in the 143 nonhomologous proteins are removed. As the example in Table 2 illustrates, a data set containing nonhomologous proteins would only include 2CPP and 1L13. 1L18 would be dropped from the data set due to homology with 1L13. By incorporating the unique context found in 1L18, the example data set size increases by more than 10%.

**Table 2.** *Sample unique amino acid contexts over three proteins*

PDB entry	Residues 26-58
2CPP (Poulos et al., 1987)	<u>SAGVOEAWAVLOESNV<del>P</del>DLVWTRCNGGHWIAT</u>
1L13 (Alber et al., 1987)	<u>TIGIGHLLTKSPSLNAAKSELDKAIGRNCNGV</u>
1L18 (Alber et al., 1987)	<u>TIGIGHLLTKSPDLNAAKSELDKAIGRNCNGV</u>

<sup>7</sup>It is conceivable that hundreds of homologous proteins can differ by a single amino acid in a specific region. The repetition of the same context with a minor variance would bias statistical and neural network systems much the same way as entire homologous proteins would. Fortunately, the PDB currently contains no more than a few similar variants of the same region from homologous proteins.

### **C. Data set implementation**

The data set used in this research contains 483 proteins collected from the PDB. These account for nearly 100,000 amino acids, of which 40,363 amino acids are in unique contexts and 143 proteins are composed entirely of unique windows of seven amino acids. The proteins used in this research, and their unique context locations, are listed in Appendix A.

The data set has been divided into training and testing sets. The training set, used to compute the likelihoods for the system, contains 433 proteins, accounting for 31,582 unique contexts. The testing set contains the remaining 50 proteins which are composed entirely of unique contexts and which are not included in the computation of the likelihoods. The distribution of amino acids in secondary structures within the data sets (both training and testing sets) is provided in Table 3.

**Table 3.** *Distribution of amino acids in data set*

	Percent occurrence			Percent of total amino acids	
	helix	sheet	coil	total data set	nonhomologous data set
Ala (A)	48%	18%	34%	8.6%	8.7%
Arg (R)	44%	21%	34%	4.2%	3.2%
Asn (N)	30%	17%	53%	4.5%	4.5%
Asp (D)	35%	15%	51%	6.2%	5.6%
Cys (C)	28%	32%	41%	1.8%	1.5%
Gln (Q)	42%	24%	35%	3.6%	3.6%
Glu (E)	48%	16%	36%	5.9%	4.8%
Gly (G)	20%	21%	59%	8.3%	8.6%
His (H)	34%	22%	44%	2.1%	2.4%
Ile (I)	37%	37%	27%	5.5%	4.5%
Leu (L)	46%	27%	27%	8.1%	8.3%
Lys (K)	41%	20%	39%	6.3%	6.7%
Met (M)	48%	26%	27%	2.0%	1.5%
Phe (F)	35%	31%	34%	3.9%	3.8%
Pro (P)	17%	15%	69%	4.3%	4.5%
Ser (S)	30%	21%	49%	6.3%	7.7%
Thr (T)	27%	30%	44%	5.6%	6.4%
Trp (W)	37%	29%	33%	1.5%	1.5%
Tyr (Y)	33%	30%	37%	3.6%	3.5%
Val (V)	31%	40%	29%	7.0%	7.4%
Unk <sup>a</sup>	10%	10%	80%	5.2%	1.3%
<b>Total</b>	<b>35%</b>	<b>25%</b>	<b>40%</b>	<b>100.0%</b>	<b>100.0%</b>

<sup>a</sup> Unk accounts for unknown or uncommon residues in the data sets.



## IV. ACCURACY METRICS

In order to determine the effectiveness of the prediction process, a variety of accuracy metrics is employed. An overall correctness metric,  $Q_1$ , is used to provide a general utility measurement. This metric provides no specifics about the prediction's effectiveness, but rather summarizes the prediction performance. The  $Q_{predict}$  and  $Q_{observe}$  metrics are useful for determining over- and under-predictions. A correlation coefficient,  $C_{coef}$ , is used to determine the relationship of correct and incorrect predictions.

Predictions analyzed in this research determine the presence or absence of a secondary structure feature when provided with a portion of a primary sequence. A positive instance indicates that the feature is present; a negative instance indicates the absence of the feature. We define the terms correct positive (CP) and correct negative (CN) to indicate a correct prediction of the secondary structure when compared with the actual structures found in the PDB entry. Similarly, a false positive (FP) and false negative (FN) indicate incorrect predictions. A FP indicates an over-prediction; the feature is predicted to be present while it was observed to be absent. A FN indicates an under-prediction; the feature is predicted to be absent even though it was observed to be present.  $N$  represents the total number of predictions.

### A. Overall correctness: $Q_1$

The overall correctness metric,

$$Q_1 = \frac{CP+CN}{CP+CN+FP+FN} = \frac{CP+CN}{N}, \quad (8)$$

compares the total number of correct predictions to the total number of predictions ( $N$ ). This simple metric provides only a general summary of the prediction; it cannot

be used to determine weaknesses in the prediction process due to over- or under-prediction of features.

The  $Q_1$  metric is applicable only to single-structure prediction. Since related prediction models determine multiple secondary structures, no other prediction system uses this metric. We have duplicated the single neural network system of Qain and Sejnowski (1988) for use as a baseline for comparison purposes.

### B. Combined correctness: $Q_3$

Most prediction systems determine all secondary structures. The combined metric,

$$Q_3 = \frac{CP_\alpha + CN_\alpha + CP_\beta + CN_\beta + CP_c + CN_c}{N}, \quad (9)$$

(Qain & Sejnowski, 1988), similar to  $Q_1$ , provides a simple success rate for the combined helix ( $\alpha$ ), sheet ( $\beta$ ), and coil ( $c$ ) prediction. This metric is commonly used by prediction models to provide a general measurement of the prediction accuracy.

### C. Over-prediction: $Q_{predict}$

$Q_{predict}$  compares the number of correctly predicted amino acids in a structure with the total number of predictions:

$$Q_{predict} = \frac{CP}{CP + FP}. \quad (10)$$

This frequently-used metric determines the amount of over-prediction. It is unaffected when the system under-predicts a structure, but is influenced by the number of false-positive predictions.

#### D. Under-prediction: $Q_{observe}$

In contrast to  $Q_{predict}$ ,  $Q_{observe}$  measures the amount of under-prediction by comparing the number of correctly predicted amino acids in a structure with the total number of amino acids observed in the structure:

$$Q_{observe} = \frac{CP}{CP+FN}. \quad (11)$$

A prediction system which fails to identify observed helices performs poorly with this metric. Like  $Q_{predict}$ ,  $Q_{observe}$  is a commonly-used metric.

#### E. Correlation coefficient: $C_{coef}$

More complex than  $Q_1$ , the correlation coefficient (Matthews, 1975),

$$C_{coef} = \frac{CP \times CN - FP \times FN}{\sqrt{(CP+FP)(CP+FN)(CN+FP)(FP+FN)}}, \quad (12)$$

compares the number of correct predictions with the number of false predictions. In general,  $C$  determines the relationship between the occurrence of correct predictions and the absence of incorrect predictions. A high correlation indicates few false predictions, while a low correlation shows a high degree of error from the prediction system.

## V. MODIFICATIONS TO THE NEURAL NETWORK SYSTEM

The basic neural network model for protein folding prediction is based on the NETtalk system (Sejnowski & Rosenberg, 1986). This system showed that a neural network system can identify high-level structure (pronunciation) from a low-level pattern (the alphabet). Current protein prediction models based on this system have included some modifications to the basic system, but divulge little of the information stored within the prediction system. Consequently, many aspects of the NETtalk system, including implicit assumptions, have not yet been challenged, and little has been done to analyze the information content of the neural network system.

In this chapter the basic NETtalk system is modified in order to optimize the system and determine the factors important to the network. We introduce the concept of a null amino acid to maintain a constant input norm for the perceptron. The information content, describing both information location and importance, is derived from the weight matrix analysis. Methods for analyzing the weight matrix include “poking holes” in the input window, denying inputs, and applying variable window sizes. Enhancements to the inputs are tested by associating amino acid properties with the input window. Multiple output systems, such as majority prediction systems, are discussed. The final section of this chapter summarizes the implicit assumptions and limitations of the NETtalk-based approach.

### A. Null amino acid representation

The input size of the prediction system is generally consistent: a seven amino acid window maintains seven active system inputs. This consistent input size gives the system a constant norm over the input vector. But when the input window contains the protein terminals, the norm changes since fewer inputs (amino acids) are available for the window. The net result is that all amino acids are treated with equal input weights except the amino acids near the protein terminals. This change in input

**Table 4.** Comparison of null amino acid effectiveness

	$Q_I$	$Q_{predict}$	$Q_{observe}$	$C_{coef}$
Tested on total proteins				
with Nul	65.92%	60.44%	58.10%	0.350
without Nul	65.78%	58.05%	60.01%	0.339
Tested on terminals only				
with Nul	73.35%	97.78%	62.92%	0.742
without Nul	72.33%	96.90%	64.62%	0.694

weight can effectively confuse a single perceptron.

We introduce the concept of a null amino acid (Nul) to maintain a consistent input window size. When no amino acid is available for a window position, Nul is used as a placeholder in the window position. The null amino acid plays a similar role to the “space” letter in the NETtalk system.

As illustrated in Table 4, the null amino acid has little effect on overall protein prediction, but significantly improves the terminal regions. This is because protein terminals, where Nul is used, account for about 12.5% of the entire data set. When tested strictly on the protein terminal regions, the prediction accuracy increases noticeably. Nul correlates with increases in  $Q_I$ ,  $C_{coef}$ , and  $Q_{predict}$ , while causing a minor drop in  $Q_{observe}$ . The loss of  $Q_{observe}$  accuracy appears to be due to a more conservative prediction from the system.

Unless noted otherwise, all neural network systems mentioned in the remainder of this chapter implement the null amino acid.

## B. Weight matrix analysis

The weight matrix of the neural network stores the information learned by the system. Unfortunately, many input factors influence the weight matrix, including

magnitude, quantity, and type. These factors are combined within the weight matrix, making analysis of the learned information difficult. As a means to identify distinctive elements and determine their significance within the weight matrix, specific influencing factors have been identified and withheld from the system. By isolating these factors, the effects on the network's performance can be observed.

Two specific factors that affect the prediction system have been identified and examined. Using "windows with holes" suppresses a specific position of the input window. This allows the importance of a specific input position to be observed. Alternately, the use of asymmetrical windows identifies the influence of neighboring positions and the secondary structures which may be learned by the system.

We hypothesize that different positions within the sliding window have different degrees of importance. Under this hypothesis, it is conceivable that an amino acid at specific position in the sliding window has a stronger influence on structural determination than an amino acid at a different position, regardless of the amino acid types. This hypothesis is tested in two ways. First, the raw weight matrix of a trained neural network system is graphed, equating the weight magnitudes with the window positions. Second, each position within the window is suppressed and the resulting network accuracy is compared with the unsuppressed system. For testing these hypotheses, a neural network with a 15 amino acid input window is trained on helix structures.

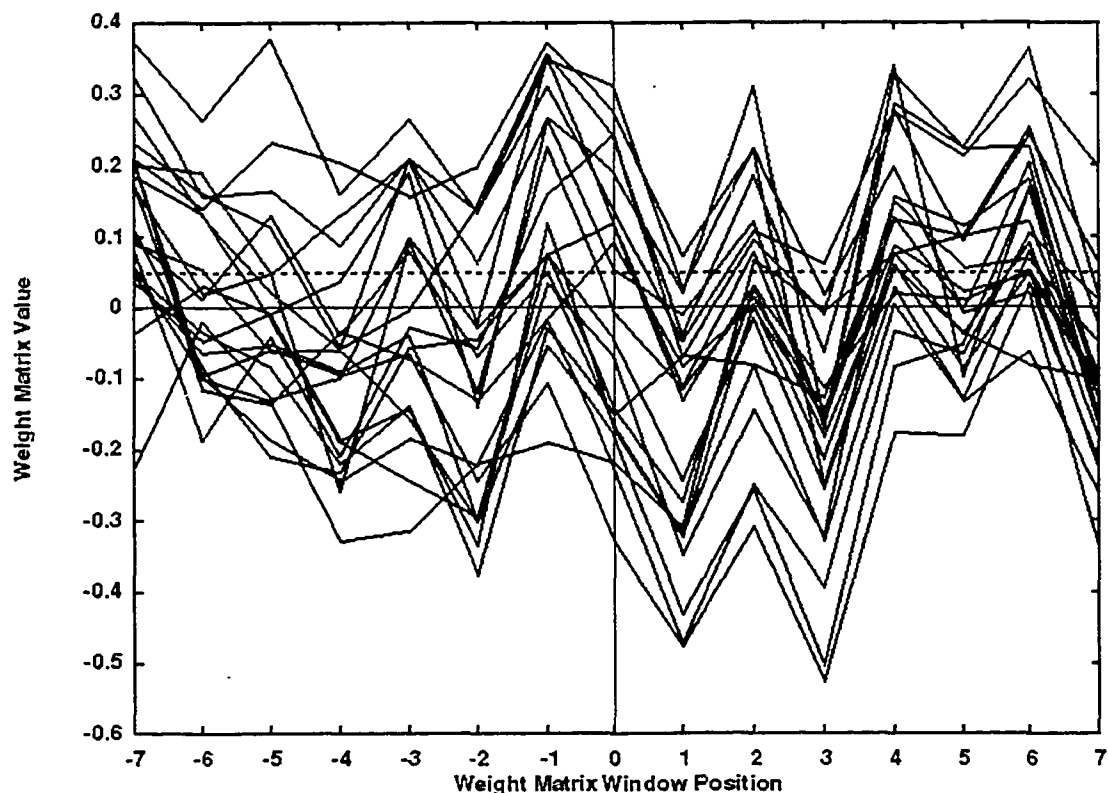
### *1. Weight matrix position analysis*

Figure 11 shows average weight matrix values for each window position. The system was trained with the null amino acid and an input window of 15 amino acids. The system only predicts helices: each position is predicted as a helix or not-helix. Each line in the figure represents one amino acid, but the individual amino acids are not labeled. While the raw value of the weight matrix is unimportant (it can be scaled without affecting the system) the relative values indicate the degree of importance

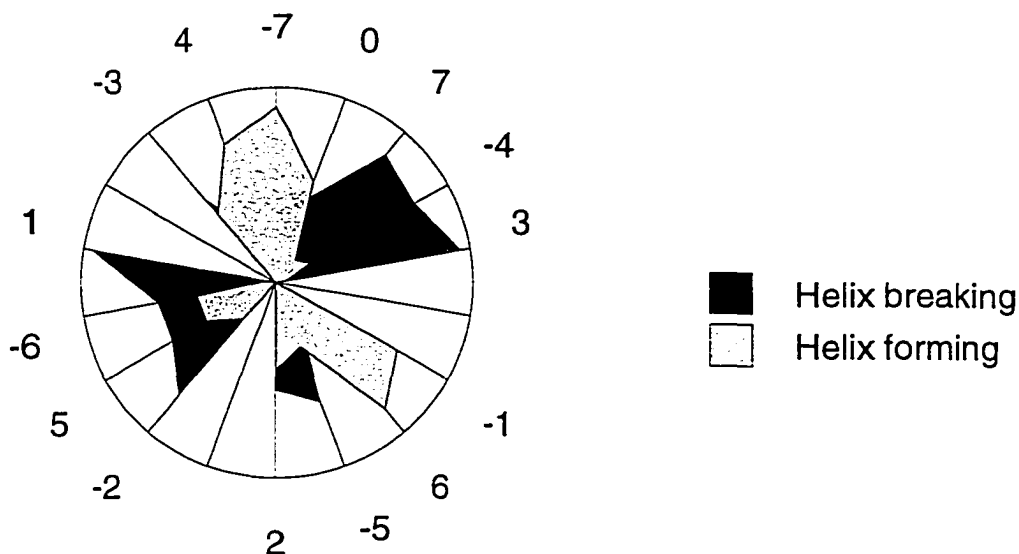
within the system. Window positions  $-2$  through  $+4$  appear to represent inputs that are considered important to the system. These inputs have large magnitudes and can be readily classified as helix-forming or helix-breaking positions. Specifically, positions  $-3$ ,  $-1$ ,  $0$ ,  $+2$ , and  $+4$  appear to be helix forming positions, while positions  $-2$ ,  $+1$ , and  $+3$  appear to be helix breaking position.

The window positions between  $-3$  and  $+7$  show strong similarities in weight magnitudes, regardless of the amino acid. In contrast, the positions between  $-7$  and  $-4$  appear fairly randomized, possibly representing noise in the system.

When the importance of each position is placed around the helix wheel (Fig. 12), it appears that the neural network system clusters helix forming and helix breaking window positions. This suggests that the system matches strong helix forming patterns (or weak helix breaking patterns) around the helix wheel with helix



**Fig. 11.** A trained neural network weight matrix, ordered by window position. The system threshold is 0.04.



**Fig. 12.** Clusters of helix forming and helix breaking positions around the helix wheel. Areas indicate the number of amino acids with weights above the threshold, as determined by the neural network system. The helix wheel appears divided into four distinct regions: two helix forming regions and two helix breaking regions.

identification.

Additionally, Qain and Sejnowski (1988) showed that there is no perceptible difference in accuracy when the system is trained for 50 iterations or 10,000 iterations over the training set. Advanced learning techniques, such as simulated annealing, appear unsuccessful in further reducing the prediction error. One possible explanation could be the helix wheel positioning. When the weight positions are displayed along the helix wheel during training, the helix forming and breaking regions appear to rotate. Although the forming and breaking positions appear at right-angles along the wheel, there is a virtually-infinite number of “best” positions. Error metrics, such as the mean squared error, attempt to identify the best weight matrix for the lowest error. Unfortunately, the mean squared error of this system does not vary; an increase would suggest overtraining, when the system begins to memorize the training set, while a decrease would indicate a better weight matrix. The system appears best trained when the mean squared error stabilizes, in as few as 20 iterations

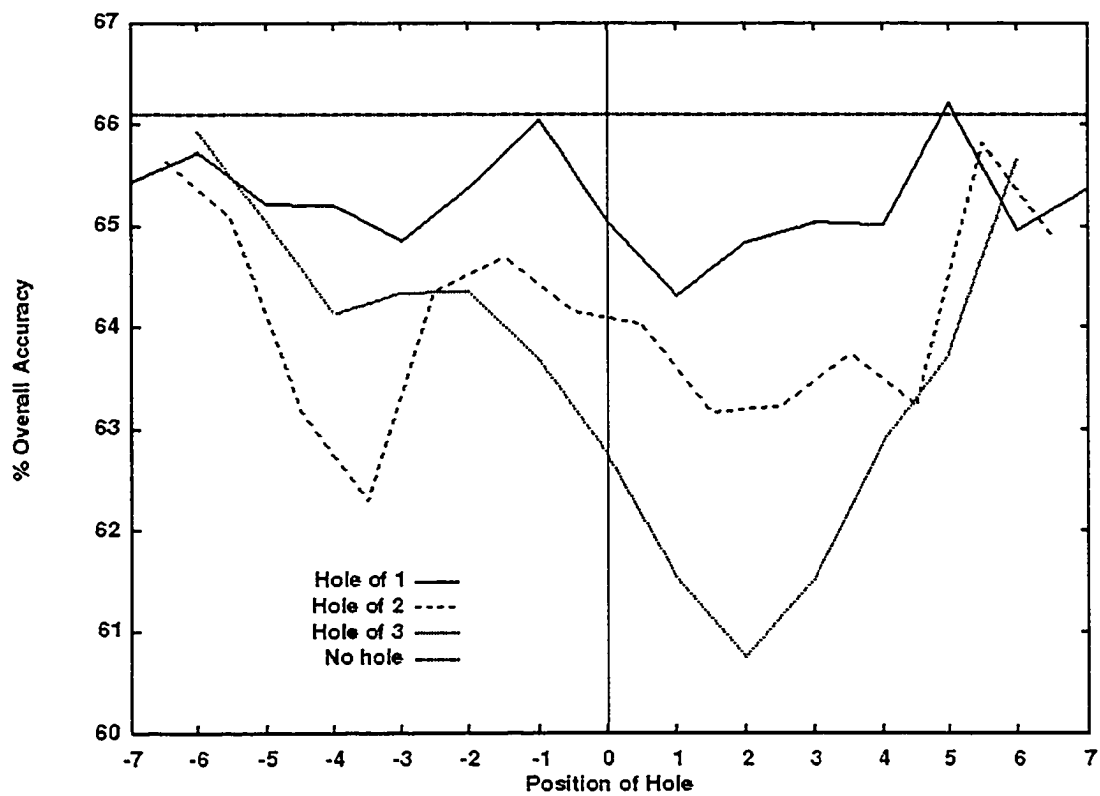


through the training set.

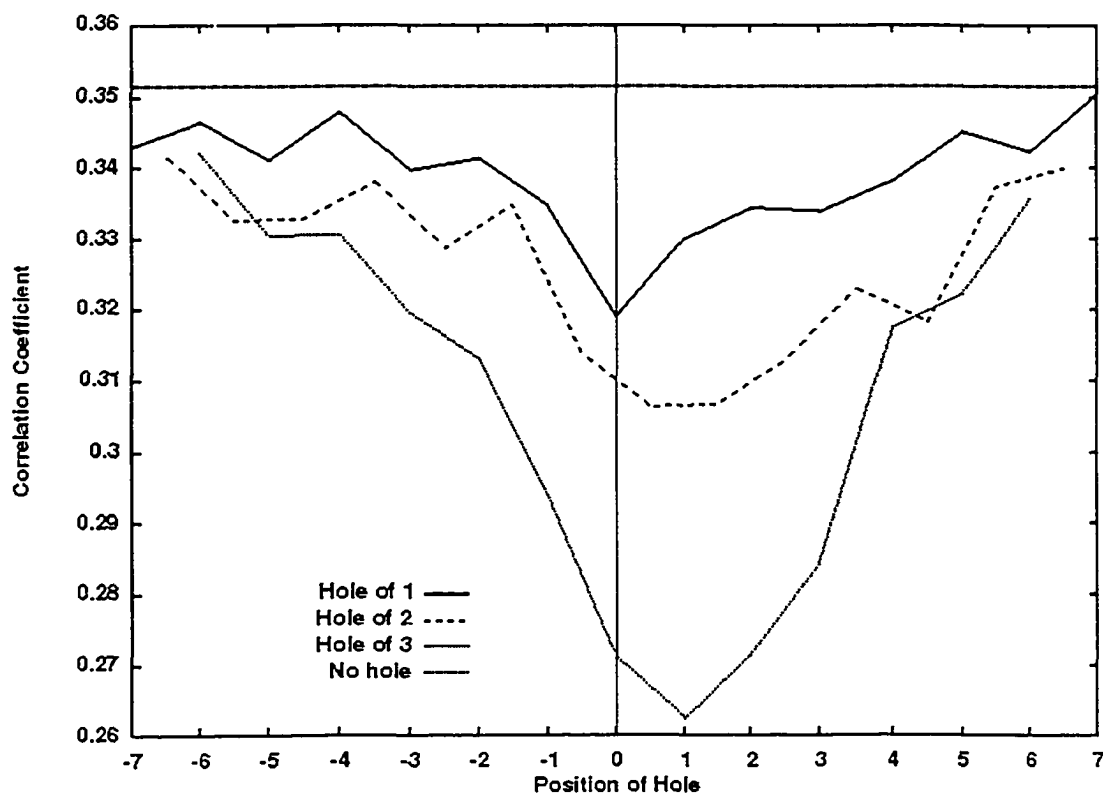
## 2. *Windows with holes*

Training the system with suppressed inputs, or holes in the window, allows the effects of each window position to be identified. Figures 13-16 show the effects of suppressing up to three positions from the input of a 15 amino acid window. Positions 0 through +4 closely correspond with the position's importance as determined by the neural network weight matrix. These positions have the strongest detrimental affect when removed from the system. Although not obvious from the observed weight matrix, positions -1 and +5 have virtually no effect on the system's performance.

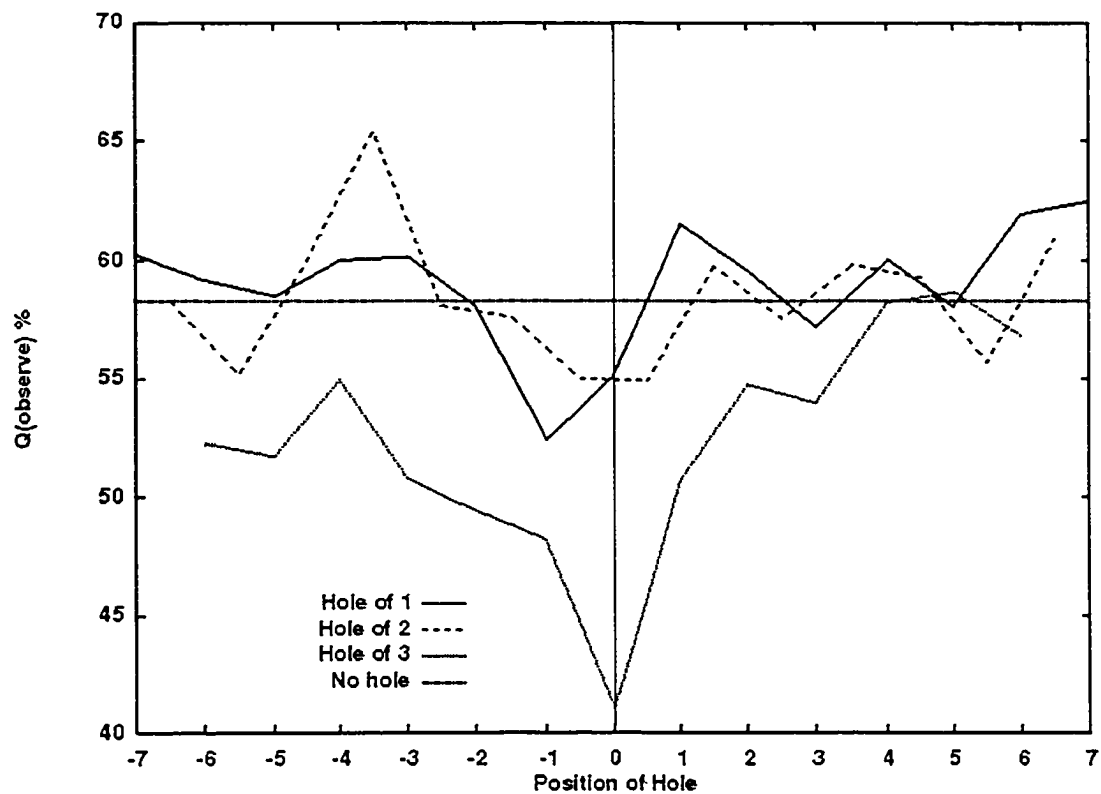
The window positions between -7 and -3 show possible noise in the neural network system. Similarly, the network trained with holes shows an inconsistency over the same region: a two amino acid hole at positions -4 and -3 performs significantly worse than either a one or three amino acid hole in the same position. This inconsistency in performance probably denotes the presence of noise in the system. Absence of positions in same region also causes an increase in  $Q_{observe}$ , indicating fewer missed helices, and a decrease in  $Q_{predict}$ , indicating fewer correct predictions.



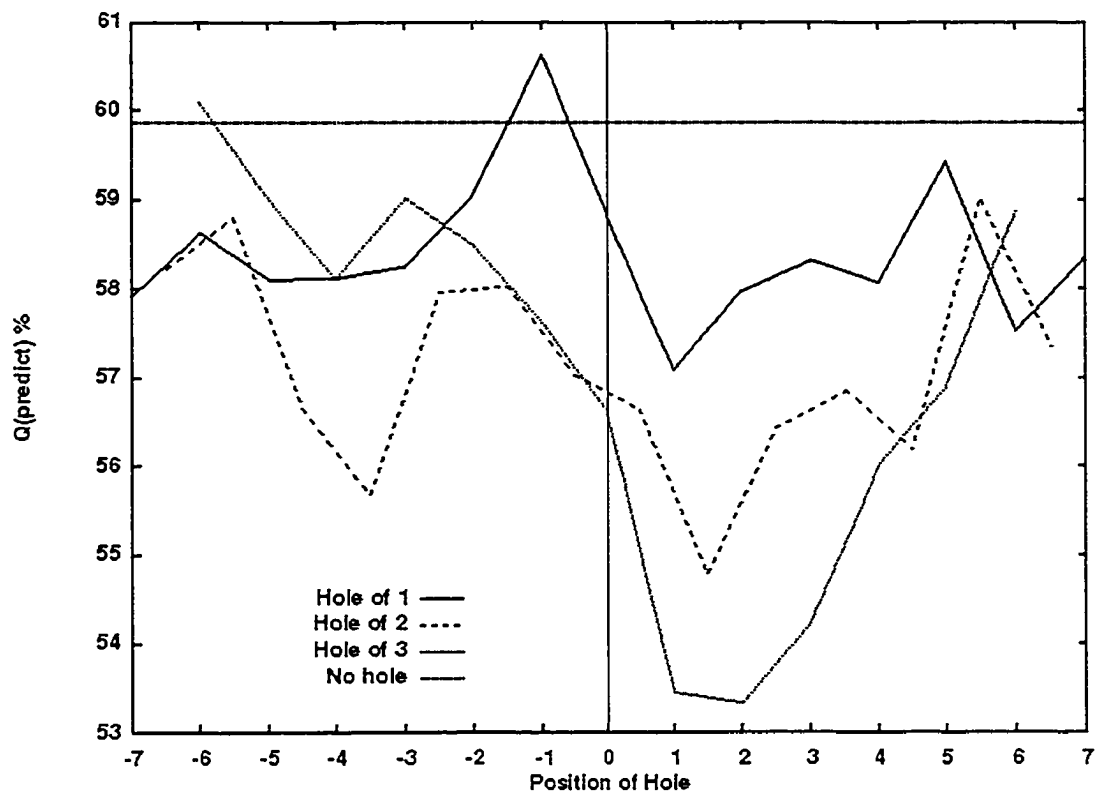
**Fig. 13.** The effect of suppressed positions (holes) within the input window, measured with the metric  $Q_i$ . Holes representing the suppression of one, two, and three amino acids are presented.



**Fig. 14.** The effect of suppressed positions (holes) within the input window. The results are measured with the  $C_{coef}$  metric. Holes representing the suppression of one, two, and three amino acids are presented.



**Fig. 15.** The effect of suppressed positions (holes) within the input window as presented by the  $Q_{\text{observe}}$  metric. Holes representing the suppression of one, two, and three amino acids are presented.



**Fig. 16.** The effect of suppressed positions (holes) within the input window, using the  $Q_{\text{predict}}$  metric. Holes representing the suppression of one, two, and three amino acids are presented.

In contrast to window positions  $[-7, -3]$ , positions  $[-2, +7]$  appear to have a high level of information content, maximized at positions +1 and +2. This indicates that the predictions formed by the neural network system are strongly influenced by the amino acids on the C-terminal side of the input window.

### *3. Asymmetrical window methodology*

We hypothesize that information concerning secondary structure location within the sliding window of amino acids is neither symmetrically distributed around the central prediction point nor minor in the amount of extraneous information. Under this hypothesis, a symmetrical window around the predicted position is not necessarily optimal for the system. A non-optimal window may add a substantial amount of noise to the system, greatly effecting the prediction accuracy. In addition, we hypothesize that the basic single sliding window approach is incapable of learning sheet or coil secondary structures.

In the original NETtalk system (Sejnowski & Rosenberg, 1986), a symmetrical window of seven letters was used to learn the pronunciation of the central letter. The window size of seven was determined by simple means: through the trial of various window sizes, seven letters performed the best. In both the Qain-Sejnowski (1988) and Holley-Karplus (1989) secondary structure prediction systems, a variety of window sizes, ranging from 3 to 21 amino acids, was tested. Each of these systems used a symmetrical window around the prediction point.

Neural networks can generally identify noise in the system by quickly discerning which inputs to the system are useful and which are not. When excessive noise is applied to the neural network, the ability to identify useful information becomes more difficult and the prediction accuracy may suffer. In prior NETtalk-based systems, it was assumed that either the information content within the window was symmetrically distributed around the predicted element, or the amount of extraneous information (noise) was minor. Thus, we hypothesize that the information

content is neither symmetrically distributed nor minor in systems with large sliding windows.

NETtalk-based prediction systems are used for determining helix, sheet, and coil locations from a single window of the primary sequence. The prediction of each secondary structure is independent, so conflicting predictions are resolved by awarding the structure to the strongest prediction. If, for example, a particular window is predicted as 0.28 helix, 0.27 sheet, and 0.26 coil, then the prediction would be a helix due to the larger likelihood. We hypothesize that the simple neural network system cannot learn sheets or coils, but instead learns different ways to represent helix (or not-helix) structures.

To test these hypotheses, the window is divided into three distinct regions. Through the modification of these regions, each hypothesis can be tested.

The first window region,  $R_I$ , represents the position being predicted. This region is one amino acid in length and located between the two other regions. The second region,  $R_N$ , contains all positions that are on the N-terminal side of the predicted region. Similarly, the third region,  $R_C$ , contains all positions that are on the C-terminal side of the predicted region. The entire input window can be described as  $R_N+R_I+R_C$ . A symmetrical window of seven amino acids would be represented as 3+1+3, while an asymmetrical window of seven amino acids could be described as 0+1+6, 1+1+5, 2+1+4, 4+1+2, 5+1+1, or 6+1+0.

The middle region,  $R_I$ , is always represented as one amino acid in length. Conceptually,  $R_I$  may be larger than a single amino acid, representing a larger predicted region. However, in implementation, a large  $R_I$  is equivalent to a small  $R_I$  with larger  $R_N$  and  $R_C$  regions. For example, the asymmetrical window 2+3+4 is equivalent to 2+(1+1+1)+4, or simply 3+1+5. This holds true for all protein regions except at the protein's terminals, where a structure can only be predicted when all of the large  $R_I$  is contained within the protein. Systems that predict based on large  $R_I$  regions perform similarly to those with small  $R_I$  and larger  $R_N$  and  $R_C$  regions.

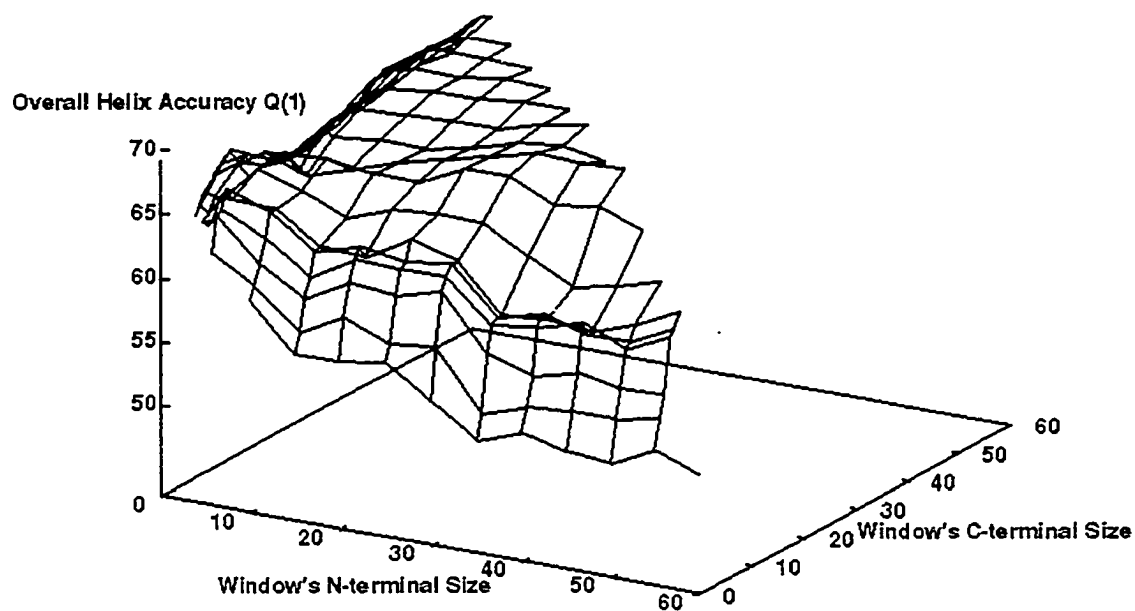
#### 4. *Asymmetrical window analysis*

The neural network system was trained on a variety of asymmetrical window sizes. All windows of  $R_N+1+R_C$  from seven to 61 amino acids in length were tested. Through the training of the neural network system on a variety of asymmetrical window sizes, the information content per terminal was determined, including which structures were applicable to this implementation of the sliding window approach.

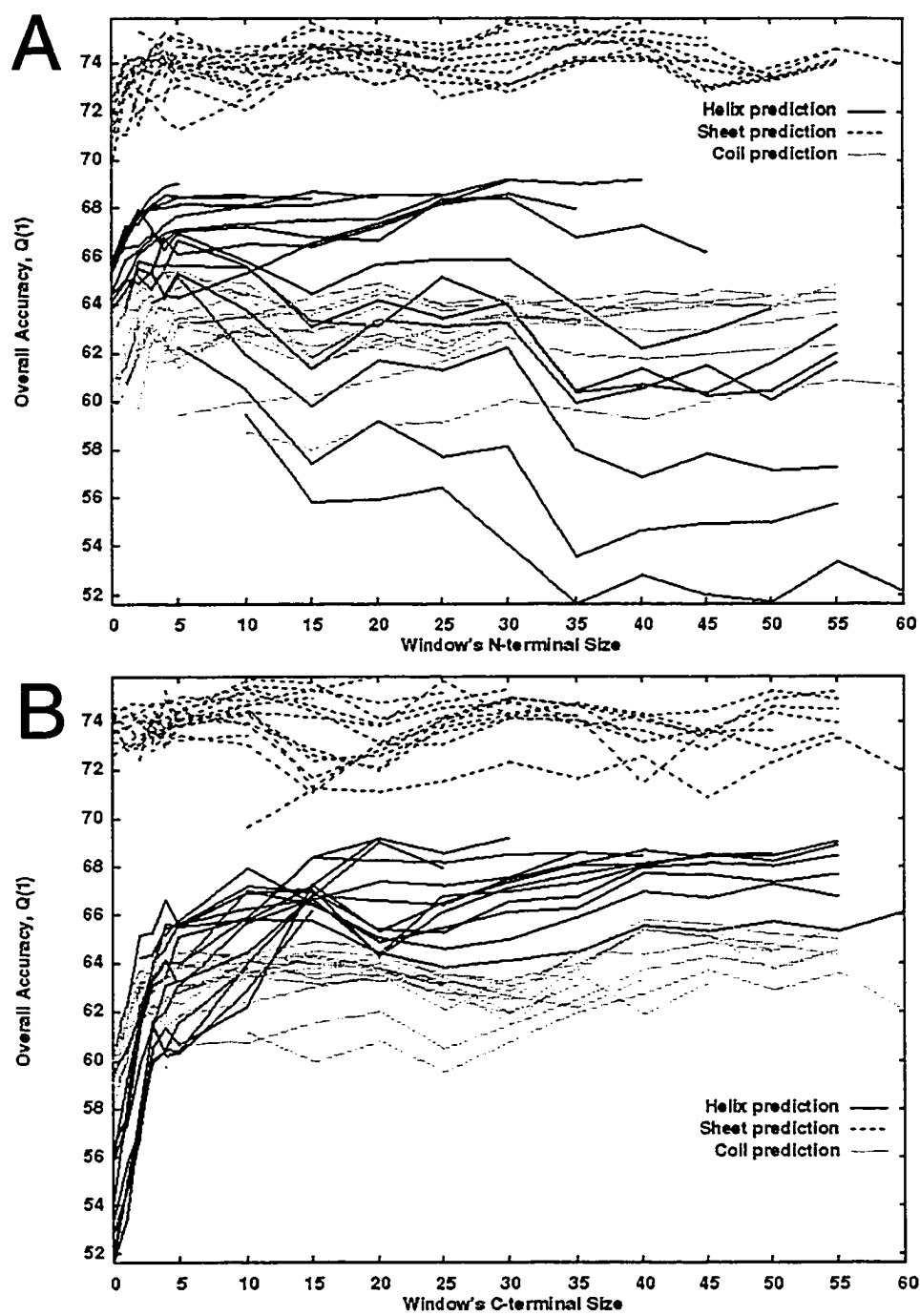
As illustrated in the helix prediction (Fig. 17), the effects of the  $R_N$  and  $R_C$  regions are asymmetrical. A small  $R_C$  region has a strong detrimental influence on the entire helix prediction, while having a small  $R_N$  region has a minimal effect on the system's performance. This may indicate that factors which begin the helix are less important than factors which form the C-terminal.

NETtalk-based prediction systems are commonly used to predict sheets and coils, as well as helices. Figures 18A and 18B view the surface described in Figure 17 along the window's N- and C-terminals,  $R_N$  and  $R_C$ , respectively, including accuracies for helix, sheet, and coil predictions. This allows for identification of the region's influence on the prediction.





**Fig. 17.** The effects of variable asymmetrical window sizes on the prediction accuracy of the sliding window neural network model. The  $Q_i$  accuracy of the system shows the three-dimensional surface from variable window sizes on helix prediction.



**Fig. 18.** The effects of variable asymmetrical window sizes on the prediction accuracy of the sliding window neural network model. **A:** the effects of variable N-terminal sizes on the  $Q_1$  accuracy of the system. **B:** the effects of different C-terminal sizes on the prediction

The neural network seems able to identify some factors involved in determining helix location. This is evident by the dramatic effect changing the window regions has on the helix prediction accuracy. A small  $R_C$  region significantly reduces the effectiveness of the helix prediction, suggesting a high degree of information content. The minimum effective  $R_C$  length is around 2 amino acids, or half a loop in a helix; the optimal  $R_C$  size is around 10 amino acids, or 3 helix loops. Although the  $Q_I$  metric continues to increase over larger  $R_C$  sizes, the  $Q_{observe}$  values drop due to the excessive noise introduced by a large window size.

Although the effect of a small  $R_N$  region is not as dramatic as  $R_C$ , it seems optimal at two or three amino acids (one helix loop). An N-terminal longer than seven amino acids (2 helix loops) appears to introduce excessive noise into the system, reducing its performance.

By using an asymmetrical window of 2+1+10, the simple accuracy metric  $Q_I$  maximizes at 67.9%, with a correlation coefficient,  $C_{coef}$  of 0.37.

For sheet prediction (Fig. 18A and B), the simple accuracy metric,  $Q_I$ , does not vary significantly for any size window. The  $Q_I$  metric measures around 75% over all window sizes. Amino acids are involved in sheets 25% of the time. By predicting “not-sheet” most of the time, a constant accuracy of 75% can be achieved. Although the neural network is used to predict sheets, this prediction alone is incorrect nearly as often as it is correct. This lack of variation suggests the single sliding window neural network system is incapable of determining sheet locations. One possible explanation is that a sheet depends on bonds between two or more non-adjacent segments in the primary sequence. Since the single sliding window only views a fragment of the sheet, the network is unable to learn the entire structure. To determine a sheet’s location properly, multiple sliding windows would be required, increasing the system’s complexity exponentially.

In contrast, the variable window size does appear to contain information related to helix and coil prediction (Fig. 18A and B). In both predictions, the effects of variable window sizes seem closely related; both show maximum information

stored within the same region lengths. This close relationship suggests an inability to learn coil location and a recognition of “not-helix” structures.

Based on the accuracy from variable window sizes, we can support our hypothesis: the information stored in the C-terminal side of the predicted window appears to have a larger importance in the prediction process than the amino acids along the N-terminal side. Furthermore, a single sliding window seems unable to predict sheets and coils; the NETtalk-based system appears only capable of determining helix locations. Consequently, the remainder of this research focuses strictly on helix determination.

### **C. Association with amino acid properties**

The quality of input into the perceptron directly affects the system’s learning ability. If, for example, the input is cryptic, then the system must learn to decrypt the input before evaluating the data. In contrast, data that is too simplistic or uninformative may not provide enough information to the system. Although the sliding window encoding is not cryptic, it may be too simplistic by overlooking key elements that may enhance the prediction process. In particular, known amino acid attributes, such as hydrophobicity, may provide useful information that the system cannot otherwise readily identify.

Some prediction systems, such as Chou-Fasman (1974) and Fauchere-Pliska (1983), use observed data to generate structural predictions. It is conceivable that this information cannot be readily derived from the position sequence. We hypothesize that associating these attributes with the positional information can provide additional useful information for the neural network system. To test this hypothesis, the statistical data from Chou-Fasman and Fauchere-Pliska were combined with each amino acid’s occurrence within the sliding window. This effectively adds “attributes” to each position within the window. Additionally, each amino acid’s molecular weight was used as an attribute.

To apply the amino acid attributes to the sliding window system, the input to the network was modified from an array of possible amino acids to the amino acid's single value attribute. For a seven amino acid window, the NETtalk system normally received 154 inputs: 7 window positions  $\times$  22 amino acid types (20 common amino acids + Unk + Nul). When using amino acid attributes, the same seven amino acid window system would only have seven inputs, one for each attribute. The attribute values were combined with the sequence position, incorporating 161 total inputs to the system.

The results of applying amino acid attributes to the NETtalk-based sliding window system (Table 5) show a substantial change in the system's performance. Without using the sequence position, both Chou-Fasman and Fauchere-Pliska predict more accurately than the position sequence input. This indicates that the attribute's information is useful to the prediction system.

Combining the attributes with the position sequence dramatically improves the generated predictions. This indicates that the information learned from the attributes is sufficiently different from the information learned from the position sequence. In addition, the information is complementary, accounting for the increase in prediction accuracy when combined.

The content provided by the attributes appears to identify different aspects of the learned information. The prediction system using Fauchere-Pliska values, measuring hydrophathy, does not dramatically increase the accuracy when combined with the sequence information. This seems to indicate that the information content of the two inputs is similar. In addition, the  $Q_{observe}$  metric decreases in the combined system, suggesting an increase in noise or contradictory information.

Combining the Chou-Fasman system with the sequence position appears to perform similarly to the Fauchere-Pliska system. This suggests that the information learned by the combination of Chou-Fasman and sequence system may be equivalent to the Fauchere-Pliska system.

In contrast to the Chou-Fasman and Fauchere-Pliska systems, the molecular

**Table 5.** *Effects of associated amino acid attributes on helix prediction*

Attribute	$Q_i$	$Q_{predict}$	$Q_{observe}$	$C_{coef}$
Single input types				
Sequence position only	66.50%	61.20%	57.84%	0.357
Chou-Fasman	68.97%	68.60%	54.67%	0.389
Fauchere-Pliska	64.73%	60.84%	66.77%	0.395
Molecular weight	60.49%	67.49%	33.81%	0.327
Combined input types				
Sequence position + Chou-Fasman	69.57%	66.56%	52.46%	0.384
Sequence position + Fauchere-Pliska	69.50%	66.00%	54.56%	0.388
Sequence position + Molecular weight	73.50%	69.16%	60.38%	0.442
Sequence position + all 3 attributes	74.05%	72.50%	58.47%	0.460

weight attributes perform substantially better when combined with the sequence position information. This suggests that the size of the amino acid, while relevant to the position, correlates with important information that cannot be directly derived from the sequence position observations.

#### **D. Implicit assumptions and limitations in the neural network system**

The implementation of the NETtalk-based neural network system for protein structure identification includes numerous assumptions which are not optimal for the prediction system. These assumptions include secondary structure determination and information content location.

Additional limitations to the system include the data set composition and size of the network system.

### *1. Secondary structure determination and information content location*

As shown in this chapter by the use of asymmetrical window regions, a single sliding window system cannot accurately predict sheets or coils; only helix identification is directly applicable. In addition, the tertiary and quaternary interactions may represent an important factor in more accurate structural identification, accounting for the 25% - 30% error in  $Q_i$  measurements.

Through the use of windows with holes, we have identified the regions of high information content with respects to the neural network system. Window positions +1 and +2 appear to have the strongest influence on the neural network prediction. Furthermore, training the neural network system with asymmetrical windows reveals a larger influence on the prediction from the C-terminal of the window, with little influence from the window's N-terminal, with an optimal window range of [-2,10].

### *2. Hidden layered systems and data set size*

As shown by Qain and Sejnowski (1988), the results of training a multi-layered neural network system are not significantly different than those of a single network system. In addition, the multi-layered system appeared to memorize the training set. Training the multi-layered system with the larger data set described in Chapter III yielded little improvement in prediction performance, although the training set was not memorized by the system as rapidly. From this, we conclude that there is probably not enough data in the training set to teach properly a hidden-layer system.

## VI. PROBABILISTIC ANALYSIS REGARDING HELIX PROPENSITIES

In this chapter we define region-specific, position-dependent propensities which quantify the likelihood that a given amino acid and its neighboring positions form the N-terminus, middle, and C-terminus of a helix. These values are combined using a Bayesian probabilistic approach to identify potential helix regions, including terminals and middles. From these potential helix regions, the pattern matching system described in Chapter VII determines the areas likely to be helices. Additionally, this approach incorporates a step-wise prediction process, allowing identification of the factors which are most significant in predicting the helix.

### A. Definitions

We hypothesize that the regions of a helix have different helical propensities. In this section we define the helical regions and the region-specific, position-dependent propensities used in this research.

#### *1. Window of amino acids*

A window of neighboring amino acids is used to define the influence range of a given amino acid. The influence of an amino acid varies with location within the window and is assumed to be insignificant outside the window. An amino acid at position  $i$  in the known primary sequence has a window covering the neighboring amino acid range  $[i-k_N, i+k_C]$ , where  $i-k_N$  and  $i+k_C$  are the window's N-terminal and C-terminal, respectively. Under this definition, the window of amino acids does not necessarily need to be symmetrical:  $k_N$  does not need to be the same as  $k_C$ . An asymmetrical window allows the prediction process to emphasize the influence of a



specific terminal.

## 2. *Regions of a helix*

We hypothesize that the helical propensity for forming the middle of the helix differs from the propensities for forming the terminals. Therefore, we classify helices into three distinct regions: N-terminals, C-terminals, and middles. N-terminals correspond to the amino acids that are at the start of helices. Similarly, C-terminals are amino acids that are at the end of helices, and middles are between the N- and C-terminals. These regions are used to define the predicted helical propensities,  $N$ ,  $C$ , and  $M$ , which denote the presence of N-terminals, C-terminals, and middles, respectively. We allow each amino acid in a primary sequence to be described by all three regional propensity values.

## 3. *Region-specific, position-dependent propensities*

Instead of a single helical propensity, each amino acid has different likelihoods to form the  $N$ ,  $C$ , and  $M$  regions of a helix. In addition, the influence on neighboring positions varies with distance. Therefore, each amino acid,  $a$ , is assigned multiple propensities to reflect its influence in forming the different helix regions,  $r \in [N, M, C]$ , at each position within the window,  $i \in [-k_N, +k_C]$ . This propensity is represented as the conditional probability:  $P(r \text{ at } 0 \mid a \text{ at } i)$ .  $P(N \text{ at } 0 \mid a \text{ at } i)$  indicates the likelihood that window position 0 is a helix N-terminal when window position  $i$  is the amino acid  $a$ . Similarly,  $P(C \text{ at } 0 \mid a \text{ at } i)$  and  $P(M \text{ at } 0 \mid a \text{ at } i)$  indicate the propensities for helix C-terminal and middle regions.

For a window ranging over  $[-7, +7]$ , each amino acid would have a total of 45 propensities: 15 N-terminal, 15 C-terminal, and 15 middle propensities. For example, Proline (Table 6) has  $M$  propensities that vary by more than 0.16 across a the window and are not linearly distributed; the strongest  $M$  region propensities are far from

**Table 6.**  $P(r \in [N, M, C] \text{ at } 0 \mid \text{Proline at } i)$  over the window  $i \in [-7, +7]^a$ 

Region, $r$	Window position, $i$														
	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
$N$	0.15	0.14	0.15	0.16	0.15	0.20	0.22	0.19	0.16	0.14	0.09	0.05	0.05	0.06	0.08
$M$	0.31	0.29	0.27	0.27	0.27	0.22	0.20	0.16	0.15	0.16	0.17	0.21	0.23	0.25	0.28
$C$	0.09	0.09	0.08	0.07	0.06	0.05	0.07	0.11	0.14	0.18	0.21	0.19	0.17	0.17	0.16

<sup>a</sup> The data presented in this table was observed from the training set, described in Chapter III.

position 0, indicating a preference to be found outside of the middle region. Proline's N- and C-terminal propensities also vary dramatically across different window positions. Proline at window position 0 shows a strong preference towards preceding the helix N-terminal, rather than appearing after (comparing 0.22 at position -1 with 0.05 at positions +4 and +5). Similarly, the data indicates a preference for appearing after the C-terminal. In general, each amino acid is assigned multiple propensity values, one for each region, at each window position.

## B. Computing likelihoods

Ideally, we would like to quantify the helix-forming probability of position  $i$  in a protein given a window of the primary sequence:  $P(\text{position } i \text{ is } r \in [N, M, C] \mid \text{primary sequence in window } [i - k_N, i + k_C])$ . The notation  $P(A \mid B)$  is the probability of the event  $A$  occurring given that the condition  $B$  is true. In general, we want to determine region-specific propensities from a given sequence in the window,  $w = \langle a_{k_N}, a_{k_N+1}, \dots, a_0, \dots, a_{k_C} \rangle$  where  $a_j$  is the specific amino acid  $a$  at window position  $j$ , using the conditional probability:  $P(r \mid w)$ . Bayes' theorem provides a means for deriving such posterior probability,  $P(r \mid w)$ , from the conditional probabilities,  $P(w \mid r)$ , and the priors  $P(r)$  and  $P(w)$ :

$$P(r|w) = \frac{P(r) \times P(w|r)}{P(w)} \quad (13)$$

The probabilities  $P(r)$  and  $P(w)$  are determined by observation of the training set.  $P(r)$  is the observed occurrence rate of the region  $r$ ;  $P(w)$  is the observed occurrence rate of the amino acid amino acid sequence,  $w$ .  $P(w)$  combines the occurrence likelihoods of each amino acid in the protein sequence:

$$P(w) = \prod_{j=k_N}^{k_C} P(a_j) = P(a_{k_N}) \times P(a_{k_N+1}) \times \dots \times P(a_{k_C}). \quad (14)$$

$P(a_j)$  represents the observed occurrence rate of the amino acid,  $a$ , occurring in window position  $j$ . Although  $P(a_j)$  may be similar to the occurrence rate of the amino acid,  $P(a) = P(a_0)$  when  $a$ ,  $a_0$ , and  $a_j$  are the same type of amino acid (e.g.,  $a$ ,  $a_0$ , and  $a_j$  are all alanine),  $P(a_0)$  is not necessarily the same as  $P(a_j)$ . When observing each amino acid's occurrence in each window of the training set, it is possible for window position 0 to be on or near the protein's terminus. No amino acid is available for the window position  $j$  when protein position  $i+j$  is located beyond the protein's terminus. This unavailability at the protein's terminus leads to a difference between  $P(a_j)$  and  $P(a_0)$  when both represent the same type of amino acid.

Determining the helical prediction directly from  $P(w | r)$  is impractical. For a small window containing 7 amino acids, this would require a minimum of  $3 \times 20^7$  conditional probabilities, one for each unique sequence of 7 amino acids in each region. To resolve this problem, the conditional independence assumption is employed. The conditional independence assumption states that two amino acids at different positions within the window,  $a_0$  and  $a_k$  ( $k \neq 0$ ), are conditionally independent in a helix region,  $r$ , when  $P(a_0 | a_k, r) = P(a_k | r)$ , or equivalently,  $P(a_0, a_k | r) = P(a_0 | r) \times P(a_k | r)$ . The conditional independence assumption greatly reduces the conditional probabilities needed for a Bayesian inference; hence, it is used quite frequently in the

application of Bayes' theorem. Based on this assumption, the conditional probability  $P(w | r)$  in Equation 13 can be computed from the product of  $P(a_j | r)$ . To see this, we first replace  $w$  with its definition:

$$P(w|r) = P(a_{k_N}, a_{k_N-1}, \dots, a_0, \dots, a_{k_C} | r). \quad (15)$$

Based on the conditional independence assumption stated above, we have:

$$\begin{aligned} P(w|r) &= P(a_{k_N} | r) \times P(a_{k_N-1} | r) \times \dots \times P(a_0 | r) \times \dots \times P(a_{k_C} | r) \\ &= \prod_{j=k_N}^{k_C} P(a_j | r). \end{aligned} \quad (16)$$

Because the prior probability,  $P(w)$ , can be expressed as  $P(w|r) \times P(r) + P(w|\neg r) \times P(\neg r)$ , Equation 13 can be rewritten as:

$$P(r|w) = \frac{P(a_{k_N}, \dots, a_{k_C} | r) \times P(r)}{P(a_{k_N}, \dots, a_{k_C} | r) \times P(r) + P(a_{k_N}, \dots, a_{k_C} | \neg r) \times P(\neg r)}; \quad (17)$$

$$P(r|w) = \frac{\{P(a_{k_N} | r) \times \dots \times P(a_{k_C} | r)\} \times P(r)}{\{P(a_{k_N} | r) \times \dots \times P(a_{k_C} | r)\} \times P(r) + \{P(a_{k_N} | \neg r) \times \dots \times P(a_{k_C} | \neg r)\} \times P(\neg r)}. \quad (18)$$

The conditional probabilities are determined by observing the training set.  $P(a_j | r)$  refers to the occurrence of amino acid  $a$  at position  $j$  when position 0 is region  $r$ . For example,  $P(\text{Ala at } +2 | N)$  determines the likelihood that window position +2 is Alanine whenever position 0 is a helical N-terminal. Similarly,  $P(a_j | \neg r)$  refers to the likelihood that the region at position 0 is not  $r$ . The prior  $P(\neg r)$  is equivalent to  $1 - P(r)$ .

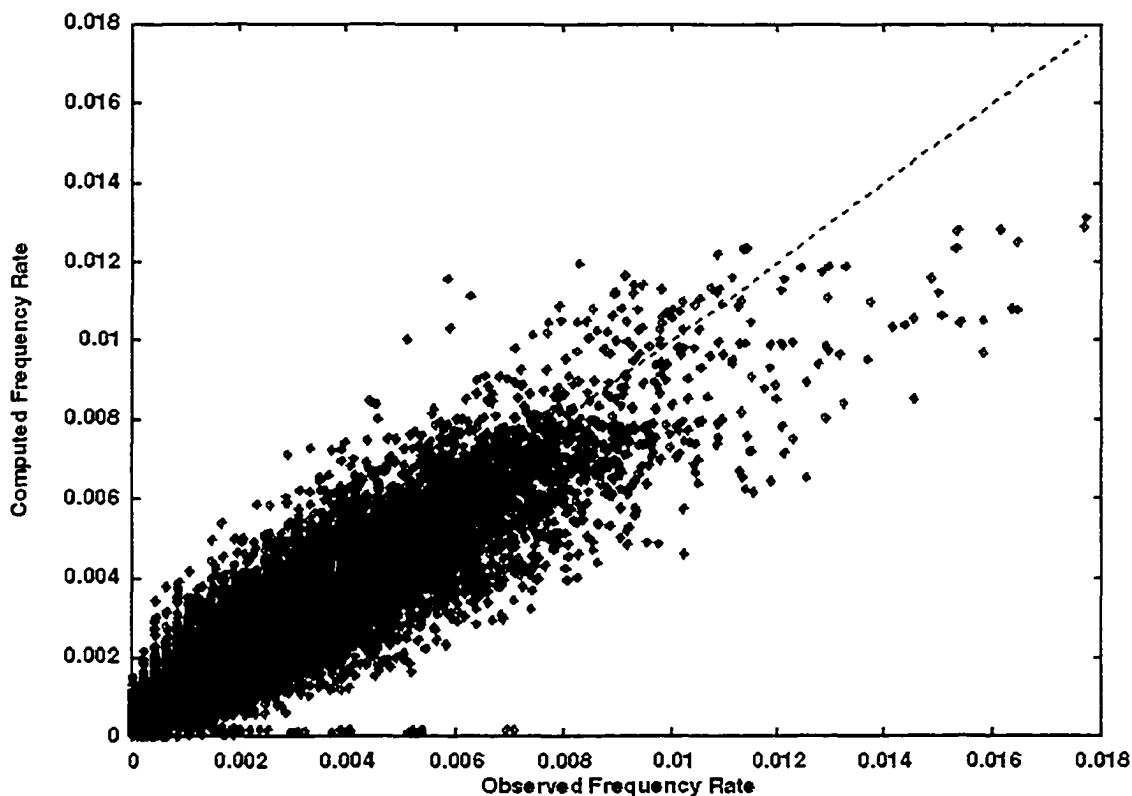
For each position,  $i$ , in a protein in the training set, a window  $[i-k_N, i+k_C]$  is constructed. The likelihood that the amino acid at position  $j \in [i-k_N, i+k_C]$  within a

window, with region  $r$  at position 0, are observed. These conditional probabilities are denoted:  $P(a_j | N)$ ,  $P(a_j | M)$ , and  $P(a_j | C)$ . The prior probabilities for the three helix regions are  $P(N)$ ,  $P(M)$ , and  $P(C)$ .

### C. Testing the conditional independence assumption

For testing the validity of this assumption, we compare the computed product of individual conditional probabilities from the training set,  $P(a_0 | r) \times P(a_k | r)$ , with the observed joint conditional probabilities,  $P(a_0, a_k | r)$ . To remove homology,  $a_0$  is required to be in a unique primary sequence context.

The pair-wise population of amino acids, where window position 0 is in a unique context with a neighborhood range  $k \in [-5, +5]$ , yields a data population size of 26,460 pairs of amino acids (this includes correlations with Unk, the unknown amino acid). These pairs include position 0 as an N-terminal, C-terminal, and middle of a helix, as well as not N-terminal, not C-terminal, and not-middle regions. The mean difference between the observed and computed pairwise conditional probabilities is 0.0000000071429 ( $7.1429 \times E^{-9}$ ) with a standard deviation of  $7.2896 \times E^{-4}$  (Fig. 19). The coefficient of determination,  $r^2$ , is 0.8232, showing a strong correlation between the product of two conditional probabilities and their joint conditional probabilities. This high degree of correlation allows us to confidently assume that these amino acids combine helical propensities conditionally independently.

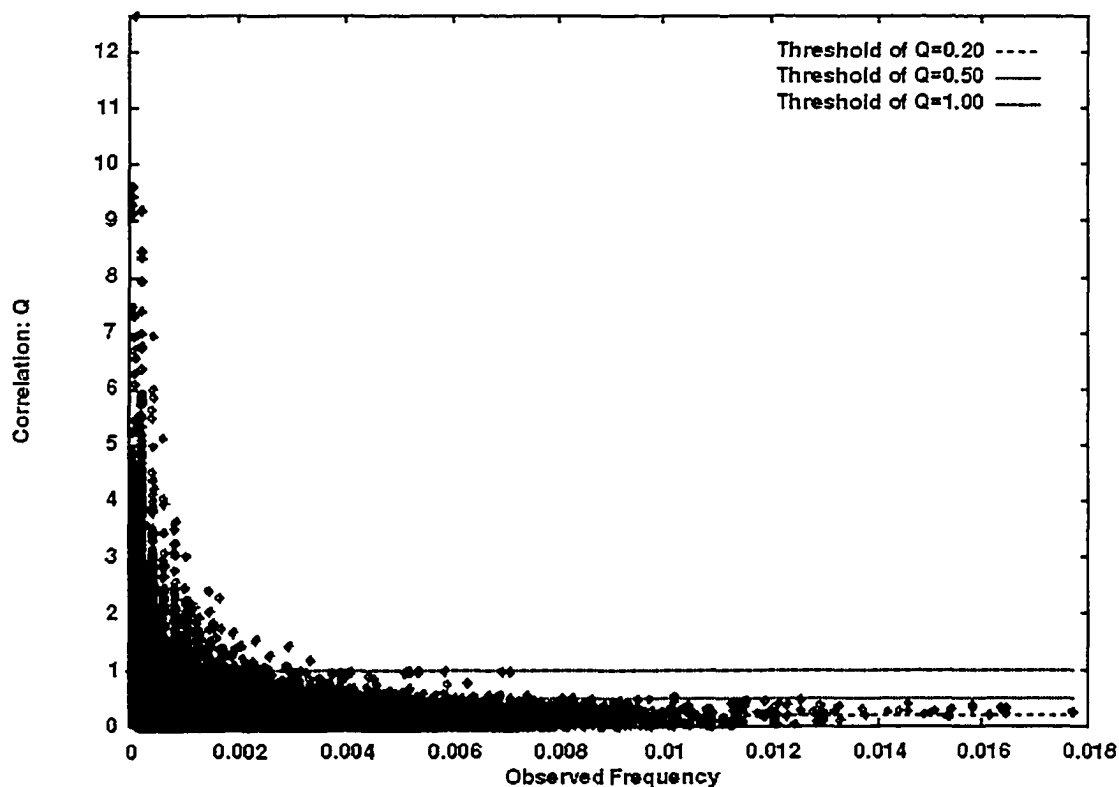


**Fig. 19.** Independence analysis comparing computed occurrence rate of amino acid pairs in helical regions with the observed frequency. The high degree of similarity between the observed and computed probabilities suggests that amino acid helical propensities combine independently.

Although most pairs of amino acids seem to be conditionally independent, it is possible for some outliers in the data to be conditionally dependent. Outliers within the independence assumption analysis were determined using a quotient,

$$Q = \frac{\|P(a_0, a_k | r \text{ at } 0) - P(a_0 | r \text{ at } 0) \times P(a_k | r \text{ at } 0)\|}{P(a_0, a_k | r \text{ at } 0)}; k \neq 0. \quad (19)$$

At a cutoff of  $Q=1$ , outliers account for less than 2% of the available data. Each of the outliers contained infrequent amino acid combinations, including Unk, Trp, Cys, His, and Phe. Extremely low joint conditional probabilities in the



**Fig. 20.** The outliers from the independence analysis are due to infrequency between the computed pairs. The outliers do not support the hypothesis that amino acid helical propensities are a dependent interaction. The threshold values of 0.20, 0.50, and 1.00 represent 25%, 50%, and 90% of the data points.

denominator of Equation 19 result in large variances of the quotient (Fig. 20).

#### **D. Analyzing probabilistic information regarding helix formation propensities**

In this section we describe the probabilistic information collected in this study and its implications for a novel probabilistic representation of helix propensity. The insight gained from this study has an important impact on the design of our methodology for identifying helix patterns, discussed in Chapter VII. Graphs of the propensity measurements are available in Appendix B.

Using the collected frequency of occurrences, we analyzed four independent

representations from each amino acid's helical propensity: the peak, N-terminal preference and cycles, C-terminal preference and cycles, and the scaled range.

### 1. *Peak definition*

The *peak*,  $P(M \text{ at } 0 | a_k)$ , determines the basic likelihood of an amino acid,  $a_k$ , forming a helix. Since amino acids are assumed to combine independently, it is feasible for an amino acid to modify neighboring position helix likelihoods more than its own position. For example, Leu has a larger helical propensity at position +2 than at position 0. The highest (or lowest) helical propensity given the amino acid at position 0 determines the peak and shows the amount of skew in the amino acid's area of influence. We hypothesize that the combination of these neighboring influences determines the helical propensity at a specific position within the primary sequence. Few amino acids are observed peaking at, or having symmetry around, position 0. For example, Ala and Pro are observed to have peaks at position +1 and Arg peaks around +3. The width and magnitude of the peak varies for each amino acid, showing an asymmetrical region of influence.

### 2. *N-terminal preference and cycles*

The peak is not capable of providing information about the terminals because the helix terminals may appear at any position within the window. Aligning all the helices by the N-terminals,  $P(a_k | N \text{ at } 0)$ , identifies cycles and *N-terminal preferences*. When viewed from the aligned N-terminals, many amino acids display clear *N-terminal cycles* which oscillate every 3.6 amino acid positions, strongly correlating with the helix wheel. In particular, it should be noted that many amino acids show a preference toward a particular side of the helix wheel, dependent upon the terminal. This seems to correspond with known ambipathic stability configurations (Kyte & Doolittle, 1982; Chakrabarty & Baldwin, 1995). Some



amino acids only show cycles when seen from a specific terminal. For example, Ala does not show cycles when seen from the N-terminal, but displays strong helix propensity oscillations every 3.6 amino acids when viewed from the C-terminal.

### 3. C-terminal preference and cycles

Similar to the N-terminal preference and cycles, aligning all the helices by their C-terminals allows us to determine helical propensities for the C-terminal, as well as propensity cycles correlating with the helix wheel. The aliphatics, Val, Leu, and Ile, as well as Ala, Lys, and Trp show distinct cycles when viewed from the C-terminal.

### 4. Scaled range definition

To determine whether an amino acid has a general preference toward the beginning, middle, or end of the helix, we scaled each helix to a uniform length, providing a common *range*. The scaled helices range from 0% at the N-terminal to 100% at the C-terminal. A larger range,  $s \in [-10\%, 110\%]$ , is used to determine propensities “outside” the helical structure. Because helices vary in length, a single amino acid in the helix usually accounts for more than 1% of the scaled helix. Determining  $P(a_i \text{ in scaled helix } | \text{ scaled helix})$  shows that each amino acid clearly contains a region of high propensity within the scaled helix. Unlike the peak, amino acids with similar physical properties seem to have similar range distributions. For example, the acids Asp and Glu both prefer the N-terminal, while the bases Lys and Arg show no strong N-terminal preference. Other amino acids, such as Pro and Gly, are seen as strong helix breakers, even though Pro is also observed as a strong helix starter.

## *5. Relation to amino acid and protein structures*

The region-specific, position-dependent probabilities are different and unique, without expressing contradictory information. By relating known high-level helical features with these probabilities, we can develop heuristics for determining likely helix locations. We illustrate the relationship between the probabilistic information and known structures by comparing the data with known amino acid classifications and helix structures, including helix dipoles and rotational location along the helix wheel.

### a. Amino acid similarities and classical classifications

Classical amino acid classifications are based on the physical characteristics of the residues. Similarities between physical characteristics have guided methods for determining nonidentical homologous structures (Needleman & Wunsch, 1970; Dayhoff et al., 1983). Because the propensities were measured without references to the amino acid's physical molecular conformation, correlating similar likelihoods to similar molecular structure provides a strong argument towards the acceptability of the data.

The classical classifications included hydrophobicity, charge, and structures. We compared each of these classifications with the region-specific, position-dependent probabilities.

#### (1) Hydrophobicity

Most proteins studied exist in aqueous solution. The hydrophobic amino acids tend to move away from the ambient water molecules, while the hydrophilic amino acids are attracted to water. The hydrophobic amino acids include Ala, Met, Cys, Phe, Ile, Leu, and Val; the hydrophilic amino acids include Pro, Tyr, His, Gln, Asn,

Glu, Asp, Lys, Arg<sup>8</sup>.

The collected likelihoods show a strong similarity between amino acids with similar hydrophobicity values. Met and Cys both contain sulfur and both have a peak values near 0.04, around positions +1 and +2. They show sharp declines in value toward the C-terminus. Phe, Ile, Leu, and Val all appear to be strong helix formers and display strong N-terminal cycles. Ala also appears as a strong helix former but shows no N-terminal preference. Instead, Ala, with its single-carbon side chain, shows C-terminal cycles similar to Leu.

The hydrophilic amino acids can be further distinguished by other traits such as charge and aromatic structure.

## (2) Charge

The charged amino acids contain charged side chains. These include the negatively charged Asp and Glu, and positively charged Lys, Arg, and His. Each of these residues shows skewed, dual peaks. The most extreme peak, Asp, appears to have a strong helix-forming peak between positions -9 and -5, and a helix-breaking peak at position -1. The other charged residues show two peaks, but not as extreme as Asp.

In addition to skewed, dual peaks, the charged residues also demonstrate strong terminal preferences and cycles. The negatively-charged amino acids appear to be attracted toward the N-terminal of the helix and show strong N-terminal cycles. Similarly, the positively-charged residues are attracted to the negatively-charged helix C-terminus and show C-terminal cycles.

---

<sup>8</sup>Some amino acids are either hydrophilic or hydrophobic, depending on the scale used (Kyte & Doolittle, 1982; Engelman et al., 1986). For this reason, Trp, Thr, Gly, and Ser are considered ambiguous and are not included.

### (3) Structures: aromatic and aliphatic

The aromatic residues, Tyr, Trp, and Phe, contain carbon rings. Physically, Tyr and Trp are very similar. In the collected likelihoods, they appear very similar; they are both weak helix formers with peaks skewed toward the C-terminus. Phe contains a double ring structure and appears to have almost no peak, no terminal preference, and is not a strong helix former or breaker.

The aliphatic residues, Val, Leu, and Ile, contain a forked carbon structure. Each of these has strong helix-forming tendencies and displays strong N-terminal cycles.

#### b. Correlation with helix dipole

The helix dipole moment is a charged bias along the helix. Each of the computed probabilistic propensities indicates a terminal preference for the charged amino acids, supporting a dipole identification. The negatively-charged amino acids, Asp and Glu appear to prefer the helix N-terminus, which is positively charged. Similarly, the positively-charged amino acids, Lys, Arg, and to a lesser degree, His, indicate a preference toward the negatively-charged C-terminus.

#### c. Correlation with helix wheel

Strong correlations exist between the region-specific position-dependent propensities and the amino acid's positioning along the helix wheel. As seen from the N-terminal conditional probabilities, Ile, Leu, Val, Glu, Asp, and Phe all demonstrate cyclical variations in positional propensity approximately every 3.6 amino acids from the terminal. Similarly, Leu, Ile, Phe, and Ala show cyclical variations when viewed from the C-terminal of the helix.

These cyclical variations appear to indicate three factors concerning helix

structure stability. First, the existence of these cycles suggests that the amino acids are not evenly distributed around helices; some amino acids show a preference toward specific sides of the helix. This preference may assist in the development of amphipathic helices, increasing the helix's stability as clusters of amino acids with similar hydrophobicity align along specific sides of the helix. The amino acids with strong cycles from both terminals (Ile, Leu, and Phe) may be used as indicators of helical stability. A single position may provide strong helix-forming tendencies from both terminals, neither terminal, or from only one terminal, leading to their respective strong, weak, or moderate formation stability.

Most cycles do not begin at the terminals, but generally start after the first helix loop (after 3.6 positions from the terminal). This suggests that the location of the cycle is related to factors which terminate a helix. These factors could include the type of terminal amino acid, the system energy, or tertiary interactions.

Finally, not all amino acids have cycles. A lack of cycles for an amino acid indicates no preference toward a specific side of the helix or the factors which terminate helices. Additionally, some amino acids do not demonstrate cycles from both terminals. Val and Glu do not have C-terminal cycles, but show strong N-terminal cycles. Ala shows only C-terminal cycles. These amino acids suggest that the factors causing the N- and C-terminals are distinguishable and independent.

## VII. HELIX PATTERN METHODOLOGY

The posterior probabilities computed by the Bayesian inference system describe the likelihood that a particular position in the primary sequence is involved in an N-terminal, C-terminal, or middle region of a helix. These raw prediction likelihoods are commonly thresholded into boolean structure states; a predicted position is either a helix or a not-helix. Simple boolean thresholding may introduce problems such as spurious errors in the prediction. To refine the raw likelihoods, heuristics based on known helix formation factors are employed. The simplest heuristic, confirmation, removes spurious helix predictions from the raw likelihoods. Filling heuristics allow for small gaps in the predicted helix regions, while removal heuristics are applied to spurious classifications. Trimming heuristics are used to prevent excessively long helices. For determining the final likelihood that the predicted region is a helix, a verification heuristic is applied.

### A. Thresholded helix prediction definition

The raw likelihood values derived from the prediction systems represent a measurement of helix propensity. Although comparable with each other, these raw propensities are not necessarily defined on a linear scale. For example, a propensity of 0.28 may be very close to 0.27 and yet not near 0.29. In the case of the neural network system, a sigmoidal threshold may be used, causing an exponential difference between values. Determining the presence or absence of a helix may be difficult since the threshold may vary with each primary sequence window. As illustrated in Chapters V and VI, the presence of an amino acid in a particular position of the window may significantly affect the propensity of the position being predicted. This effect may also cause variations in the “correct” threshold for helix determination.

To simplify the prediction process, a fixed threshold is applied to all raw

propensity values. This assumes that the amount of variation, if any, in the threshold due to amino acid positions is a minor factor and can be classified as noise in the prediction process. If the raw propensity value is larger than the threshold, the structure is considered present. Similarly, a raw value lower than the threshold indicates the absence of the structure.

Similar to the neural network system, the Bayesian inference model compares the system output with a threshold value when determining the predicted structure. To simplify the Bayesian inference model, the comparison threshold is fixed to the prior value of the structure. For helix prediction, amino acids are expected to occur in helices 35% of the time.

Problems can arise from the usage of fixed threshold values. Because the predictions are determined independently, spurious errors may occur. These errors may indicate a single primary sequence position as being a helix or not-helix in sharp contrast to the adjacent positions. Additionally, a small set of amino acids, such as two adjacent Alanines or Prolines, may heavily bias a series of adjacent positions, causing a tapering of the raw propensity values. This tapering may force values to lie just above or below the threshold, skewing the final boolean prediction. Finally, because the propensities are determined independently, the final predicted helix region may be physically unstable.

## **B. Knowledge-based refinement schemes**

The application of heuristics which compare adjacent prediction values helps resolve issues raised from using fixed threshold values. These heuristics assist in the identification of spurious values and identify possible regions of physical instability. Four heuristics are identified: filling, removal, confirmation, and verification.

## 1. Filling heuristics

Helices are known to consist of multiple amino acids; a single amino acid cannot be a helix. Through the use of filling heuristics, spurious helix and not-helix predictions are removed. There are two separate filling heuristics: absolute-fill and scaled-fill. Combining and using these heuristics raises a number of issues including order of precedence and likelihood of benefits for the heuristics.

### a. Absolute-fill heuristic

The absolute-fill heuristic identifies regions of gaps between adjacent helix predictions. For the sequence of *helix - not-helix - helix*, *H-H*, the not-helix is considered spurious and changed to a helix (Table 7). This approach is defined as an absolute-fill, since the filling does not take the raw propensity values into account. If, for example, the not-helix has a propensity near zero, indicating a very strong not-helix, it is still transformed into a helix. This heuristic assumes that small gaps in the prediction are due strictly to spurious errors.

When viewing helices as a physical structure, it is very possible for a single amino acid within the structure to lower stabilization in the helix. This allows the inclusion of unmodified *H-H* patterns in the observed helices. It is expected that the not-helix in this pattern has a propensity slightly below the cutoff threshold. This case is corrected by the absolute-fill heuristic. Other common helix formations, such as the helix-turn-helix, may contain a single amino acid with a very low propensity which breaks an otherwise long helix. In this case, the absolute-fill heuristic will incorrectly connect the two helices.



## b. Scaled-fill heuristic

Similar to the absolute-fill, the scaled-fill heuristic transforms spurious not-helix predictions into helix predictions. But unlike the absolute-fill, the scaled-fill utilizes the raw propensity values. Just as a single cutoff threshold is applied to identify helix positions, a lower “filling” threshold is used to resolve spurious not-helices. When the raw propensity value of a not-helix in a *H-H* pattern is slightly below the cutoff threshold, the position is changed to a helix. But, if the raw not-helix value is far below the threshold then it is left unchanged. The filling threshold determines the definition of “slightly below” and “far below,” allowing *H-H* patterns in the prediction and representing a less extreme alternative to the absolute-fill.

## 2. Removal heuristics

To resolve cases of spurious helix classification, *-H-*, a removal heuristic is applied (Table 7). All cases of spurious helix classifications are removed regardless of the propensity value since a helix must be longer than a single amino acid. Therefore, *-N-*, *-M-*, and *-C-* are removed since there cannot be a terminal or middle without a helix.

The filling and removal heuristics introduce two issues to the prediction system: heuristic effectiveness and heuristic ordering in ambiguous cases. These heuristics are expected to include at most no additional error in the prediction. Assuming a random prediction with helices occurring 35% of the time, *-H-* patterns will occur approximately 15% of the time, while the *H-H* pattern occurs less than 8% of the time. Based on this, the removal heuristic is expected to be applied more often than the filling heuristics, removing single amino acid helices from the prediction and increasing the correctness of the random prediction. Since predictions are not considered random, it is expected that these heuristics will, at worst, introduce about the same number of incorrect and correct predictions.

**Table 7.** *High-level knowledge-based rule set*

Heuristic	Simple rule	Transition rule
Filling heuristics <sup>a</sup>	<i>H-H - HHH</i> <i>N-N - NNN</i> <i>M-M - MMM</i> <i>C-C - CCC</i>	<i>N-M - NNM, NMM</i> <i>M-C - MMC, MCC</i> <i>N-C - NNC, NCC, NMC</i>
Removal heuristics	<i>-H- - ---</i> <i>-N- - ---</i> <i>-M- - ---</i> <i>-C- - ---</i>	
Trimming heuristics	<i>NNN - -NN</i> <i>CCC - CC-</i>	

<sup>a</sup> The filling heuristics represent absolute-fill rules. For the scaled-fill rules, each transformation is compared with the scaled-fill threshold.

Heuristic ordering must be considered to resolve ambiguous cases where the order of application can cause different prediction results. Since the removal heuristic, which reduces over-predictions, is expected to occur more often than the absolute- and scaled-fill heuristics, preference is given toward over-predictions. When predicting helices, we assume that an over-prediction is better than an under-prediction. To resolve ambiguous cases of *H-H-* and *-H-H*, the removal heuristics are applied after the filling heuristics.

### 3. *Trimming heuristics*

The helix structure is made stable by hydrogen bonding between the helix loops. Each helix loop contains 3.6 amino acids. The trimming heuristic uses this information to shorten extraneous helix terminals.

The Bayesian inference system determines the likelihood of helix terminals and middles. If a terminal is found to be longer than three or four amino acids, then the extra amino acids are considered excessive and removed from the prediction (Table 7). More than three or four amino acids in a predicted terminal are expected to

be over-predictions. Since there are 3.6 amino acids per turn, there is no place for the excess amino acids to bond. For example, suppose the system predicts many N-terminals followed by the remainder of the helix. This would correlate to a large number of bonds from the N-terminal to a middle or C-terminal, which in reality either do not exist or are too far away to form stable bonds.

The trimming heuristic is similar to the Ncap and Ccap hypothesis (Aurora et al., 1994), in which specific terminating sequences exist at the ends of helices. Only a small terminating sequence is necessary for stabilizing the helix; longer Ncap or Ccap regions are extraneous and can be omitted.

#### 4. Confirmation heuristic

The confirmation heuristic is used to validate helices using known structural information. Based on the likely N-terminals ( $N$ ), C-terminals ( $C$ ), and middles ( $M$ ) determined by the Bayesian inference method, a helix region is expected to be ordered from the N-terminal to the C-terminal;  $N$  regions before  $M$  and  $C$ , and  $C$  preceded by  $N$  and  $M$ . This is denoted by the regular expression:  $N^*M^*C^*$ , where the asterisk denotes "zero or more occurrences."

Because an  $\alpha$ -helix has 3.6 amino acids per turn, it is expected that the helix terminal is no shorter than two amino acids. Under this hypothesis, the predicted helix pattern must have at least two amino acids in the N-terminal and at least two amino acids in the C-terminal. In between the terminals, there can be any number of  $N$ 's,  $M$ 's, and  $C$ 's as long as they are found in order. The shortest possible helix pattern is expected to be  $NNCC$ , matching the smallest helices containing four amino acids, while the general helix pattern is  $NN(N^*)(M^*)(C^*)CC$ . Predictions that do not match the general pattern are not considered helices. For example, it is possible for a helix region to be predicted which contains only middle elements and no terminals. In this instance, it is hypothesized that the helix cannot form since it has no likely beginning or end. Similarly, a predicted helix region that is missing a single terminal

or is out of order is also not considered a helix.

When used in conjunction with the trimming heuristic, the confirmation heuristic shortens the terminals of the helix region definition to  $NN(M^*)CC$ .

### *5. Predicted helix verification heuristic*

The confirmation heuristic assumes that a predicted region matching the helix pattern correctly defines a helix. Since the components of the helix are determined independently, it is plausible for a region to match the helix pattern without forming a stable helix. For example, it may be possible for a predicted helix to have strong terminals and middles, but have them incorrectly rotated with respect to each other, causing structural instability. For example, in ambipathic helices the division of hydrophobic and hydrophilic amino acids along the helix wheel provides stability. However, if there is a rotation in the middle of the helix, it is plausible for the interactions between the misaligned ambipathic N-terminal and C-terminal to destabilize the helix.

The prediction verification heuristic views the helix as a whole, determining the likelihood that the entire structure is a helix. Because helices vary in length, each helix is scaled to a uniform size, from 0% at the N-terminal to 100% at the C-terminal. Rather than viewing specific positions in the helix, percentages within the helices are used.

The occurrence rate of each amino acid at a specific percent within the scaled helices of the training set are used to compute the probability of the predicted scaled helix. The likelihood of each amino acid's being at a specified percentage of the scaled helix is compared with a fixed threshold denoting the entire region's helical propensity. A predicted helix region will have a scaled helix propensity above the verification threshold.

### C. Helix pattern methodology implications

Heuristics based on the helix pattern methodology are expected to improve the Bayesian regional classifications. The filling heuristics are expected to improve  $Q_{predict}$  by allowing for marginal region classifications. Alternatively, the removal, confirmation, and verification heuristics should improve  $Q_{observe}$  by removing regional classifications.

## VIII. IMPLEMENTATION OF BAYESIAN INFERENCE SYSTEM WITH KNOWLEDGE-BASED POSTPROCESSING

This chapter focuses on a specific implementation of the Bayesian inference system for helix prediction, discussed in Chapter VI, with the helix pattern heuristics described in Chapter VII. Specific choices for the system configuration are provided, and the effectiveness of the implemented prediction model is compared with other prediction systems. Although the novel systems in this chapter are meant more as a proof of concept than as a final prediction system, their performance and accuracy are comparable to other black-box prediction systems.

### A. Regional probability computation

The Bayesian inference system is used to determine three helical regions which are expected to have different helical propensities. These regions,  $N$ ,  $M$ , and  $C$ , are determined by the observed occurrence rates of the amino acids in specific window positions. Because region  $M$  is expected to be fully contained in the helix, a symmetrical window of 5+1+5 amino acids is used. This assumes that the  $M$ -region amino acid is influenced by approximately 1.5 helix loops on each side. An amino acid is considered to be in the  $M$ -region only when the conditional probabilistic propensity is greater than the prior odds for being a helix:

$$M: P(\text{helix} | \text{window}[-5, +5]) > 0.35. \quad (21)$$

As shown by the neural network system in Chapter V, a symmetrical window is not necessarily optimal. For terminal regions, it is hypothesized that one or two amino acids beyond the helix terminal strongly influence the terminal formation. For terminal prediction, this hypothesis is combined with the conjecture that an amino acid in a helix is influenced by 1.5 helix loops on each side. The N-terminal region,

$N$ , is predicted when the conditional probabilistic propensity of the asymmetrical window 1+1+5 is larger than 0.20. Similarly, the C-terminal region,  $C$ , uses an asymmetrical window of 5+1+1:

$$N: P(N\text{-terminal}\backslash\text{window}[-1,+5]) > 0.20, \quad (22)$$

$$C: P(C\text{-terminal}\backslash\text{window}[-5,+1]) > 0.20. \quad (23)$$

## B. Implementation of Bayesian inference system with heuristic refinement

The results of applying the knowledge-based postprocessing methods to the Bayesian inference system are shown in Table 8. The standard threshold system determines helix location from a single cutoff value. The single propensity value is the same as the middle region,  $M$ , in the 3-region threshold system and is used for all heuristics. When applying the confirmation heuristic, the 3-region threshold system is used with the predicted  $N$ ,  $M$ , and  $C$  components.

### 1. Baseline: standard threshold

The baseline used for comparing heuristic effectiveness is a fixed-threshold conditional probabilistic system. This system is similar to other Bayesian approaches for helix prediction (Klinger & Brutlag, 1994; Goldstein et al., 1994) in which a single helix prediction,  $M$ , determines the helix location. Only the filling and verification heuristics are applicable since the trimming and confirmation heuristics require identification of terminal regions.

The standard threshold prediction system is 69.3% accurate ( $Q_1$ ) but frequently over-predicts helices, as illustrated by the low  $Q_{predict}$  value and the separation of 15.8% between  $Q_{predict}$  and  $Q_{observe}$ . A similar value in  $Q_{predict}$  and  $Q_{observe}$  would reveal noise in the prediction rather than a lack of prediction from the system.

**Table 8.** *Effectiveness comparison of heuristics*

Heuristic	$Q_I$	$Q_{predict}$	$Q_{observe}$	$Q_{observe} - Q_{predict}$	$C_{coef}$
<b>Single-region standard threshold</b>	69.6%	60.3%	76.1%	15.8%	0.4499
with filling heuristics	69.6%	60.3%	76.5%	16.2%	0.4506
with verification heuristic	70.7%	62.0%	71.2%	9.2%	0.4389
with filling and verification heuristics	70.6%	61.8%	72.1%	10.3%	0.4397
<b>3-Region threshold using confirmation heuristic</b>	70.6%	63.4%	74.3%	10.9%	0.4393
with filling heuristics	70.4%	63.0%	75.2%	12.2%	0.4374
with trimming heuristics	71.5%	64.8%	73.9%	9.1%	0.4523
with verification heuristics	70.9%	63.7%	74.0%	10.3%	0.4416
with filling and trimming heuristics	71.4%	64.4%	74.8%	10.4%	0.4516
with filling, trimming, and verification heuristics	71.4%	64.5%	74.4%	9.9%	0.4517

The 15.8% difference shows a potential problem with the prediction system due to lack of prediction, and not random error from the prediction model.

## 2. Application of the confirmation heuristic

When using a three region prediction system, the confirmation heuristic becomes applicable for determining helical regions. Application of the helical pattern  $NN(N^*)(M^*)(C^*)CC$  increases the overall accuracy while the correlation coefficient decreases. This is due to the over-prediction and under-prediction ratios; while the number of over-predictions decreases, the number of under-predictions increases. The difference between  $Q_{predict}$  and  $Q_{observe}$  drops from 15.8% in the baseline to 10.9%, showing a smaller discrepancy between noise in the prediction and under-/over-predictions.



### 3. Application of the filling heuristics

The scaled-fill and absolute-remove heuristics were applied to the single threshold model and the three-region confirmation heuristic system. The scaled-fill cutoff threshold was set to 95% of the original threshold. For example, the  $M$  cutoff threshold of 0.35 used a scaled-fill threshold of 0.33, allowing values slightly below the absolute threshold to be filled in. Because the scaled-fill increases the number of helix predictions, it is expected to increase  $Q_{observe}$  by decreasing the number of under-predictions. Similarly, the absolute-remove heuristic is expected to increase  $Q_{predict}$  by decreasing the number of over-predictions.

While the single threshold system only considered patterns of  $H-H$  for the scaled-fill heuristic, the three-region model fills many different patterns. In the three-region model, spurious not-helix predictions may be transformed into any of the three prediction regions. For example,  $N-N$  can become  $NNN$  and  $M-M$  can become  $MMM$ . Mixed fill sequences, such as  $N-M$ ,  $N-C$ , and  $M-C$  are also considered. For the scaled-helix heuristic, a spurious not-helix with a terminal propensity of  $0.20 \times 95\%$  determines a terminal region and a middle propensity larger than  $0.35 \times 95\%$  determines a middle region.

By applying these predictions, there is a strong increase in the  $Q_{observe}$ . Each implementation of the fill heuristics increases the  $Q_{observe}$  value. This suggests that a large amount of error in the prediction system is due to spurious not-helix predictions.

Unlike  $Q_{observe}$ ,  $Q_{predict}$  does not increase when the filling heuristic is applied, and in some cases it slightly decreases. The decrease appears to be due to over-prediction caused by the scaled-fill heuristic. Because the amount of decrease is small (approximately 0.4% in most cases), it indicates that the amount of error removed by the removal heuristics is nearly equivalent to the error added by the scaled-fill heuristic.

Application of the filling heuristics appears to increase the amount of error due to non-prediction. The difference between  $Q_{observe}$  and  $Q_{predict}$  increases when the

filling heuristic is applied. This suggests a reluctance in the system for predicting helices, although the helices it predicts have a high accuracy. Since the difference does not decrease, the amount of error in the system due to random noise appears to be reduced.

#### *4. Application of the trimming heuristics*

The trimming heuristic is only applicable to the three-region prediction system since it is used to reduce the size of the terminal regions. This heuristic is always used with the confirmation heuristic, reducing the helix pattern from  $NN(N^*)(M^*)(C^*)CC$  to  $NN(M^*)CC$ .

As hypothesized, the application of the trimming heuristic increases the  $Q_{predict}$  metric by reducing the number of over-predictions. The  $Q_{predict}$  metric increases by about 1.4% while the  $Q_{observe}$  metric decreases by about 0.4%, indicating that the amount of error introduced by the removal of real helix terminals is much less than the amount of incorrect helix predictions that were removed. In all cases, the trimming heuristic appears to lessen the difference between  $Q_{observe}$  and  $Q_{predict}$  indicating less error due to non-predictions than due to noise.

The correlation coefficient appears to increase due to the trimming heuristic, not the filling heuristic. This suggests that the baseline system generally misses helices by predicting incorrect terminal locations.

When the trimming heuristic is used in conjunction with the filling heuristic, all accuracy metrics increase. This suggests that the effects from the two heuristics are independent and mutually beneficial to the prediction model.

#### *5. Application of the helix verification heuristic*

After the three-region prediction model's confirmation heuristic determines a helix location, the helix verification heuristic is applied. This heuristic views the

entire helix as a whole and determines the overall likelihood of formation.

In implementation, the helix verification heuristic appears to accept nearly all helices. This heuristic increases  $Q_{predict}$  and decreases  $Q_{observe}$  by approximately the same amount, suggesting that the amount of error due to over-prediction is the same as the amount of under-prediction.

The helix verification heuristic only removes valid predictions and cannot increase the number of helix predictions. Because this heuristic reduces the difference between  $Q_{observe}$  and  $Q_{predict}$ , we can conclude that it only adds more noise to the prediction system by removing valid predictions and is generally not beneficial.

### C. Comparison with other prediction models

The results of the Bayesian inference model with three distinct regions and refinement heuristics has been compared with implementations of other prediction models. The published results from other prediction models use different data sets for determining accuracy. Also, few of the published works use all of the accuracy metrics. Thus, the accuracy of a particular method may vary due to the training set.

To set a baseline for the comparisons, Qain and Sejnowski's single-layer neural network system (1988) has been implemented on the data sets described in Chapter III and used by the Bayesian inference systems. This reimplemention is referred to as Reference 1. Although only the correlation coefficient is available from the original publication, it appears to be the same as the correlation coefficient in Reference 1. Reference 2 denotes the implementation of a modified single neural network system using the null amino acid and an asymmetrical window of 2+1+10. Reference 3 is the modified neural network system which includes the amino acid attributes in the system input, described in Chapter V.

With the exception of the Bayesian inference systems, each of the prediction models was originally designed to determine the main secondary structure: helices, sheets, coils, and in some cases, turns. The results presented in Table 9 represent only

**Table 9.** *Comparison of prediction models*

Prediction Model	$Q_I$	$Q_{predict}$	$Q_{observe}$	$C_{coef}$
<b>Statistical Models</b>				
Chou and Fasman (1974)	-	-	-	0.25
Robson and Suzuki (1978)	-	-	-	0.31
<b>NETtalk-based Neural Network Systems</b>				
Reference 1: Qain and Sejnowski	65.4%	58.1%	60.0%	0.34
Qain and Sejnowski, 1988 (1 net)	-	-	-	0.35
Reference 2: Qain and Sejnowski + null amino acid	67.9%	62.1%	54.6%	0.37
Qain and Sejnowski, 1988 (2 nets)	-	-	-	0.41
Holley and Karplus, 1989	63.2%	59%	-	0.41
Reference 3: input includes attributes	74.1%	72.5%	58.5%	0.46
<b>Bayesian Inference</b>				
Standard threshold	69.6%	60.3%	76.1%	0.45
3-regions + filling and trimming heuristics	71.4%	64.4%	74.8%	0.45
<b>Hybrid Systems</b>				
GOR (Garnier et al., 1978)	68.4%	-	-	0.48
DSC (King & Sternberg, 1996)	70.1%	-	-	0.58
Jury (Rost & Sander, 1993b)	-	72%	73%	0.60

the helix prediction components.

### *1. Comparison with statistical models*

Although the Bayesian inference system and classical statistical models are both used in probabilistic information, the Bayesian inference systems clearly have a significantly higher correlation coefficient. An explanation for this large difference involves the way the priors are determined and combined. For the Chou-Fasman (1974) and Robson-Suzuki (1978) systems, a specific physical factor, such as

hydrophobicity, was identified as being important. The probabilistic values strictly reflect the known physical factor. The combinational approach implements how these physical factors combine.

In contrast to the classical statistical models, the Bayesian inference system does not explicitly take physical attributes into account. Instead, this probabilistic system only uses the observed positions within known helices. The combinational method does not explicitly account for physical factors. It is assumed that any significant physical factors will affect the observed occurrences.

The performance gain by using the Bayesian inference system indicates that hydrophobicity, and other physical factors, are not necessarily the determining factors in helix identification.

## *2. Comparison with neural network models*

The neural network systems of Qain and Sejnowski (1988) and Holley and Karplus (1989), as well as the Reference systems, perform better than the classical statistical approaches, but not as well as the Bayesian inference systems. Although the Holley-Karplus system does appear more accurate than the Reference systems, this may be due to their testing method. As mentioned by Rost et al. (1993), Holley and Karplus did not cross-validate the predictions; the same data set was used for training and testing.

Although the neural network and Bayesian systems only use positional information, their combination methods are very different. A single neural network system cannot learn a generalized Bayesian combination. Rather than learning  $P(\text{helix} \mid \text{window of amino acids})$ , the single neural network can only learn  $P(\text{window of amino acids} \mid \text{helix})$ . A neural network with a single hidden layer can learn a Bayesian combination, but as shown by Qain and Sejnowski, there is currently an inadequate amount of data available for training a complex system.

An additional benefit of the Bayesian inference system is the step-wise

combinational approach. Neural network systems act as black-boxes, making it difficult to identify what factors directly affect the prediction. In the Bayesian inference system, each predicted position can be traced to the specific conditional probabilities and amino acids which became the determining factors.

### *3. Comparison with hybrid systems*

The hybrid systems combine many independent approaches to form a single prediction. The assumption is that each system's prediction is based on partially unique factors. By combining the prediction schemes, the hybrid system can benefit from the unique aspects of the individual components. Each of the hybrid systems performs substantially better than the Bayesian inference system. This is most likely due to the predictions generated. In the Bayesian inference systems, only helical positions are used. Thus, tertiary interactions, such as a sheet dominating a segment of the primary sequence, are explicitly ignored. In contrast, hybrid systems predict helices, sheets, and coils, allowing the system to determine the dominating factors. If the hybrid system can even marginally predict sheets or coils, then the helix prediction will significantly improve.

## IX. CONCLUSION AND FUTURE RESEARCH

The research presented in this dissertation focused on the application of computer science techniques in the field of theoretical biochemistry. This interdisciplinary study analyzed current black-box neural network systems and applied information from the analysis into a novel step-wise (white-box) prediction system. This system provided insight into the prediction process and performed comparably well with existing prediction models.

### A. Conclusion on basic neural network with sliding window

This research focused on the analysis of the NETtalk-based neural network system. Special emphasis was placed on identifying the aspects of the prediction problem that the system could and could not learn. In this work we have hypothesized and supported the following:

1. A consistent input window norm enhances the predictions at the protein terminus. This was supported by the use of a null amino acid which represents “no amino acid available.” Through the use of the null amino acid, terminal predictions increased the correlation coefficient by 0.05.
2. The neural network system places more importance on specific positions in the sliding window. Through weight matrix analysis, we have shown that the neural network clusters important window positions around the helix wheel; strong matches around the helix wheel appear to form helices.
3. When predicting the middle of the sliding window, the neural network does not necessarily place maximum importance on the center of the window. Training the system on “windows with holes” and observing the amount of loss in prediction accuracy, we have shown that more

importance is placed on window position +2 than on position 0. Additionally, we have shown that some positions, such as the extreme N-terminus positions, actually add noise into the system and result in less accurate predictions.

4. The neural network system does not place equal emphasis on the sides of the sliding window; a symmetrical sliding window is not necessarily optimal. By training the neural network on various asymmetrical window sizes, we have determined that more emphasis should be placed on the C-terminus than on the N-terminus. The optimal window size appears to have a two amino acid long N-terminus and a C-terminus with ten amino acids: a window of 2+1+10.
5. The information learned by training on specific amino acids and their positions is different and independent of the information learned by associating known propensity measurements with the window positions. This was demonstrated by training the neural network system with Chou-Fasman (1974), Fauchere-Pliska (1983), and molecular weight propensities. By combining these propensities with the amino acid in each window position, the prediction accuracy significantly increased.

## **B. Significance of the research**

The significance of this research impacts the fields of Computer Science and Theoretical Biochemistry. In the field of Computer Science, this research has identified approaches for extracting relevant information from a neural network system and relates this information with the problem domain. In addition, this research has developed a knowledge-based approach for refining Bayesian-based classifications.

The impact of this research in the field of Theoretical Biochemistry include



region-specific, position-dependent helical propensities. These propensities are used in a novel two-stage, step-wise prediction system. This system provides insight into the helix folding process and performs comparably well with existing black-box prediction systems. The insight provided by this system includes helix overlapping and adjacency.

### **C. Insight into the Bayesian prediction system**

The high accuracy and step-wise prediction process of the Bayesian inference system provides insight into the factors used to determine helix location. The most general hypothesis within this prediction system concerns helix structures. We hypothesize that each amino acid has different propensities towards the three regions in a helix. The data collected and the prediction system's high degree of accuracy strongly support this hypothesis.

#### *1. Helix overlapping and adjacency*

Sample output of the prediction system (Table 10) illustrates helix overlapping and adjacency. For example, in 1AAT (Torchinsky et al., 1982) a helix is observed between residues 312-341. This helix contains a bend at residues 319 (D) and 320 (N). The bend corresponds with the predicted adjacent helices, 311-319 and 320-342 (Fig. 21), denoted by the starting N-terminal at residue 320. In addition, residues 320-342 appear to contain overlapping helices, where the start of one helix (330-334) is fully contained in the middle of another helix.



**Fig. 21.** The protein 1AAT as displayed by Rasmol (Sayle, 1994). The helix along residues 312-341 is highlighted and the bend at residues 319 (Asp) and 320 (Asn) is labeled.

**Table 10.** Example of helix prediction: IAAT (Torchinsky et al. 1982)

	0	AASIFAAVPR	APPVAVFKLT	ADFREDGDSR	KVNLGVGAYR	TDEGQPWVLP	VVRKVEQLIA	GNGSLNHEYL	PILGLPEFRA	NASRIALGDD
P(N w)	NN	NN	N NN NNNNN		N N		NNN	-NN	-NNN N NN	NN
P(M w)	MMMMMM		M MMM MMMMMMMMM				MMMMMM		M M	MMM MMMMM
P(C w)	CC CC		CC C CCC	CC		C	CCCC --		C	CCCCCCC
Pred H	HHHHH		H HHHHHHHH				HHHHHHHH		HHHHHH	HHHHHHHHHH
Real H		Aaaaa	aaaa				Aaaaaaaaa	a	Aaaa	aaaa
	90	SPAIAQKRVG	SVQGLGGTGA	LRIGAEFLRR	WYNGNNNTAT	PVYVSSPTWE	NHNSVFM DAG	FKDIRTYRYW	DAAKRGLDLQ	GLLSDMEKAP
P(N w)	NNN		NN	NNNN			N	NnN NN	NNN N N NN	NNNNNN
P(M w)	MMMMMM			MMMMMMMMMM	M		M MMMM	MMMM	MMMM	MMMMMMMMMM
P(C w)		CCCC		CC C CCC-				C C		CCCCCCC
Pred H	HHHHHHHHHH			HHHHHHHH	HHH				HHH	HHHHHHHHHH
Real H	Aaaaa		Aaaa	aaaaaaaaaa			Aaaaaaaaa		Aaa	aaaaaaaa
	180	EFSIFILHAC	AHNPTGTDPT	PDEWKQIAAV	MKRRCLEPFF	DSAYQGFASG	NLEKDAWAVR	YFVSEGFELF	CAQSF SKNFG	LYNERVGNLS
P(N w)	N		--	NNNNNNNNN	NN		NN	NNNNN	NNNNN	N N
P(M w)	MMMMMMMM			MMMMMMMM	MMMM		MMMMMMMMMM	M MM MMMM	MM	Mm
P(C w)	C	CC C--		CCC	CCCCCC		C	CC C C C	CCC CC	CC
Pred H	HHHHHHHHHH	H		HHHHHHHHH	HHHHHHH		HH	HHHHHHHHH	H	
Real H			A	aaaaaaaaaa	aaaa		Aaaaaaaaa	aaaa		
	270	VVGKDEDNVQ	RVLSQMEKIV	RTTWSNPPSQ	GARIVATTLT	SPQLFAEWKD	NVKT MADRVL	LMRSELRSRL	ESLGT PGTWN	HITDQIGMFS
P(N w)	NNNNN	NNNN		-NN		-NNNN	N N NNNNN	NNNNNN N		
P(M w)	MMMM	MMMMMMMM			M		MMMMMM	MMMMMMMMMM	MMMMMMMMMM	M
P(C w)		CCCCCC	C			CCC	C C CC	CCCCCCCC	CCC-	
Pred H	HHHHHHH	HHHHHHHHH	H			HHHHHHHHH	HHHHHHHHH	HHHHHHHHH	HHH	
Real H	Aaaaa	aaaaaaaaaa	aaa		Aaaaaaaaa	a Aaaaaaa	aaaaaaaaaa	aaaaaaaaaa	aa	Aaaa
	360	FTGLNPKQVE	YMIKEKHIYL	MASGRINMCG	LTTKNLDYVA	KSIHEAVTKI	Q			
P(N w)	--NNN	NNN NN			NN	NN				
P(M w)	MMM	MMMMMMMM	MMMMMM		MMMM	MMMMMMMMMM	M			
P(C w)		CCCCCCcC	CCCC	C C		CCCC	C C C			
Pred H	HHHHH	HHHHHHHHH	HHHHH		HHHHH	HHHHHHHHH	H			
Real H	Aaaaa	aaaaaa			Aaaa	aaaaaaaaaa	a			

<sup>a</sup> The primary sequence uses the common amino acid single letter name. "N" and "C" show N- and C-terminals identified by the cutoff values, P(N) and P(C). Application of the filling heuristics are denoted by "n" and "c". "M" marks the predicted middle regions. Hyphens shows predictions removed by the trimming heuristics. "A" and "a" locate the start and length of the observed  $\alpha$ -helices.

## 2. Step-wise approach issues

Through the use of a step-wise approach, the exact factors necessary for determining each prediction are available. This approach has the potential for allowing rapid identification of potentially stable and unstable helices in a protein, as well as identifying the predicted cause of the structural formation. A step-wise prediction approach appears essential for reverse-engineering proteins and determining the effects of mutagenesis.

### D. Limitations of the Bayesian inference system and areas for future research

The Bayesian inference system makes a number of explicit and implicit assumptions concerning helix formation. This system assumes helix formation occurs independently with respect to other secondary structures. The high number of false-helices which coincide with other secondary structures suggests that this is a poor assumption. Although the actual amount of tertiary structure interaction is as yet unknown, it may be possible for different secondary structures to dominate sections of the primary sequence. The high number of false-helices may also be due to other factors, such as an unidentified helix aspect. It is also plausible that the false-helices are actually correct for an initial minimum energy folding state, as suggested by Boczko and Brooks (1995).

The observed conditional probabilities correlate well with some known helical patterns, such as the helix wheel. Other attributes, including each amino acid's dipole moment within a helix, may also correlate with the observed data. If there is a correlation, then these physical factors may also be addressed by additional heuristics.

Each amino acid is assigned a common window size. For example, when using a window of 2+1+5, all amino acids use windows of 2+1+5. It may be possible for some amino acids to be optimal with larger or smaller windows. As shown by  $P(M_k | w_0)$  in Chapter VI, different amino acids have different widths of effect; some

are narrow and some are very wide. By combining each amino acid with different window sizes, it may be possible to increase the prediction accuracy.

Along with using a common window size, the Bayesian inference system weighs all regions equally. This assumption could be too general; it may be possible for some regions to dominate others, such as N-terminals having more importance than C-terminals.

The thresholding of the raw conditional probabilities into boolean attributes may also limit the effectiveness of the prediction system. The use of a fuzzy logic system instead of or in conjunction with the boolean thresholding and Bayesian inference may assist in increasing the predictions accuracy while maintaining the step-wise approach.

As shown in the neural network analysis, combining amino acid attributes with Bayesian inference system may improve the prediction. Additionally, hybrid prediction approaches, which combine multiple prediction systems, still out-performs the Bayesian inference system even though the Bayesian inference system appears to be more accurate than any single prediction method. Combining this system with a hybrid approach could dramatically increase the prediction's accuracy, but would lose its step-wise insight.

## REFERENCES

- Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein Data Bank. In: Allen FH, Bergerhoff G, Sievers R, eds. *Crystallographic databases- Information content, software systems, scientific applications*. Cambridge: Data Commission of the International Union of Crystallography. pp 107-132.
- Alber T, Bell JA, Dao-pin S, Nicholson H, Wozniak JA, Cook SP, Matthews BW. 1988. Replacements of Pro<sup>86</sup> in phage T4 lysozyme extended an alpha-helix but do not alter protein stability. *Science* 239:631-669.
- Alber T, Dao-pin S, Wilson K, Wozniak JA, Cook SP, Matthews BW. 1987. Contributions of hydrogen bonds of THR 157 to the thermodynamic stability of phage T4 lysozyme. *Nature* 330:41-47.
- Aurora R, Srinivasan R, Rose GD. 1994. Rules for  $\alpha$ -helix termination by glycine. *Science* 264:1126-1129.
- Aurora R, Rose GD. 1998. Helix capping. *Protein Sci* 7:21-38.
- Boczko EM, Brooks CL. 1995. First-principles calculation of the folding free energy of a three-helix bundle protein. *Science* 269:393-396.
- Branden C, Tooze J. 1991. *Introduction to protein structure*. New York: Garland Publishing, Inc.
- Chakrabarty A, Baldwin RL. 1995. Stability of alpha-helices. *Advances in Protein Chemistry* 46:141-176.
- Chothia C. 1992. One thousand families for the molecular biologist. *Science* 357:543-544.
- Chou PY, Fasman GD. 1974. Prediction of protein conformation. *Biochemistry*

13:222-245.

- Creighton TE. 1993. *Proteins: Structures and molecular properties 2nd edition*. New York: W. H. Freeman and Company. pp 1-113, 139-259.

Dayhoff MO, Barker WC, Hunt LT. 1983. Establishing homologies in protein sequences. *Methods in Enzymology* 91:524-545.

Delcher AL, Kasif S, Goldberg HR, Hsu WH. 1993. Probabilistic prediction of protein secondary structure using causal networks. In: *Proceedings of the Eleventh National Conference on Artificial Intelligence*. Washington, D. C.: AAAI Press. pp 316-321.

Engelman DM, Steitz TA, Goldman A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem* 15:321-353.

Fauchere JL, Pliska V. 1983. *Eur J Med Chem* 18:369-375.

Free Software Foundation, Inc. 1993. GNU diff utility, version 2.6 compiled for OS/2. Available as source code from <[www.fsf.org](http://www.fsf.org)>.

Goldstein R, Luthey-Schulten Z, Wolynes P. 1994. A Bayesian approach to sequence structure alignment algorithms for protein structure recognition. In: *Proceedings of the 27th Hawaii International Conference on System Sciences*. Los Alamitos, California: IEEE Computer Society Press. pp 306-315.

Granier J, Osguthorpe DJ, Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 120:97-120.

Hecht-Nielsen R. 1989. *Neurocomputing*. New York: Addison-Wesley Publishing Company.

- Hertz J, Krogh A, Palmer RG. 1991. *Introduction to the theory of neural computation*. Redwood City, California: Addison-Wesley Publishing Company.
- Holley LH, Karplus M. 1989. Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci USA* 86:152-156.
- Holm L, Sander C. 1993. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 233:123-138.
- Hunter L. 1993. Integrating AI with sequence analysis. In: Hunter L, ed. *Artificial intelligence and molecular biology*. Menlo Park, California: MIT Press. pp 259-288.
- Kabsch W, Sander C. 1984. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. *Proc Natl Acad Sci USA* 8:1075-1078.
- Kamtekar S, Hecht MH. 1995. The four-helix bundle: what determines a fold? *The FASEB Journal* 9:1013-1022.
- King RD, Sternberg MJE. 1996. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci* 5:2298-2310.
- Klingler TM, Brutlag DL. 1994. Discovering structural correlations in alpha-helices. *Protein Sci* 3:1847-1857.
- Koretke KK, Luthey-Schulten Z, Wolynes PG. 1996. Self-consistently optimized statistical mechanical energy functions for sequence structure alignment. *Protein Sci* 5:1043-1059.
- Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157:105-132.
- Laplace PS. 1812. *A philosophical essay on probabilities*. New York: Dover. pp 189.



Lathrop RH, Webster TA, Smith TF. 1987. Ariadne: Pattern-directed inference and hierarchical abstraction in protein structure recognition. *Communications of the ACM* 30:909-921.

Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405:442-452.

Matthews BW, Dahlquist FW, Maynard AY. 1973. Crystallographic data for lysozyme from bacteriophage T4. *J Mol Biol* 78:575.

Muskal SM, Kim SH. 1992. Prediction protein secondary structure content, a tandem neural network approach. *J Mol Biol* 225:713-727.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48:443-453.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85:2444-2448.

Poulos TL, Finzel BC, Howard AJ. 1987. High-resolution crystal structure of cytochrome P450CAM. *J Mol Biol* 195:687-757.

Qain N, Sejnowski TJ. 1988. Prediction the secondary structure of globular proteins using neural network models. *J Mol Biol* 202:865-884.

Reeck GR, Haën C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH, Zuckerkandi E. 1987. "Homology" in proteins and nucleic acids: A terminology muddle and a way out of it. *Cell* 50:667.

Robson B, Suzuki E. 1978. Conformational properties of amino acid residues in globular proteins. *J Mol Biol* 107:327-356.

Rost B, Sander C. 1993a. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA* 90:7558-7562.

Rost B, Sander C. 1993b. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 232:584-599.

Rost B, Schneider R, Sander C. 1993. Progress in protein structure prediction? *TIBS* 18:120-123.

Salys R. 1994. *RasMol 2.5: Molecular graphics visualisation tool for Windows 3.1*. BioMolecular Structures Group, Glaxo Research & Development. Greenford, Middlesex, UK.

Sejnowski TJ, Rosenberg CR. 1986. NETtalk: A parallel neural network that learns to read aloud. *John Hopkins University EE & CS Technical Report* JHU/EECS-96/01.

Smith HO, Annau TM, Chandrasegaran S. 1990. Finding sequence motifs in groups of functionally related proteins. *Proc Natl Acad Sci USA* 87:826-830.

Smith RF, Smith RF. 1990. Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc Natl Acad Sci USA* 87:118-122.

Srinivasan R, Rose G. 1995. Linus: A hierarchical procedure to predict the fold of a protein. *Proteins Struct Funct Genet* 22:81-99.

Teeter MM, Hendrickson WA. 1984. Highly ordered crystals of the plant seed protein Cranbun. *J Mol Biol* 127:219.

Torchinsky YUM, Harutyunyan EG, Malashkevich VN, Kochkina M, Makarov VL, Braunstein AE. 1982. Aspartate aminotransferase from chicken heart cytosol, three-dimensional structure and coenzyme reorientations in the active site. *Clin Biol Res* 102:13.

Waterman MS, Eggert M, Lander E. 1992. Parametric sequence comparisons. *Proc Natl Acad Sci USA* 89:6090-6093.

## APPENDIX

### A. Data sets

The data sets used in this research are from the Brookhaven Protein Databank (Abola et al., 1987) release number 62, October 1992. Although many proteins have been added, or in some cases refined, since 1992, no effort has been made to update this data set.<sup>9</sup> Modifying the data sets to be current is not expected to have a significant impact on the findings of this research. In particular, in 1992 the Protein Databank (PDB) was expected to contain no more than 5% error in sequencing. Four years later, less than 5% of the existing proteins have been updated with corrections.

The entire data set contains 483 single-strand proteins containing nearly 100,000 amino acids. Of these, 40,363 amino acids are in unique contexts and 143 proteins are composed entirely of unique windows of seven amino acids. The testing set, shown in List 1, contains 50 completely unique proteins with respect to the entire data set.<sup>10</sup> This set is only used during the testing phase.

The residue numbering throughout the PDB is not consistent; some proteins begin with residue ID "1" while others begin with residue ID -5, 15, or some other number. To simplify the data set, all protein residues have been renumbered to begin with ID "1." This renumbering only affects the amino acid residue ID and not the sequencing or location of secondary structures.

Uniqueness was determined through the use of the GNU "diff" application for comparing two files (Free Software Foundation, 1993). GNU diff works in a similar fashion to the MAT and PAM protein alignment matrices (Needleman & Wunsch,

---

<sup>9</sup>In October, 1992 there were approximately 1000 proteins in the PDB. By January, 1997 the number of proteins increased to nearly 7000.

<sup>10</sup>When two or more proteins are similar, one is considered to be "unique" while the others are considered homologous to the unique protein.

1970; Dayhoff et al., 1983) except that two different amino acids are considered different; diff allows no partial relations based on similarity between codons.

**List 1.** *Proteins and unique context locations in testing set*

PDB protein ID	Number of residues	Unique context residues
1ALI	13	1-13
2ETI	28	1-28
1GCN	29	1-29
2CBH	36	1-36
1ATF	37	1-37
2SH1	48	1-48
4TGF	50	1-50
9PTI	58	1-58
1HCC	59	1-59
1C5A	66	1-66
1UTG	70	1-70
1UBQ	76	1-76
3FLX	79	1-79
2FXB	81	1-81
1MLI	96	1-96
1APS	98	1-98
3WRP	101	1-101
4RNT	104	1-104
1ACX	107	1-107
5CPV	109	1-109
3C2C	112	1-112
2RHE	114	1-114
2MHR	118	1-118
1ALC	122	1-122
1PHY	126	1-126
8LYZ	129	1-129
4FXN	138	1-138
5MBA	147	1-147
2LH7	153	1-153
3DFR	162	1-162
5P21	166	1-166
1RBP	175	1-175
8DFR	186	1-186
2ACT	218	1-218
3DPA	218	1-218
3PGM	230	1-230
1TON	231	1-231
2PNP	289	1-289
4CCP	293	1-293
9ABP	305	1-305
6CPA	307	1-307
7TLN	316	1-316
5PEP	327	1-327
1ALD	363	1-363
6XIA	387	1-387
2PHH	391	1-391
3AAT	405	1-405
9ICD	414	1-414
6CTS	433	1-433
4GRI	461	1-461

The training set contains 433 proteins. Some of training proteins have no unique windows of amino acids, and other proteins have spurious unique contexts. The unique context locations, shown in List 2, represent the residues which, at position 0 of a 7+1+7 window, compose a unique context.

**List 2. Proteins and unique contexts used in the training set**

PDB protein ID	Number of residues	Unique context locations
1XY2	8	1-8
1ZNF	26	1-26
3CTI	29	1-29
1CBH	36	
1PPT	36	1-36
1BDS	43	
2BDS	43	1-43
1ATX	46	1-46
1CRN	46	1-46
1SH1	48	
2HIR	49	
4HIR	49	
5HIR	49	47-49
6HIR	49	1-49
6RXN	53	1-53
1ROP	56	1-56
2OVO	56	1-56
1BUS	57	
2BUS	57	1-57
6PTI	57	57
4PTI	58	1-7, 16-17, 25-28, 36, 45
5PTI	58	
7PTI	58	30, 51
8PTI	58	35
1DTX	59	1-59
1PI2	61	1-61
1NXB	62	1-62
3EBX	62	
5EBX	62	1-7, 23-26, 41-42, 50-55
1R69	63	1-63
1SN3	65	1-65
2CI2	65	1-65
2CRO	65	1-40, 48-65
1GF2	67	1-67
1CTF	68	1-68
1GF1	70	1-13, 26-45, 55-70
1PGX	70	1-70
1CTX	71	1-71
1HOE	74	11, 19-26, 40, 51-66
2AIT	74	
3AIT	74	
4AIT	74	1-74
3ICB	75	1-75
1FLX	79	
351C	82	
451C	82	1-82
1CC5	83	1-83
1HIP	85	1-85
3B5C	86	1-86
1LRP	89	1-89

**List 2. Continued**

PDB protein ID	Number of residues	Unique context locations
3FXC	98	1-98
1PCY	99	10-16, 25, 33-36, 46-47, 57-63, 85-87
2PCY	99	
3HVP	99	14, 25, 37-41, 63-67, 94-99
3PCY	99	
3PHV	99	1-99
4PCY	99	
5PCY	99	
6PCY	99	1-7, 17-30, 45-99
7PCY	100	1-100
1CYC	103	1, 12-17, 61-62, 70-75, 87-89, 103
1RNT	104	
2RNT	104	
3RNT	104	12, 39, 55-62, 85-92, 100-104
5CYT	104	1-104
1RMS	105	1-68, 78-84, 97-105
1FKF	107	1-107
1OMD	107	1-107
1YCC	107	101-107
2CDV	107	1-107
2SSI	107	1-107
2YCC	107	1-29, 37-44, 57-73, 85-107
1SRX	108	1-108
1CDP	109	
1PAL	109	
2PAL	109	
3PAL	109	
4CPV	109	8, 26, 35-40, 52, 60, 79, 91-99, 109
4PAL	109	1-29, 37-41, 49-50, 58-60, 72-109
1CCR	112	1-21, 30-78, 90-112
2C2C	112	
1HRB	113	1-96, 104-113
1APK	118	1-26, 34-53, 62-118
1CY3	118	1-118
1BPK	119	1-119
1PAZ	120	9-20, 29-37, 45-46, 54-55, 68, 78, 86-107, 120
2APK	122	1-122
3BP2	122	1, 70-73, 84
1BP2	123	62-66
2BP2	123	1-3, 59-61
2PAZ	123	1-123
4BP2	123	1-20, 40-72, 80-87, 119-123
1P2P	124	58, 67-72, 85
1RSM	124	
3FGF	124	1-124
3RN3	124	41, 61-64, 76-78, 88-92, 111-124
4P2P	124	1-124
5RSA	124	
6RSA	124	
7RSA	124	1-124
2FGF	127	1-12, 20, 41-60, 68, 104-109, 126-127



**List 2. Continued**

PDB protein ID	Number of residues	Unique context locations
2CHY	128	1-128
1LYZ	129	
2LYM	129	
2LYZ	129	
2LZ2	129	1-3, 15, 41, 73, 98-101, 121
3LYM	129	1-4, 16, 24, 79-84, 108, 119
3LYZ	129	
4LYM	129	36, 88, 97-101, 109-129
4LYZ	129	
5LYZ	129	
6LYZ	129	
1LHM	130	77, 95
1LZ1	130	86-92
2LHM	130	
3LHM	130	1-107, 117-130
1IFB	131	
2BPK	131	1-131
2IFB	131	1-131
1SNC	135	135
1ECA	136	
1ECD	136	
1ECN	136	
1ECO	136	1-136
1SNM	136	1-6, 23-37, 48, 62, 71, 92, 115, 129-136
3FXN	138	
1LE4	139	139
2SNS	141	1-141
3CLN	143	1, 143
1LE2	144	136
1LPE	144	1-144
1HBG	147	
1MBA	147	
2HBG	147	1-147
3MBA	147	
4MBA	147	
8I1B	147	1-147
2CLN	148	1-148
4CLN	148	26-28, 37-39, 54-64, 99-101, 127-148
2LHB	149	1-149
31BI	149	149
1I1B	151	1, 47-50, 106, 140
1RNH	151	1-151
21BI	151	69
41BI	151	1-6, 19, 31-46
4I1B	151	19, 31-46, 64, 83, 113-123, 134-151
5I1B	151	1-39, 49, 62, 71-75, 87, 104-114, 127-151
1LH1	153	
1LH2	153	
1LH3	153	
1LH4	153	
1LH5	153	

**List 2. Continued**

PDB protein ID	Number of residues	Unique context locations
1LH6	153	
1LH7	153	
1MBC	153	
1MBD	153	
1MBI	153	
1MBN	153	
1MBO	153	
1MBS	153	1-28, 45-66, 74, 118-132, 140, 151-153
2I1B	153	1-2, 16, 30, 39, 54, 63, 74, 86, 99-106, 126-138, 152-153
2LH1	153	
2LH2	153	
2LH3	153	
2LH4	153	
2LH5	153	
2LH6	153	
2MB5	153	95, 124, 149-153
4MBN	153	
5MBN	153	1-153
1MBW	154	1, 123
5DFR	159	
6DFR	159	16-59
7DFR	159	1-159
4TNC	160	1-2, 12, 27, 37-38, 64, 97, 113-114, 140, 149-160
5TNC	161	1-37, 47-161
1L36	162	
1L55	162	92
1L57	162	
1L59	162	
1L61	162	38
1L62	162	
1L63	162	
1L64	162	39-48
1L65	162	47
1L66	162	43
1L67	162	46
1L68	162	44
1L69	162	
1L70	162	
1L71	162	127-130
1L72	162	127
1L73	162	126-131
1L74	162	127
1L75	162	126-132
1L76	162	34, 54, 72, 162
4LZM	162	
5LZM	162	
6LZM	162	
7LZM	162	
1L01	164	155
1L02	164	157
1L03	164	157

**List 2. Continued**

PDB protein ID	Number of residues	Unique context locations
1L04	164	
1L05	164	157
1L06	164	157
1L07	164	157
1L08	164	157
1L09	164	157
1L10	164	157
1L11	164	157
1L12	164	157
1L13	164	157
1L14	164	157
1L15	164	157
1L16	164	156
1L17	164	1-3
1L18	164	1-3
1L19	164	38
1L20	164	144
1L21	164	55
1L22	164	
1L23	164	77
1L24	164	82
1L25	164	86
1L26	164	86
1L27	164	86
1L28	164	86
1L29	164	86
1L30	164	86
1L31	164	86
1L32	164	86
1L33	164	
1L34	164	96
1L35	164	9, 164
1L37	164	
1L38	164	
1L39	164	
1L40	164	
1L41	164	83, 112
1L42	164	
1L43	164	16
1L44	164	
1L45	164	
1L46	164	
1L47	164	154
1L48	164	
1L49	164	
1L50	164	
1L51	164	98, 149
1L52	164	152
1L53	164	149
1L54	164	
1L56	164	60

**List 2. Continued**

PDB protein ID	Number of residues	Unique context locations
1L58	164	143
1L60	164	
1LYD	164	1-164
2LZM	164	
3LZM	164	
2APD	169	1-169
4Q21	169	15, 26, 37, 58-75, 84, 104-117, 126, 140-151, 165-169
1TIE	170	1-170
1Q21	171	
2Q21	171	12, 170-171
1CD4	173	1-25, 33-34, 47, 58-64, 72, 112, 120-121, 132-136, 147-155, 163-173
2FCR	173	1-173
1GCR	174	1-174
2CD4	177	1-177
1FHA	180	1-156, 166-180
2STV	184	1-184
1PPD	212	43, 57, 67, 108-117, 128, 138, 169, 206-212
9PAP	212	1-212
1CLA	214	
2CLA	214	155, 194
3CLA	214	65, 155
4CLA	214	1-214
1BRD	219	1-56, 68-120, 130-185, 194-219
1SGC	227	
2SGA	227	1-227
1TGN	229	
2TGD	229	
1EST	230	
1NTP	230	33, 80, 100
1SGT	230	1-175, 183-195, 207-230
1TGB	230	
1TGC	230	
1TGT	230	
1TLD	230	
1TPO	230	
1TPP	230	
2PTN	230	
2TGA	230	
2TGT	230	
2TRM	230	8-11, 34, 47, 56-115, 130-168, 187-189, 202-209, 221-230
3EST	230	
3PTB	230	
3PTN	230	1-5, 28-40, 48-52, 66-76, 88-92, 121-125, 141-146, 158-168, 182-200, 211-216
4PTP	230	1-230
6EST	230	101-156
8EST	230	1-230
2CNA	237	11, 23, 39-47, 67, 81-89, 103-107, 116, 131, 144-150, 179- 190, 200

**List 2. Continued**

PDB protein ID	Number of residues	Unique context locations
3CNA	237	1-237
1CHG	245	12-15, 231-234
2CHA	245	10-11, 28, 39-41, 49-69, 92, 101-110, 133, 141, 149-155, 163-166, 185-209, 218-231
2GCH	245	101-176
3GCH	245	
4GCH	245	
5GCH	245	
6GCH	245	
7GCH	245	1-192, 200-245
12CA	256	
1CA3	256	
1HCA	256	
1HEB	256	194
1HED	256	194
2CAB	256	1-99, 107-187, 196-256
4CA2	256	101-156
5CA2	256	101, 158-196
6CA2	256	101-156
7CA2	256	101-156
8CA2	256	101-156
9CA2	256	1-256
1CA2	257	
1HEA	257	195
1HEC	257	195
2CA2	257	
3CA2	257	1
3BLM	260	1-260
3TMS	264	1-27, 37-174, 185-264
1S02	275	19, 27, 56-61, 88-89, 98-99, 158-159, 178-180, 197-201, 239, 251
1SBC	275	1-59, 74-79, 88-89, 97-118, 129-172, 183-185, 194-217, 241- 259, 271-275
1SBT	275	11-13, 26-33, 63, 89-95, 132, 146-147, 175-177, 205-222, 238, 253, 269-275
2SBT	275	1-275
2PRK	279	1-279
1PYP	281	1-281
1CCP	293	151-193
1RHD	293	1-293
2CCP	293	234
2CYP	293	52, 151, 190-193
3CCP	293	50, 190-193
1FNR	296	
2FNR	296	1-296
1APB	305	
1BAP	305	
1CPB	305	1-82, 90-263, 271-305
3GBP	305	1-15, 23-305

**List 2. Continued**

PDB protein ID	Number of residues	Unique context locations
6ABP	305	
7ABP	305	
8ABP	305	101, 247-253
1ABP	306	108-206
5CPA	307	37-39, 47-48, 72, 89-93, 102-109, 190, 262
2GBP	309	1-3, 14, 26-44, 58-61, 70-71, 86, 95, 104-116, 127-143, 152, 163-184, 201-211, 225-238, 250-257, 274, 283-309
3TLN	316	
4TLN	316	
4TMS	316	1-316
5TLN	316	
1LDB	317	198, 212-217
2LDB	317	1-83, 93-122, 132-317
2TS1	319	
3TS1	319	1-319
3PFK	320	
4PFK	320	1-320
1LLC	322	1-129, 140-154, 162-180, 188-322
1CMS	323	1-3, 39-71, 87-89, 99-102, 125-129, 161, 173-175, 200, 221- 225, 238-245, 253-255, 271-275, 286-305
3APP	323	1-69, 78-118, 128-323
2APR	325	1-33, 41-325
3PEP	326	1-6, 41, 52, 163, 172-174, 199-204, 255, 275-289, 298, 325- 326
4PEP	326	1-24, 37, 51-78, 97-98, 118-132, 149, 168, 176-195, 209- 214, 229, 240-253, 263, 274-278, 307-318
3CMS	327	101, 241-252
4CMS	327	1-82, 91-148, 156-289, 298-299, 309-312, 320-327
1LDM	329	1, 130-136, 184-192, 224, 237-247, 262-264, 274, 283, 294- 298
4APE	329	1-329
6LDH	330	
8LDH	330	1-330
2LDX	331	1-29, 40-157, 173-248, 256-331
3LDH	332	1-9, 22-48, 56-57, 68-71, 83, 98-110, 133-155, 163-166, 181- 266, 295-301, 311
2LIV	344	1-344
2LBP	346	1, 15-31, 42-48, 65, 76-77, 98-123, 135, 144, 157-165, 184- 186, 203-216, 229-231, 239-319, 330-346
3BCL	356	1-356
1GOX	360	1-188, 200-360
1PSG	371	1-49, 60-91, 101-114, 153-159, 169-179, 187-188, 199-206, 246-247, 268, 281, 289-293, 315-320, 342-361, 369-371
1MLE	372	1-19, 29-372
5ADH	374	9-10, 32-39, 88, 135, 159, 226-228, 354
7ADH	374	1-45, 54-78, 86-105, 113, 129-138, 148-168, 185-200, 212- 249, 259-269, 282-293, 304, 312-330, 339-354, 365-374

**List 2. Continued**

PDB protein ID	Number of residues	Unique context locations
8ADH	374	1-374
3XIA	377	1, 14-17, 31-45, 54-88, 101-105, 120-185, 195-238, 246-250, 277-295, 307, 323-331, 339-377
2XIS	386	
3XIS	386	8-19, 30-33, 43-52, 61-67, 79-88, 105-106, 126-141, 149, 173-178, 186-225, 234-246, 261-262, 274-286
1XIS	387	
4XIS	387	9-20, 31-34, 44-53, 62-68, 80-89, 106-107, 127-142, 150, 174-179, 187-226, 235-247, 262-263, 275-294, 320-322, 330-351, 360-370, 382-387
1PHH	394	101-394
2CPP	405	
3CPP	405	
4CPP	405	
5CPP	405	
6CPP	405	
7CPP	405	
8CPP	405	1-405
1AAT	411	1-411
3ICD	414	
4ICD	414	101-329
5ICD	414	
6ICD	414	101-314
7ICD	414	
8ICD	414	101-314
3PGK	416	1-416
1CSC	433	
2CSC	433	
3CSC	433	
4CSC	433	
5CTS	433	
3ENL	436	
4ENL	436	
5ENL	436	
6ENL	436	
7ENL	436	1-436
1CTS	437	32-33, 41, 69, 77-85, 104-105, 163-175, 196, 283-300
2CTS	437	9-12, 30-33, 41, 69, 77-85, 104-105, 163-175, 196, 283-300, 343, 366-368, 432-437
1TPT	440	1-440
3GRS	461	
2TAA	478	1-478
7CAT	498	1-498
1COX	502	1-502
1ACE	531	1-531
5ACN	754	
6ACN	754	1-754
1GPB	823	101-723

**List 2. Continued**

PDB protein ID	Number of residues	Unique context locations
6GPB	828	101-828
2GPB	831	101-731
8GPB	832	1-832
3GPB	833	
4GPB	833	
5GPB	833	101-833

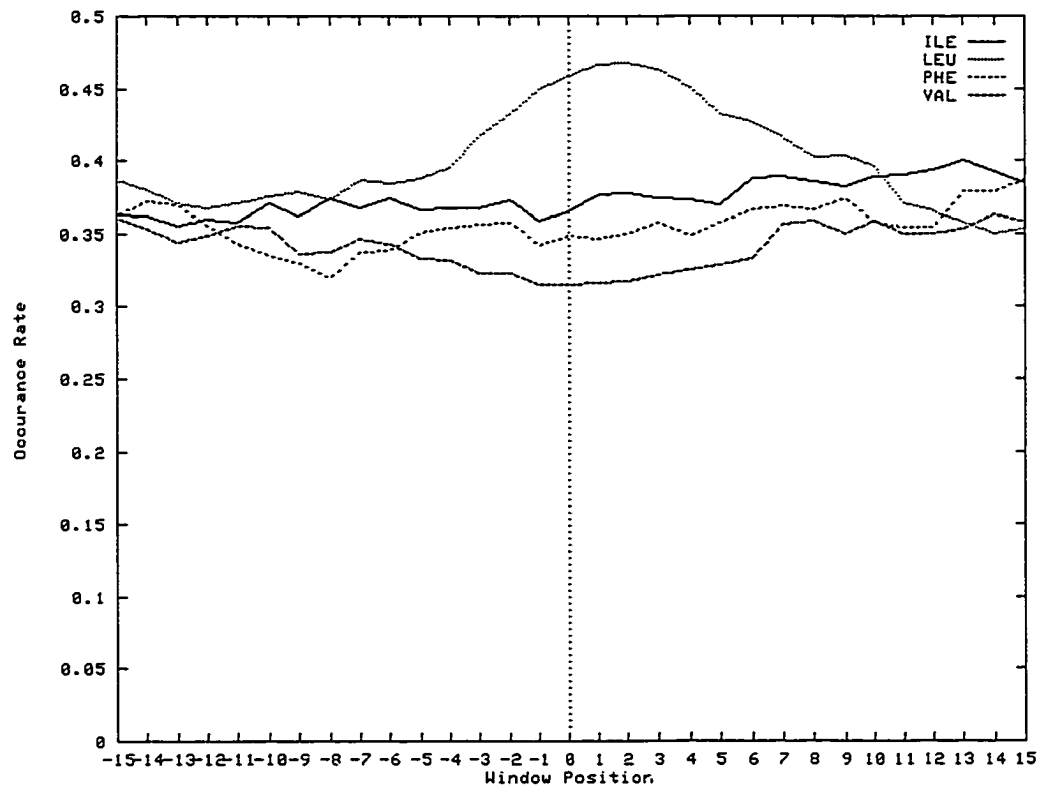


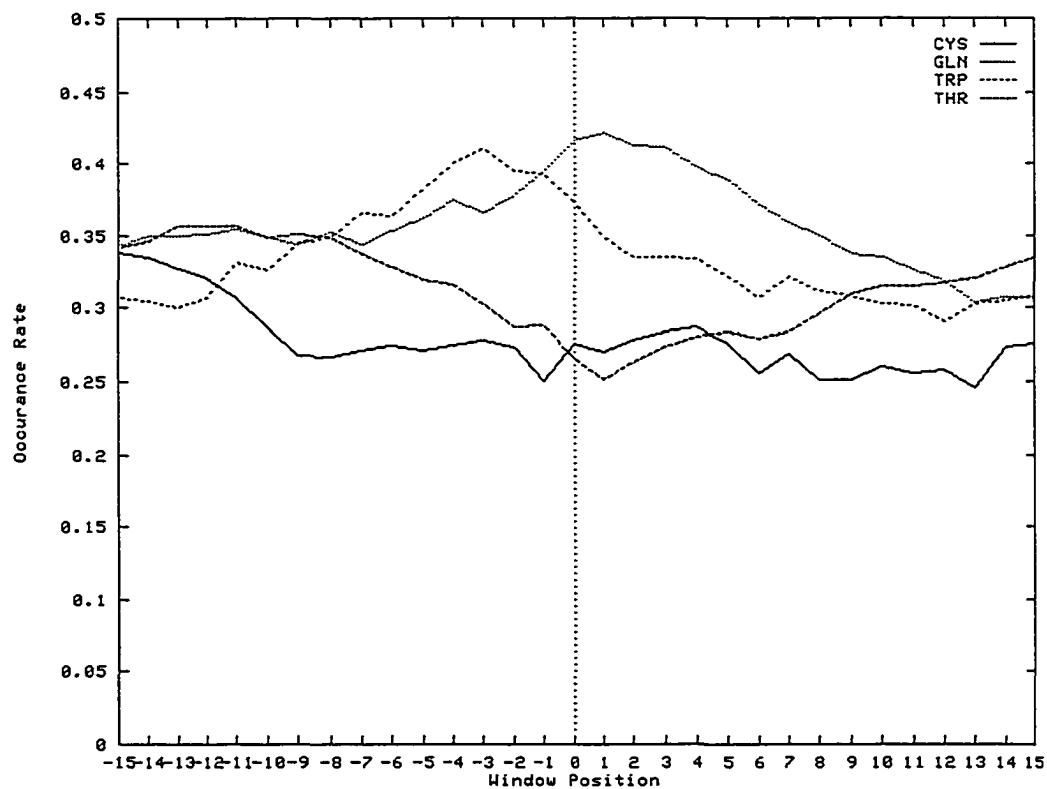
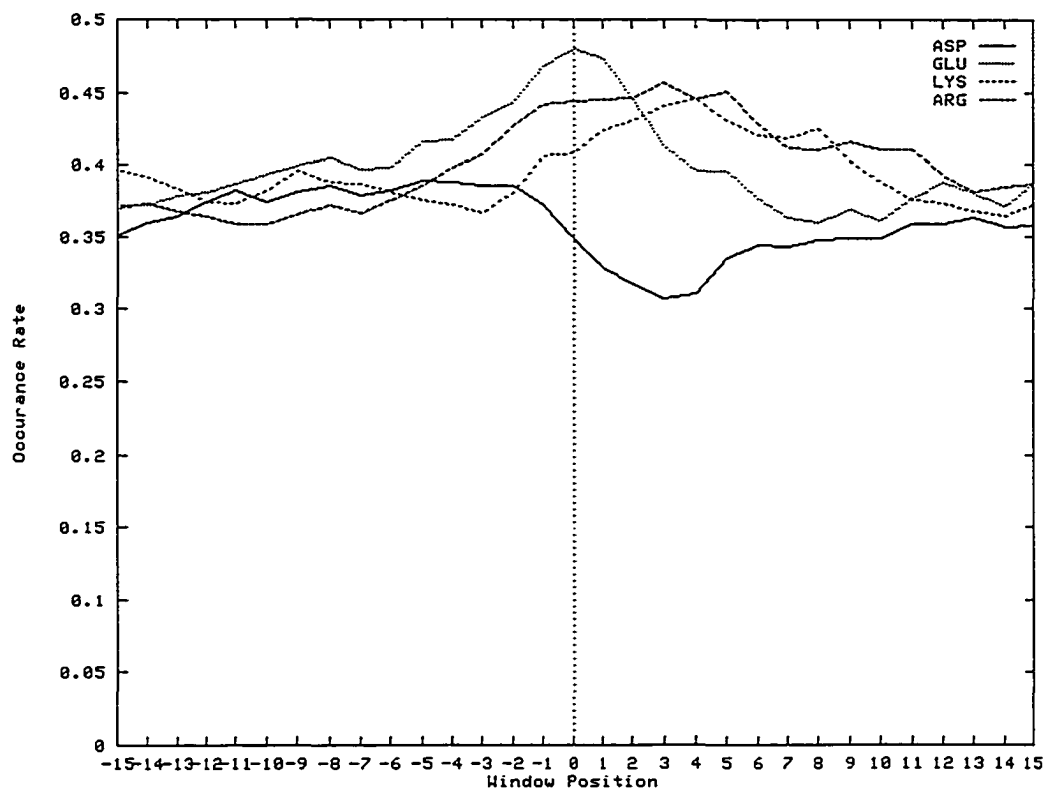
## B. Conditional probabilistic propensities

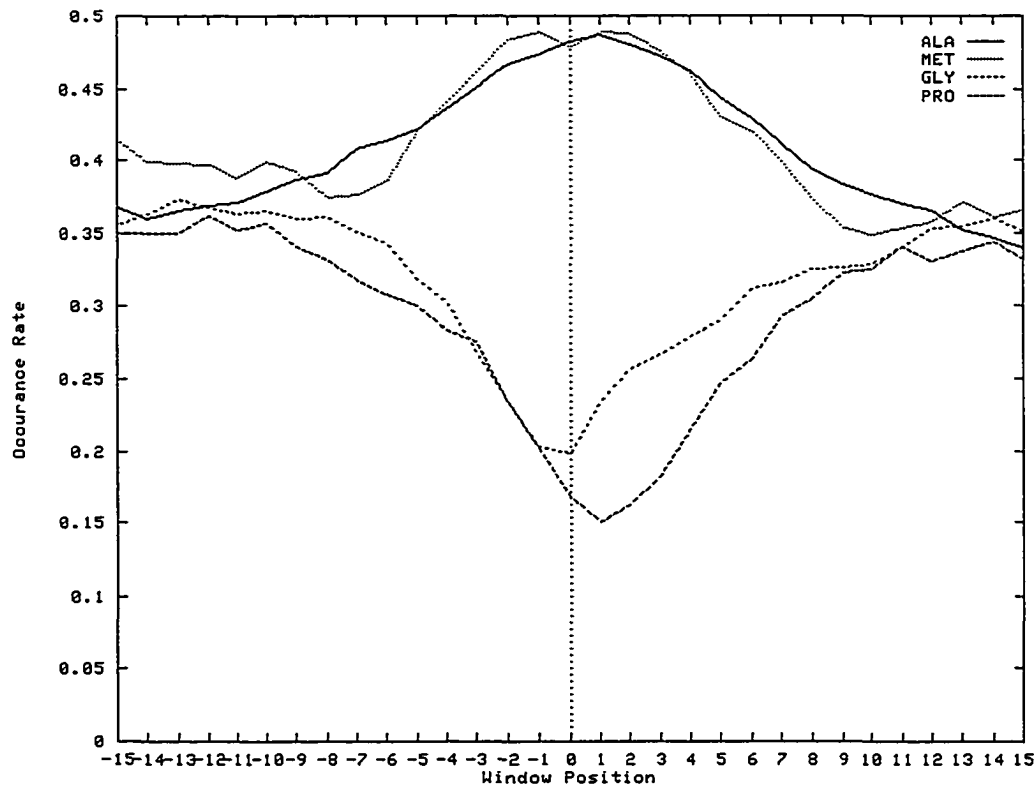
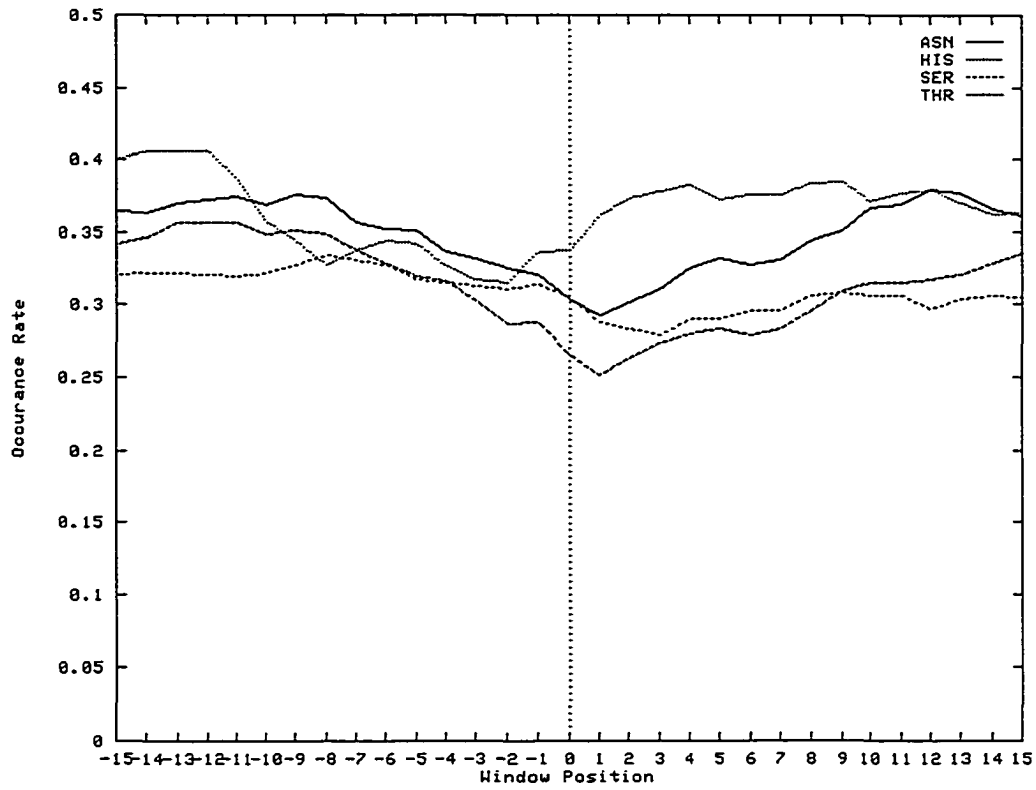
The following graphs show the helical propensities for each amino acid in each of the helix regions. The regions used are: helix ( $M$ ), N-terminal ( $N$ ), C-terminal ( $C$ ), and scaled helix ( $S$ ). The propensities for each of the twenty amino acids are presented. For readability, the amino acids have been clustered into subgraphs based on similar propensity characteristics: hydrophobics (Ile, Leu, Phe, Val), charged (Asp, Glu, Lys, Arg), polar (Cys, Gln, Trp, Tyr, and Asn, His, Ser, Thr), and “others” (Ala, Met, Gly, Pro).

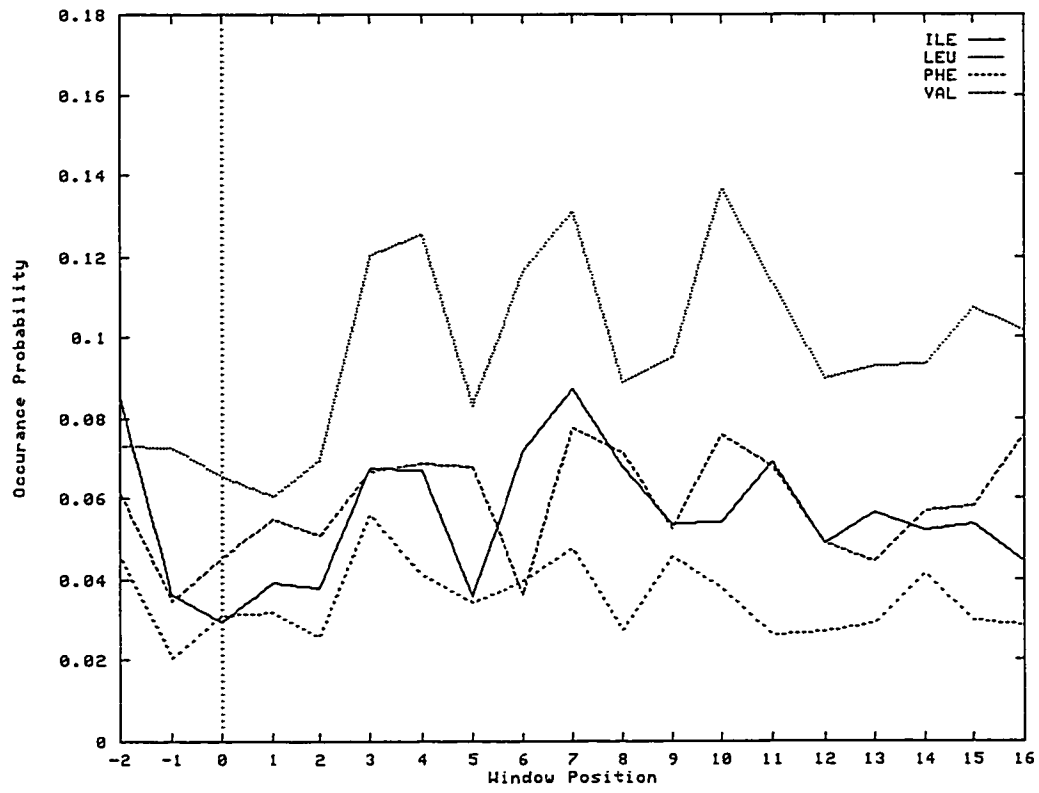
Each figure presents the likelihood of a region given a specific window position,  $k \in [-15, +15]$ . The four conditional probabilities are:  $P(M \text{ at } 0 \mid w_k)$ ,  $P(w_k \mid N \text{ at } 0)$ ,  $P(w_k \mid C \text{ at } 0)$ , and  $P(w_k \mid \text{scaled helix range } s \in [-10\%, +100\%])$ .

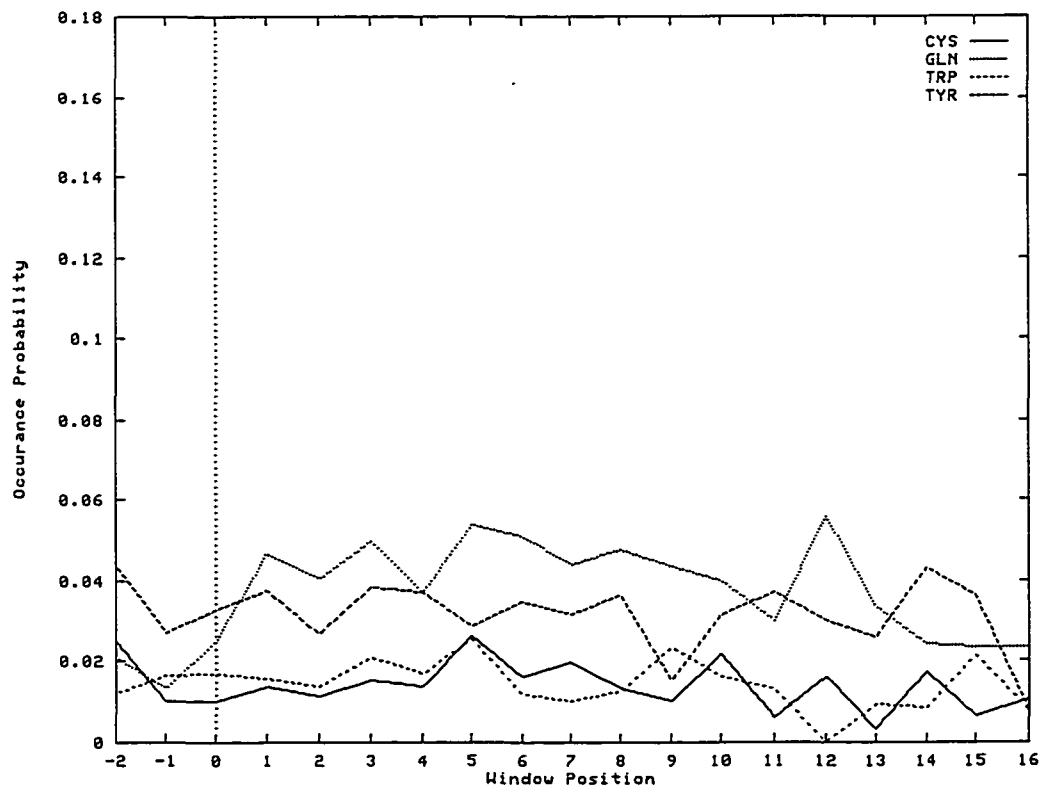
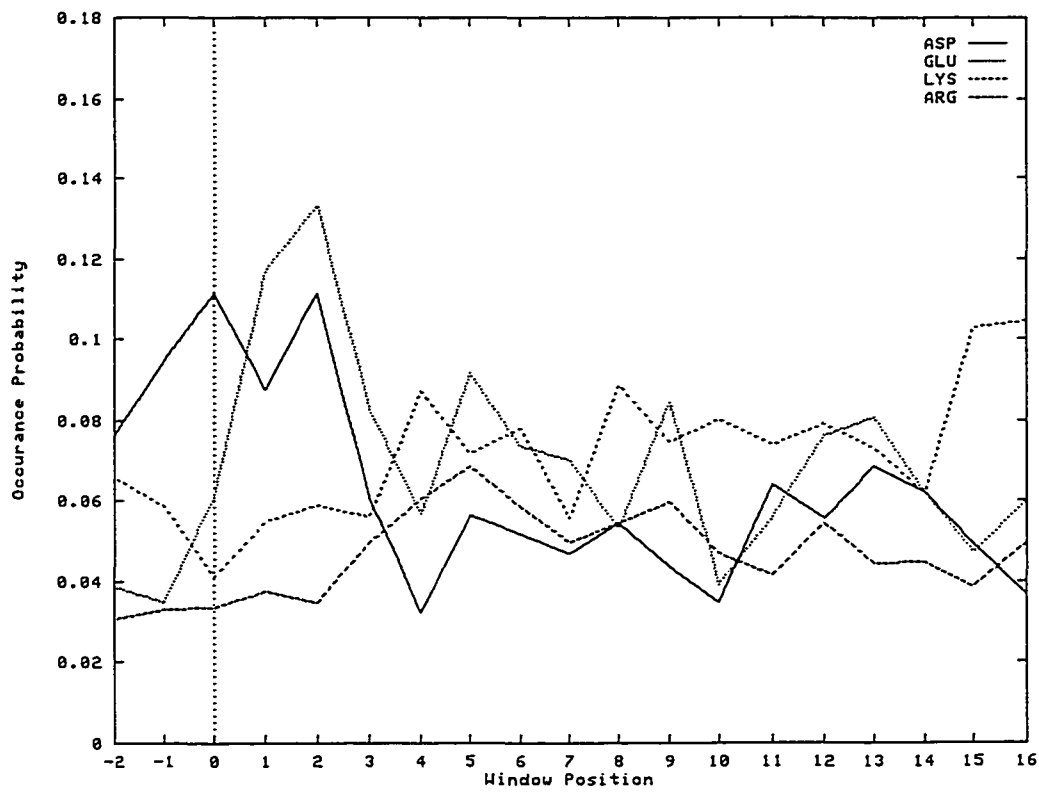
1.  $P(M \text{ at } 0 | w_k)$

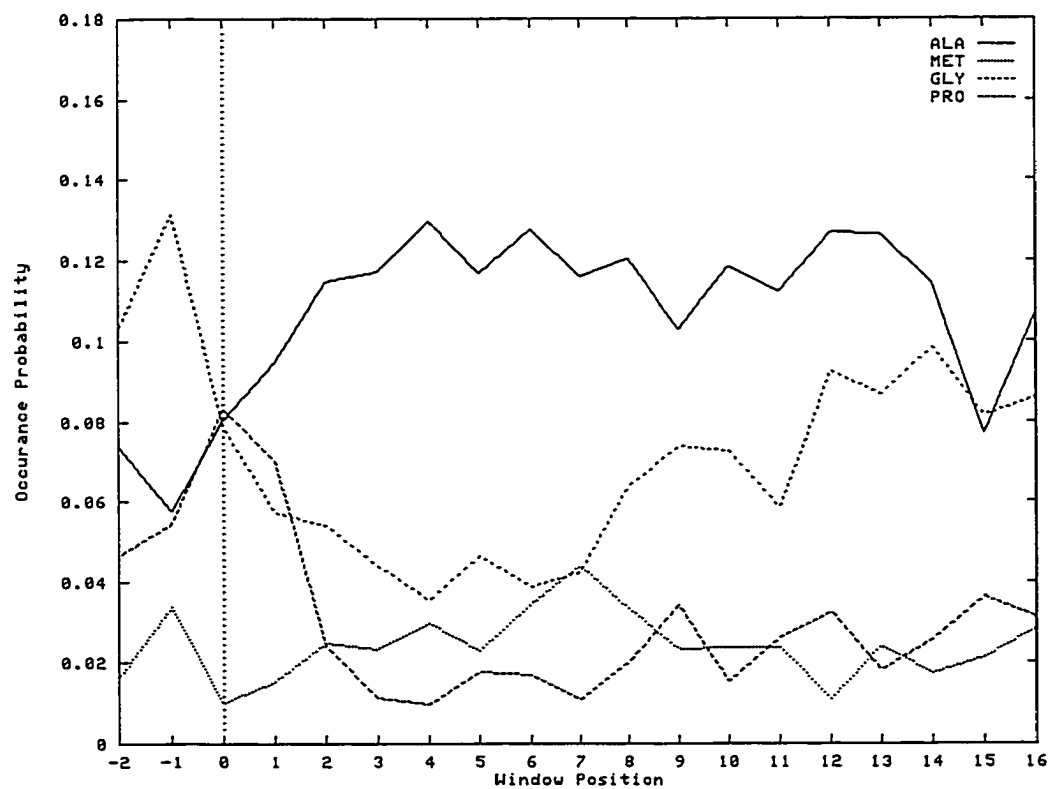
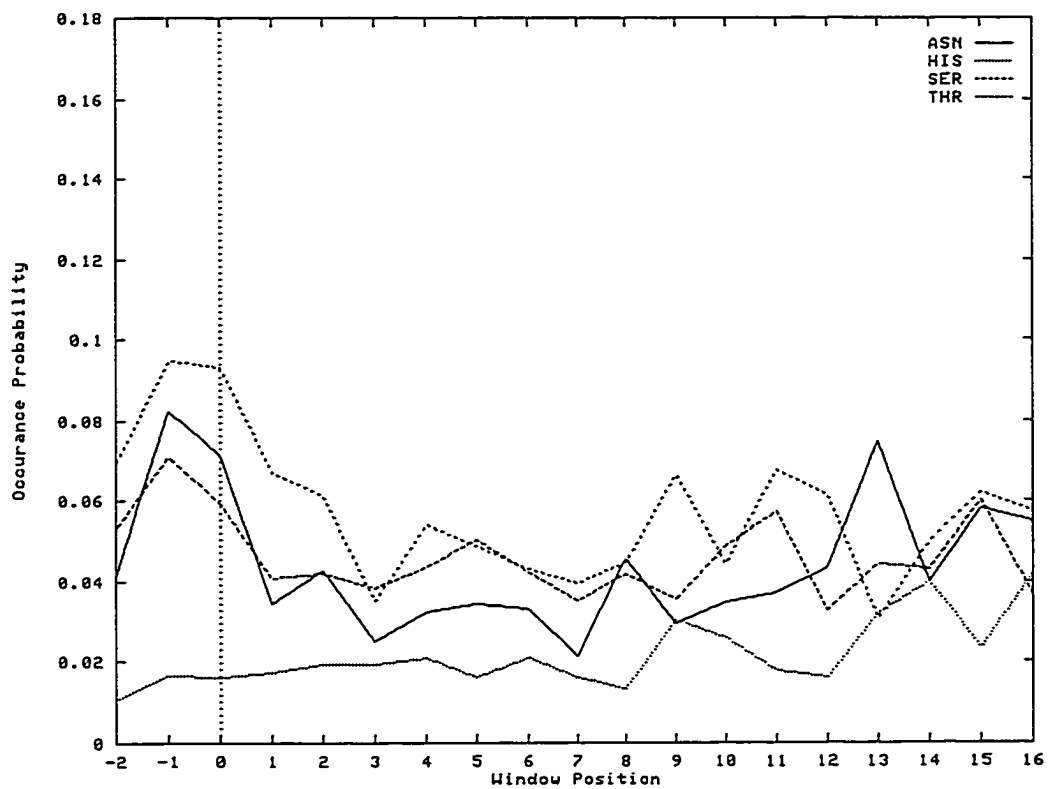


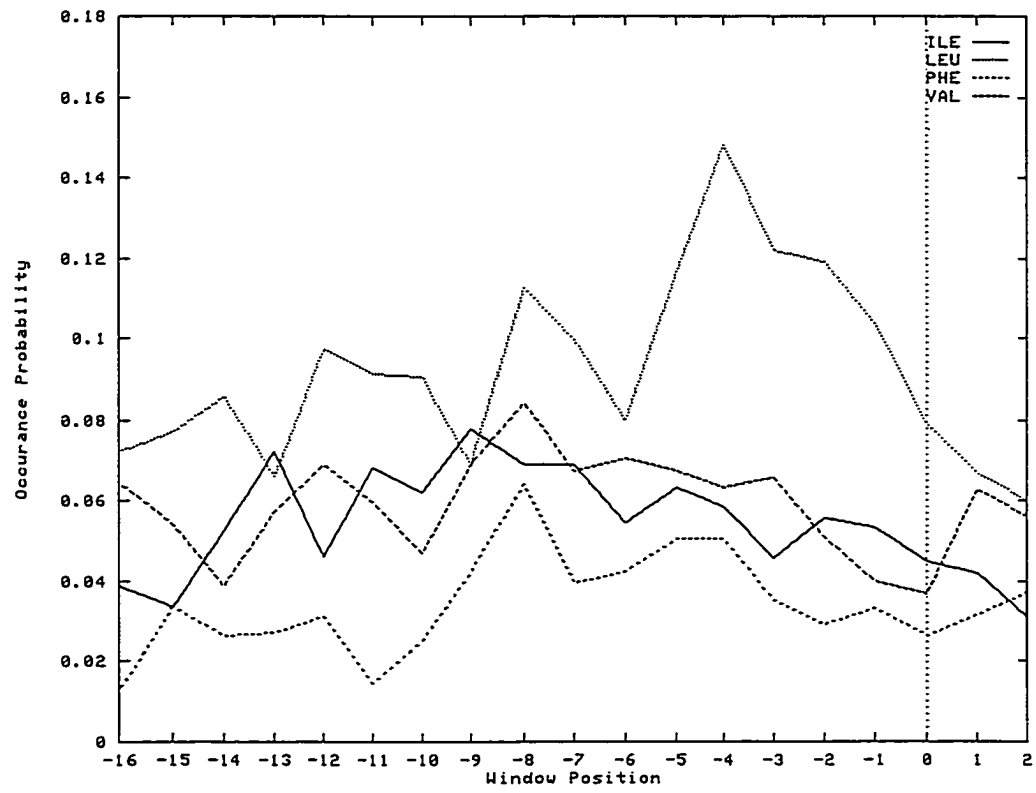




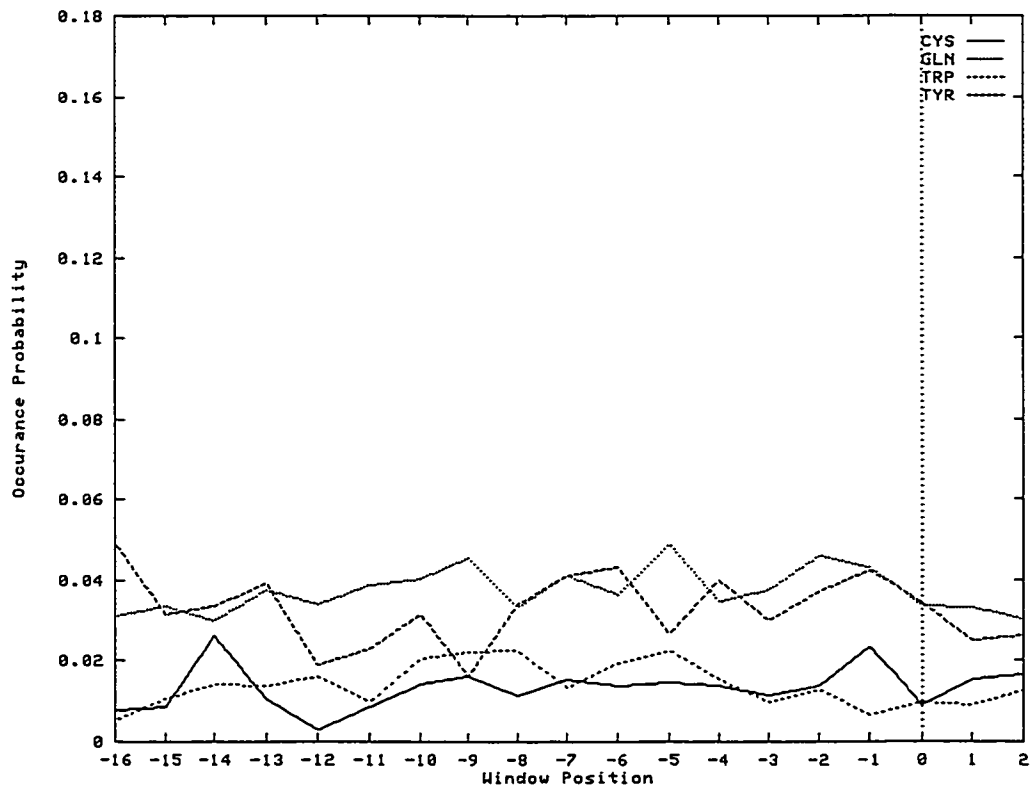
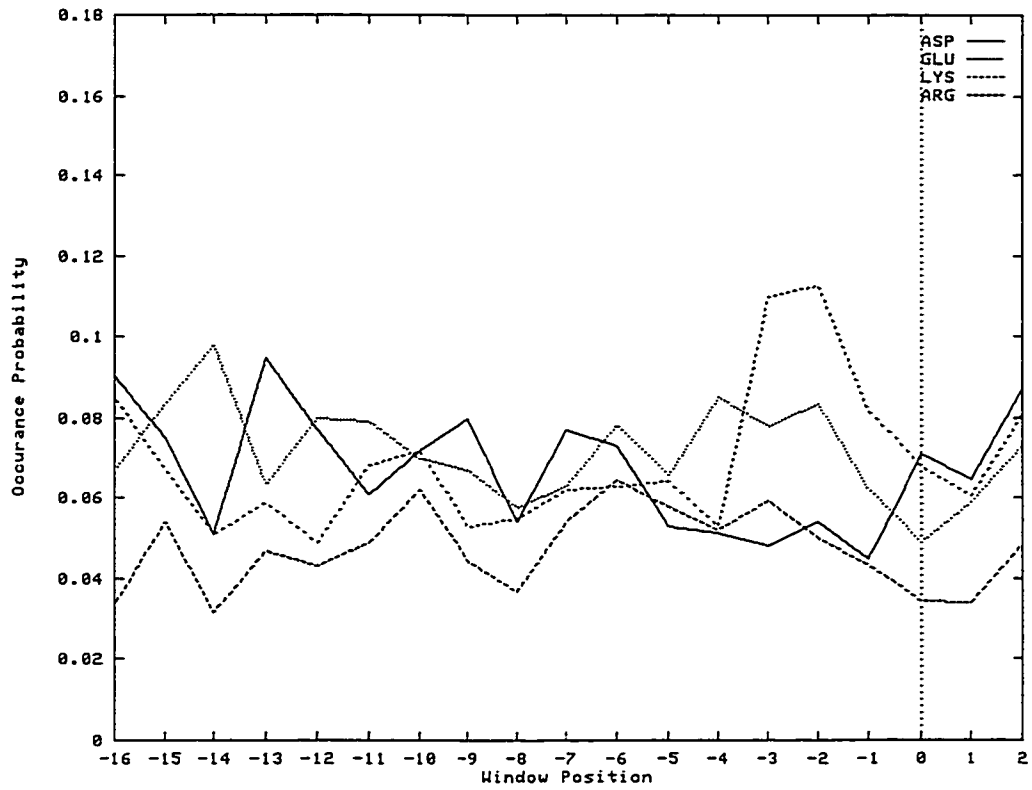
$2. P(w_k | N \text{ at } 0)$ 

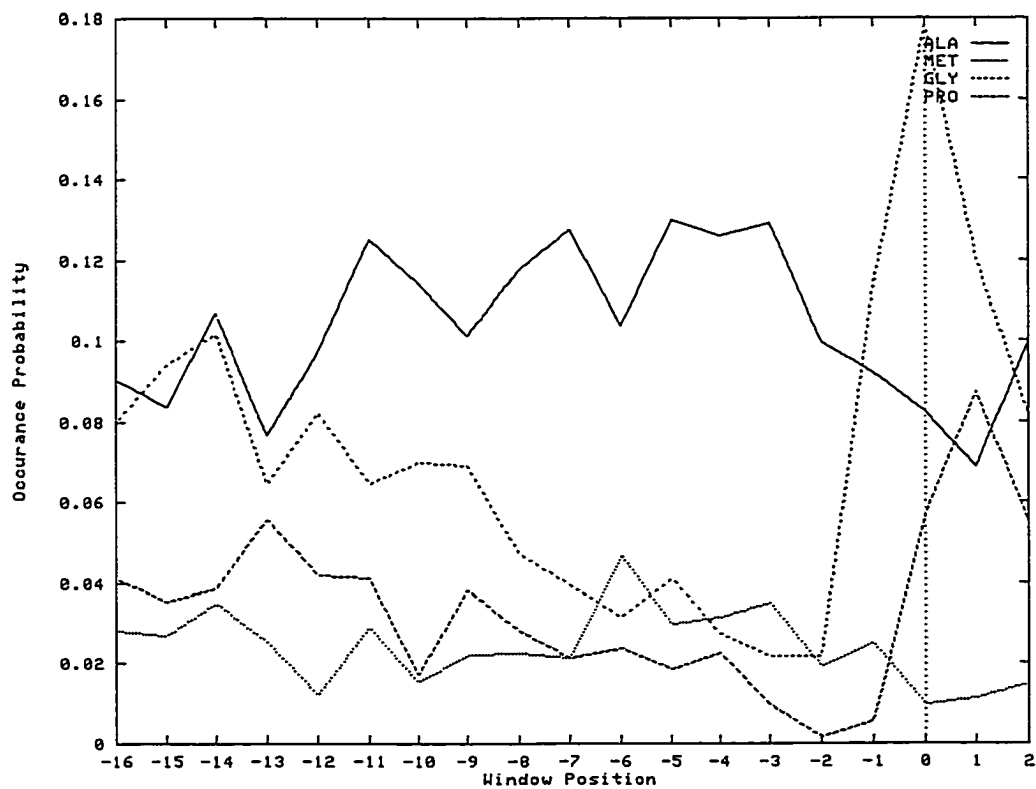
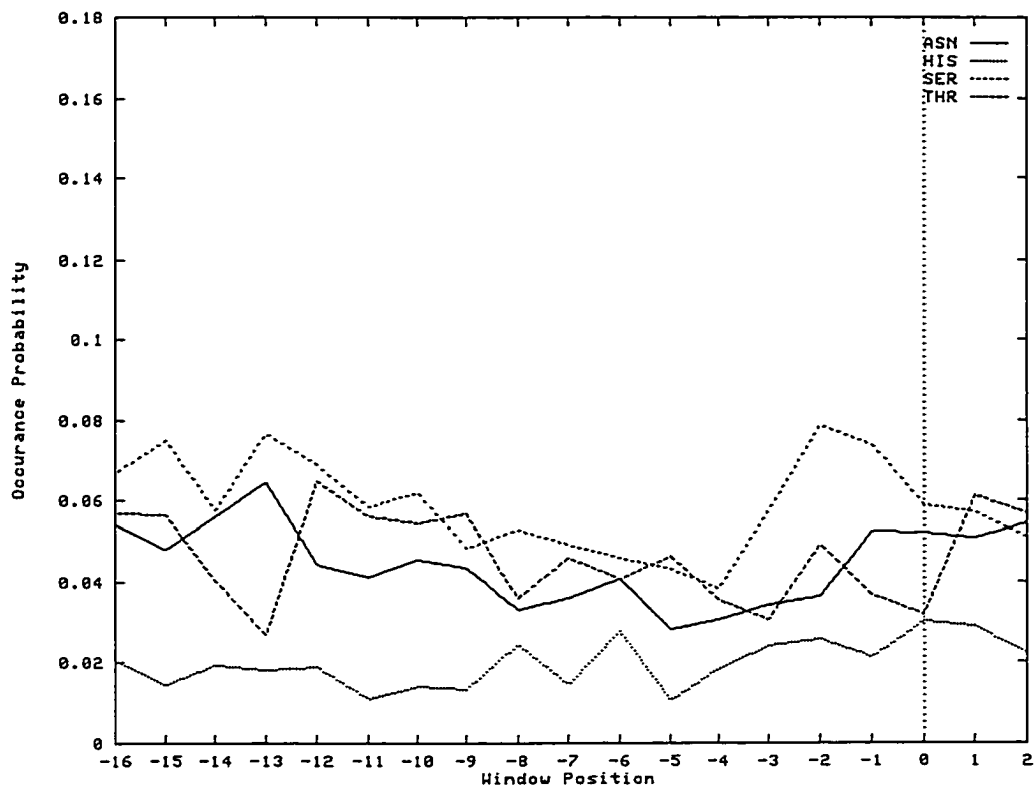


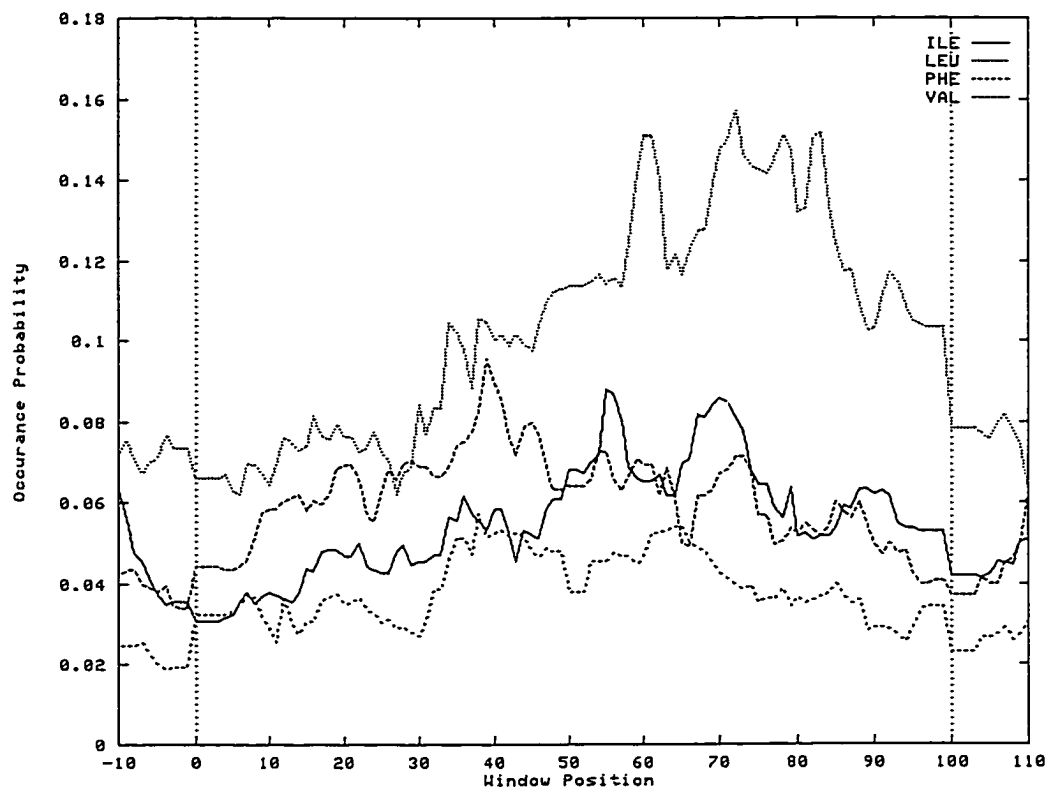


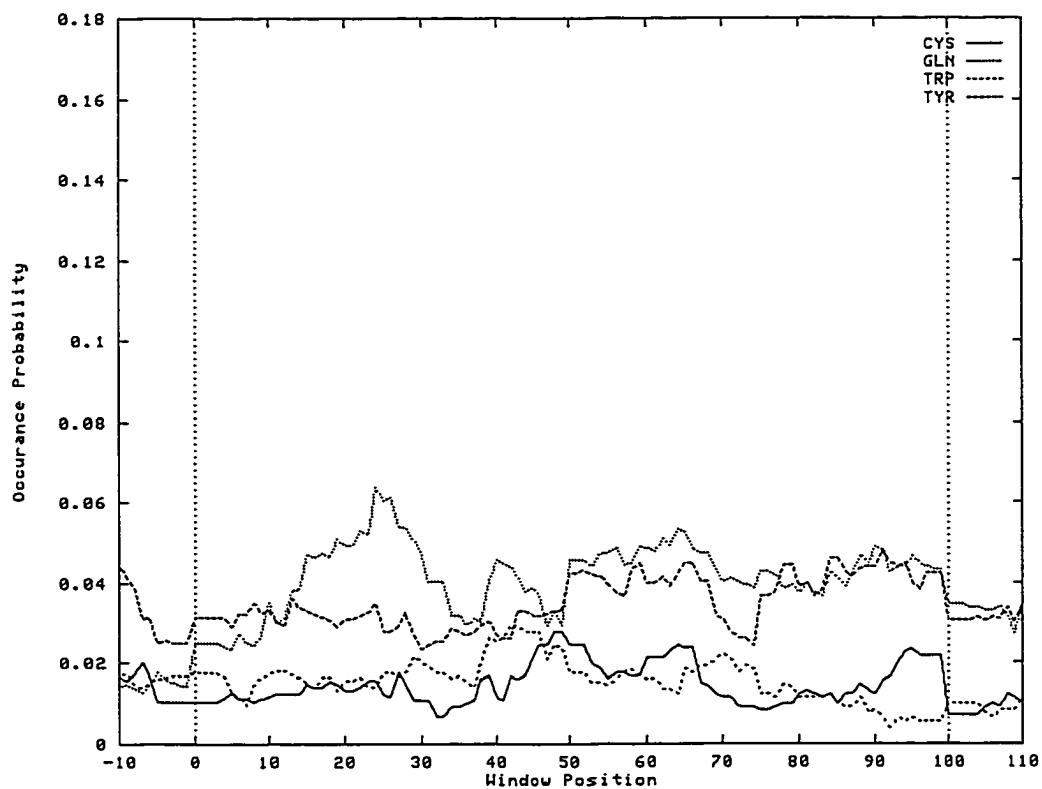
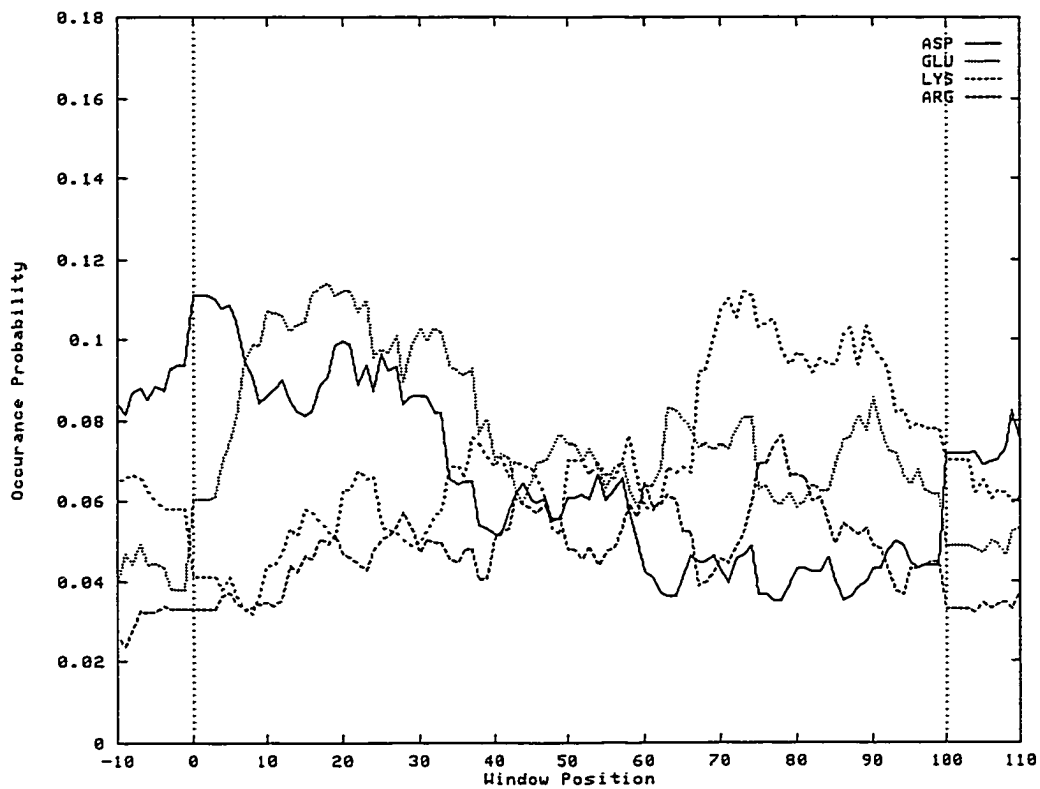
3.  $P(w_k | C \text{ at } 0)$ 

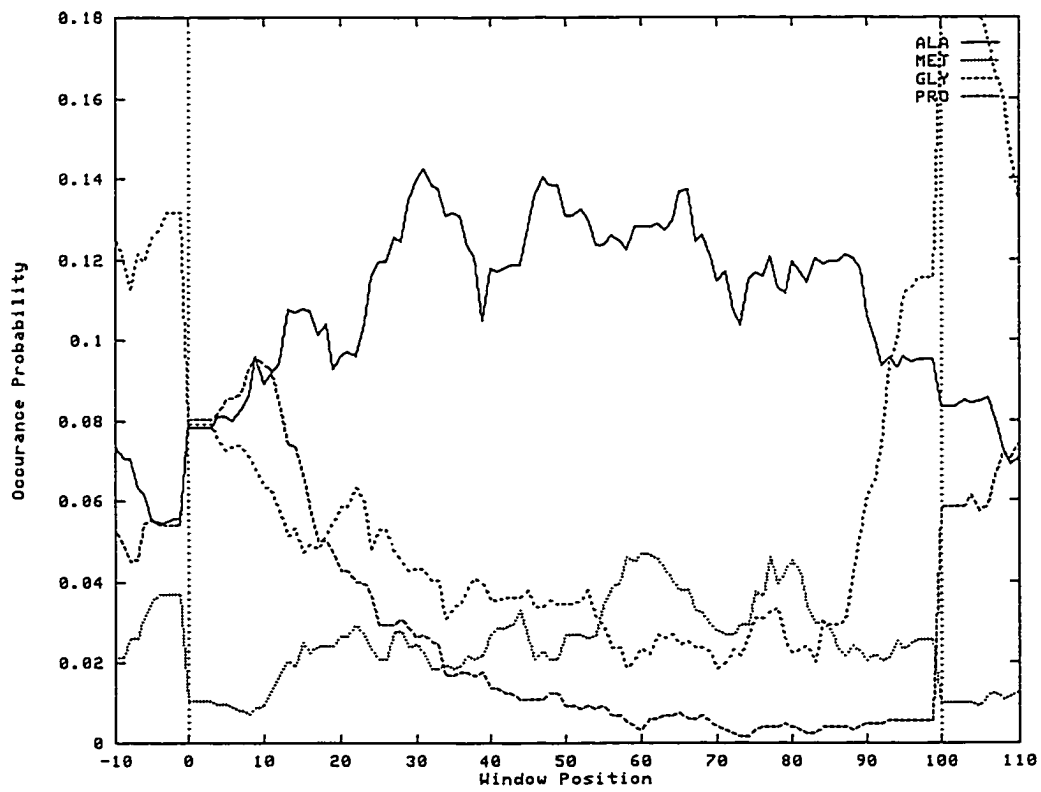
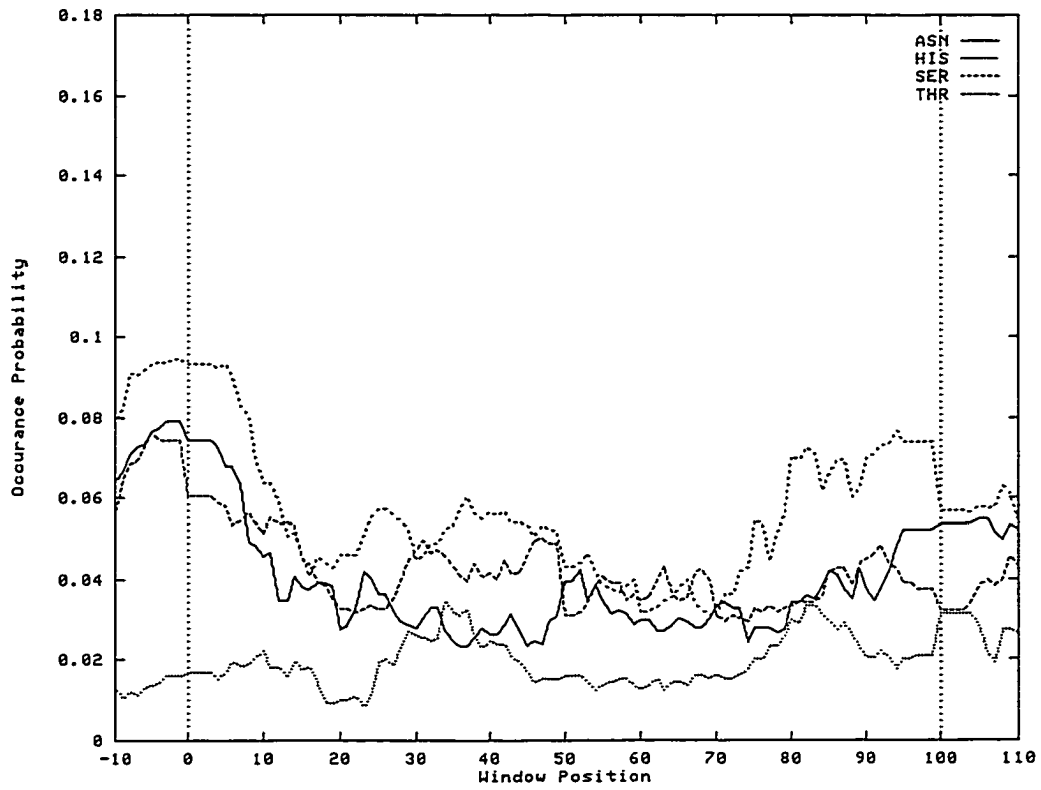






4.  $P(w_s | \text{scaled helix range } s \in [-10\%, +100\%])$ 





## VITA

### Neal Andrew Krawetz

654 Barnsley Way  
Sunnyvale, CA 94087

#### Education

*Texas A&M University*

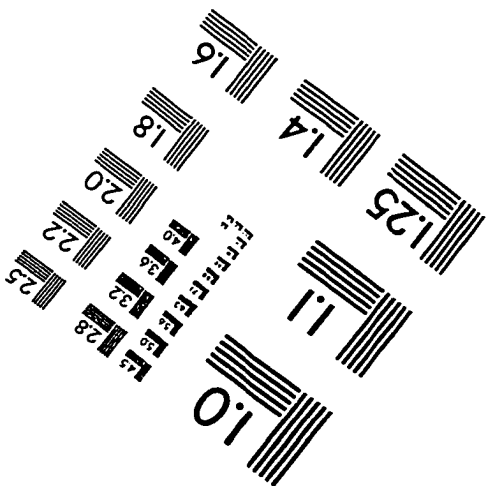
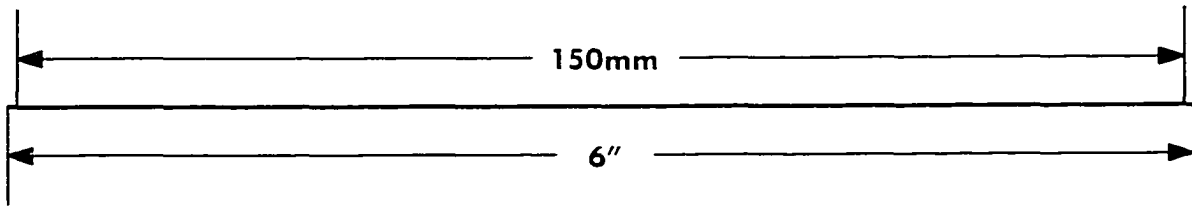
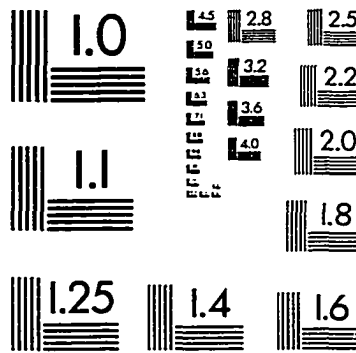
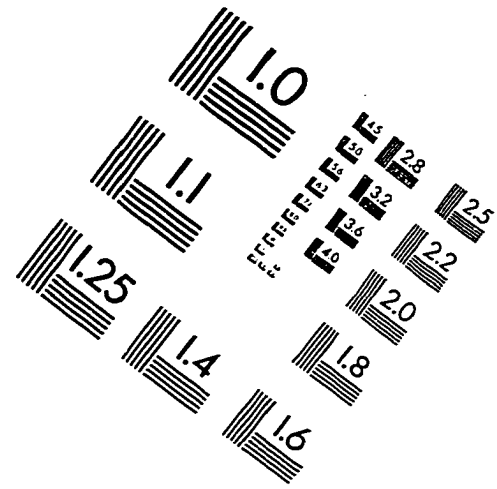
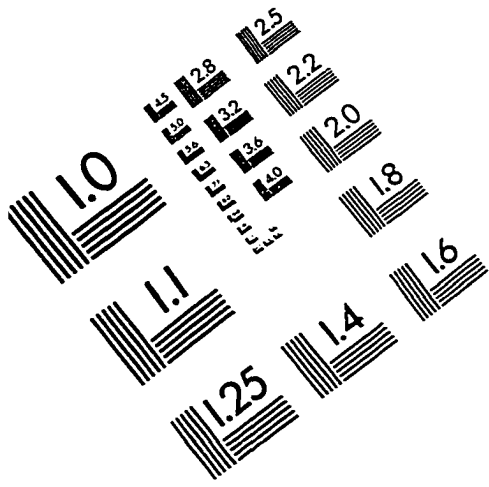
Doctorate of Philosophy Degree, Computer Science Graduate program.

Dissertation topic: *Investigation into Protein Folding Prediction of Helices using Techniques in Computer Science*. May 1998. Advisor: Dr. John Yen.

*University of California, Santa Cruz*

Bachelor of Arts Degree, Computer and Information Science Major (CIS), June 1992.

# IMAGE EVALUATION TEST TARGET (QA-3)



**APPLIED IMAGE, Inc**  
 1653 East Main Street  
 Rochester, NY 14609 USA  
 Phone: 716/482-0300  
 Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved

