



2015-05-01

Feature Identification and Reduction for Improved Generalization Accuracy in Secondary-Structure Prediction Using Temporal Context Inputs in Machine-Learning Models

Matthew Benjamin Seeley
Brigham Young University - Provo

Follow this and additional works at: <https://scholarsarchive.byu.edu/etd>



Part of the [Computer Sciences Commons](#)

BYU ScholarsArchive Citation

Seeley, Matthew Benjamin, "Feature Identification and Reduction for Improved Generalization Accuracy in Secondary-Structure Prediction Using Temporal Context Inputs in Machine-Learning Models" (2015). *All Theses and Dissertations*. 5267.
<https://scholarsarchive.byu.edu/etd/5267>

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

Feature Identification and Reduction for Improved Generalization
Accuracy in Secondary-Structure Prediction Using Temporal
Context Inputs in Machine-Learning Models

Matthew Benjamin Seeley

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of
Master of Science

Mark J. Clement, Chair
Tony R. Martinez
Bryan S. Morse

Department of Computer Science
Brigham Young University
May 2015

Copyright © 2015 Matthew Benjamin Seeley
All Rights Reserved

ABSTRACT

Feature Identification and Reduction for Improved Generalization Accuracy in Secondary-Structure Prediction Using Temporal Context Inputs in Machine-Learning Models

Matthew Benjamin Seeley
Department of Computer Science, BYU
Master of Science

A protein's properties are influenced by both its amino-acid sequence and its three-dimensional conformation. Ascertaining a protein's sequence is relatively easy using modern techniques, but determining its conformation requires much more expensive and time-consuming techniques. Consequently, it would be useful to identify a method that can accurately predict a protein's secondary-structure conformation using only the protein's sequence data. This problem is not trivial, however, because identical amino-acid subsequences in different contexts sometimes have disparate secondary structures, while highly dissimilar amino-acid subsequences sometimes have identical secondary structures. We propose (1) to develop a set of metrics that facilitates better comparisons between dissimilar subsequences and (2) to design a custom set of inputs for machine-learning models that can harness contextual dependence information between the secondary structures of successive amino acids in order to achieve better secondary-structure prediction accuracy.

Keywords: Bioinformatics, machine learning, secondary-structure prediction, amino-acid properties

ACKNOWLEDGMENTS

I wish it were possible to write a thorough, individualized, and proper expression of gratitude to every person who has contributed in some way to my education. However, I have had the privilege of learning from hundreds of friends, colleagues, teachers, and professors in both formal and informal settings throughout my life. I am indebted to all of them even though I will only list a few of them below. I would like to thank:

- Amber, my wife, for her patience, perseverance, and loyal support throughout graduate school;
- Steven Packard, my father-in-law, for his example of commitment to lifelong learning;
- Steve Perry, Alex Haymond, and Erik Ericksen for their professional mentoring;
- Nate Seeley, my brother, for always being willing to help me fix bugs in my programming projects;
- Grant Allen Seeley Jr. and Christine Seeley, my parents, for continuing to support me;
- Dee Anderson, my former work supervisor, for suggesting that I pursue a graduate degree in computer science;
- Kristin Gerdy and Jang Lee for helping me improve my writing;
- Dr. Clement, my advisor, for giving me the opportunity to do this research; and
- Dr. Martinez and Dr. Morse for serving on my graduate committee.

Table of Contents

List of Figures	viii
List of Tables	ix
1 Introduction	1
1.1 Initial Motivation and Objectives	1
1.2 Motivation for Investigating Features Based on Amino-Acid Properties . . .	3
1.3 Motivation for Investigating Contextual Features Comprising Predicted Output Classes of Neighboring Residues	5
1.4 A Note on Project Scope	8
1.5 Summary of Introduction and Thesis Statement	9
2 Related Work	10
2.1 Brief Explanation of the History of Secondary-Structure Prediction	10
2.1.1 The First Decade	10
2.1.2 The Hypothesis of Absolute Determinism	11
2.1.3 Free Energy, Levinthal’s Paradox, and Anfinsen’s Dogma	11
2.1.4 Early Statistical Models	12
2.1.5 Early Machine-learning Models and Multiple-sequence Alignment Information	13
2.1.6 The Modern Era	14
2.1.7 Use of Amino-Acid Properties in Secondary-Structure Prediction . . .	15

2.1.8	Use of Predicted Labels of Neighboring Amino Acids for Context in Secondary-Structure Prediction	15
2.2	Brief Summary of the Approaches used in this Project that have not been used Previously	16
3	Preliminary Experimental Result	18
3.1	Design for the Proof-of-Concept Experiment	18
3.2	Explanations Regarding Some Available Data Sets	20
3.2.1	The "Molecular Biology (Protein Secondary Structure) Data Set" [47]	20
3.2.2	The RS126 Data Set [14]	21
3.2.3	The CB396, CB251, and CB513 Data Sets [61]	21
3.2.4	The PSS504 Data Set [66]	22
3.2.5	The EVA6 Data Set [78]	22
3.2.6	The PLP399, PLP364, and PLP273 Data Sets [79]	23
3.2.7	The BT426 Data Set [80]	24
3.2.8	The BT823 and BT547 Data Sets [81]	24
3.2.9	The SPX Data Set [82]	24
3.2.10	The TT1032 Data Set [69]	25
3.3	Finding and Evaluating a Larger Set of Amino-Acid Properties	25
4	Primary Experimental Results	26
4.1	Evaluation of the Relevance of the Amino-Acid Properties	26
4.1.1	First Approach to Feature Selection	26
4.1.2	Second Approach to Feature Selection	27
4.1.3	Third Approach to Feature Selection	28
4.1.4	Conclusions Regarding the Use of Amino-Acid Properties for Secondary- Structure Prediction	31
4.2	Using Majority-Vote Ensembles to Raise Prediction Accuracy	34

4.2.1	First round of Majority-Vote Ensembles	35
4.2.1.1	Selecting a Set of Classifiers	35
4.2.1.2	Summary of Approaches used to Create and Verify Diversity	36
4.2.1.3	Results for First Round of Majority-Vote Ensembles	37
4.2.2	Second Round of Majority-Vote Ensembles	40
4.2.2.1	Selecting a Set of Classifiers Including More Cost-sensitive Models	40
4.2.2.2	Results for Second Round of Majority-Vote Ensembles	41
4.2.3	Conclusions and Possible Directions for Future Research Regarding Majority-Vote Ensembles that have Diversity Generated from the Three Approaches	44
4.3	Evaluating the use of Temporal Context Nodes	45
4.3.1	Relaxation	45
4.3.2	Collaborative Model Using Three High-Precision Classifiers	47
4.3.3	Conclusions Regarding Temporal Context Attributes and Directions for Future Research	49
5	Conclusion	51
5.1	Contributions to the Field of Study	51
5.2	Possible Directions for Future Work	53
	References	55
	Appendix A (Table A)	65
	Appendix B (Table B)	68
	Appendix C (Table C)	69
	Appendix D (Table D)	93

Appendix E (Table E)	99
Appendix F (Table F)	100
Appendix G (Table G)	105

List of Figures

1.1	Wenxiang Diagrams Comparing Helical and Non-Helical Regions	4
1.2	Distribution of Hydrophobic Inter-Moment Angles Found in α -Helices of Length 13aa in ss.txt	6
1.3	Distribution of Hydrophobic Inter-Moment Angles Found in Non-Structured Sequences of Length 13aa in ss.txt	6
1.4	Distribution of Lengths of All Contiguous α -Helix Sequences Found in ss.txt (measured in amino acids)	7

List of Tables

3.1	Prediction Accuracies of Several Algorithms on 3E6R Data (Ten-Fold Cross Validation)	19
3.2	List of Superseded Protein Structures and their Replacements for Data Set RS126	21
3.3	List of Superseded Protein Structures and their Replacements for Data Set CB396	22
3.4	List of Superseded Protein Structures and their Replacements for Data Set PSS504	22
3.5	List of Superseded Protein Structures and their Replacements for Data Set EVA6	23
3.6	List of Superseded Protein Structures and their Replacements for Data Set BT426	24
3.7	List of Superseded Protein Structures and their Replacements for Data Set BT823	24
3.8	List of Superseded Protein Structures and their Replacements for Data Set BT547	24
4.1	20 Attributes Selected Using the Second Approach to Feature Selection . . .	29
4.2	The 44 Attributes used in Each Individual-property arff File	32
4.3	Q_3 Accuracies Achieved by Machine-learning Models on Selected arff Files with Property Attributes	33

4.4	Exemplary Pairwise Yule’s Q Statistics for Combinations of Two Classifiers for the First Round of Majority-Vote Ensembles	38
4.5	Number of Times Each Classifier was used the First Round of Majority-Vote Ensembles	39
4.6	Exemplary Pairwise Yule’s Q Statistics for Combinations of Two Classifiers for the Second Round of Majority-Vote Ensembles	42
4.7	Number of Times Each Classifier was used the Second Round of Majority-Vote Ensembles	43
4.8	Q_3 Accuracies of Classifiers Using CB396 Training Set and RS126 Test Set with True Output Classes of 8 Neighboring Instances Used as Temporal Context Features	45
4.9	Q_3 Accuracies Achieved in Successive Iterations Using the Relaxation Process	46
4.10	Q_3 Accuracies Achieved Using Training Sets having Different Percentages Temporal Context Features Unknown	48
A.1	(Page 1 of 3) 78 Attributes Selected Using the Second Approach to Feature Selection	65
A.2	(Page 2 of 3) 78 Attributes Selected Using the Second Approach to Feature Selection	66
A.3	(Page 3 of 3) 78 Attributes Selected Using the Second Approach to Feature Selection	67
B.1	(Page 1 of 1) The 44 attributes used in Each Individual-property arff File . .	68
C.1	(Page 1 of 24) Q_3 accuracies achieved by various machine-learning models . .	69
C.2	(Page 2 of 24) Q_3 accuracies achieved by various machine-learning models . .	70
C.3	(Page 3 of 24) Q_3 accuracies achieved by various machine-learning models . .	71
C.4	(Page 4 of 24) Q_3 accuracies achieved by various machine-learning models . .	72
C.5	(Page 5 of 24) Q_3 accuracies achieved by various machine-learning models . .	73

C.6 (Page 6 of 24) Q_3 accuracies achieved by various machine-learning models . .	74
C.7 (Page 7 of 24) Q_3 accuracies achieved by various machine-learning models . .	75
C.8 (Page 8 of 24) Q_3 accuracies achieved by various machine-learning models . .	76
C.9 (Page 9 of 24) Q_3 accuracies achieved by various machine-learning models . .	77
C.10 (Page 10 of 24) Q_3 accuracies achieved by various machine-learning models .	78
C.11 (Page 11 of 24) Q_3 accuracies achieved by various machine-learning models .	79
C.12 (Page 12 of 24) Q_3 accuracies achieved by various machine-learning models .	80
C.13 (Page 13 of 24) Q_3 accuracies achieved by various machine-learning models .	81
C.14 (Page 14 of 24) Q_3 accuracies achieved by various machine-learning models .	82
C.15 (Page 15 of 24) Q_3 accuracies achieved by various machine-learning models .	83
C.16 (Page 16 of 24) Q_3 accuracies achieved by various machine-learning models .	84
C.17 (Page 17 of 24) Q_3 accuracies achieved by various machine-learning models .	85
C.18 (Page 18 of 24) Q_3 accuracies achieved by various machine-learning models .	86
C.19 (Page 19 of 24) Q_3 accuracies achieved by various machine-learning models .	87
C.20 (Page 20 of 24) Q_3 accuracies achieved by various machine-learning models .	88
C.21 (Page 21 of 24) Q_3 accuracies achieved by various machine-learning models .	89
C.22 (Page 22 of 24) Q_3 accuracies achieved by various machine-learning models .	90
C.23 (Page 23 of 24) Q_3 accuracies achieved by various machine-learning models .	91
C.24 (Page 24 of 24) Q_3 accuracies achieved by various machine-learning models .	92
D.1 (Page 1 of 6) Pairwise Yule's Q statistics for the first round of majority-vote ensembles	93
D.2 (Page 2 of 6) Pairwise Yule's Q statistics for the first round of majority-vote ensembles	94
D.3 (Page 3 of 6) Pairwise Yule's Q statistics for the first round of majority-vote ensembles	95
D.4 (Page 4 of 6) Pairwise Yule's Q statistics for the first round of majority-vote ensembles	96

D.5 (Page 5 of 6) Pairwise Yule’s Q statistics for the first round of majority-vote ensembles	97
D.6 (Page 6 of 6) Pairwise Yule’s Q statistics for the first round of majority-vote ensembles	98
E.1 Number of Times Each Classifier was used the First Round of Majority-vote Ensembles	99
F.1 (Page 1 of 5) Pairwise Yule’s Q statistics for the second round of majority-vote ensembles	100
F.2 (Page 2 of 5) Pairwise Yule’s Q statistics for the second round of majority-vote ensembles	101
F.3 (Page 3 of 5) Pairwise Yule’s Q statistics for the second round of majority-vote ensembles	102
F.4 (Page 4 of 5) Pairwise Yule’s Q statistics for the second round of majority-vote ensembles	103
F.5 (Page 5 of 5) Pairwise Yule’s Q statistics for the second round of majority-vote ensembles	104
G.1 Number of Times Each Classifier was used the Second Round of Majority-vote Ensembles	105

Chapter 1

Introduction

1.1 Initial Motivation and Objectives

Accurate protein secondary-structure prediction from amino-acid sequence data has been called the holy grail of structural bioinformatics [1]. This is due, in part, to the fact that sequence data can be extracted using relatively fast and inexpensive laboratory techniques such as Edman sequencing, while protein structural data typically has to be extracted using much more expensive techniques such as x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. To illustrate the price difference, consider that the average cost of determining a novel protein's three-dimensional *structure* was about \$138,000 (though the best lab averaged \$67,000 per protein) in 2006 [2]. A protein *sequence*, by contrast, can now be determined for just over \$100 [3]. A reliable method for predicting secondary structure from sequence data could, therefore, help researchers model a sequenced protein's three-dimensional structure quickly and inexpensively.

Many of the most effective modern algorithms for secondary-structure prediction use information from multiple-sequence alignments of homologous proteins with known structures. This is undoubtedly a sound approach for predicting structures of sequences that have many known homologues; good accuracy could probably be achieved by simply predicting that the test sequence's structural label at any given position in the sequence matches the consensus label at the corresponding position in the multiple-sequence alignment. However, "a significant number of proteins identified in genome sequencing projects have no detectable sequence similarity to any known protein" [4]. For these proteins with few or no known homologues,

it would be prudent to use a different approach—one that still leverages domain-specific knowledge in the context of a machine-learning model.

Ideally, predicting the secondary structure of a protein at a given amino-acid position would be as simple as identifying unique, short subsequences whose central amino acids always have one specific label. This approach’s effectiveness is limited, though, for two reasons: first, the number of possible permutations of 22 amino acids (with replacement) for a subsequence is exponentially large. There would, for example, be 22^{13} possible subsequences that are 13 amino acids long (i.e., of length 13aa) . Even with all the data in the Research Collaboratory for Structural Bioinformatics protein data bank (RCSB PDB) [5], the number of subsequences of length 13aa with known labels is a very small fraction of the number of subsequences that is possible. More important, though, is the fact that the RCSB data demonstrates that many identical subsequences of length 13aa have different labels when they appear in different proteins or in different contexts. One study has even demonstrated that a specific sequence of eleven amino acids folds into an alpha helix when inserted into one position of a protein, but folds into a beta sheet when inserted into a different position in the same protein [6]. Thus, even if the search space of every possible subsequence were tractable, some subsequences could only be assigned tentative majority labels; this would limit the maximum theoretical accuracy of a predictive model.

In order for a machine-learning model to generalize well to test instances that have little sequence identity with training instances, it must use some intelligent metric that can tell when dissimilar subsequences have amino acids with similar properties at identical positions. It must also be able to identify similar periodic patterns in those properties so that instances that are nearly identical, but whose attribute values are all shifted by a single position (like two successive sliding windows) can still be recognized as similar to one another. Furthermore, the model should have some means for incorporating contextual information about the predicted structures of preceding and succeeding amino acids in the protein. This

should enable the model to resolve the ambiguity that occurs when identical subsequences exhibit different structures.

For this thesis, the first aim was to investigate and refine a set of metrics that can measure the similarity between amino-acid subsequences based on quantitative properties rather than on sequence identity alone and to use these metrics to develop a custom set of input features for machine-learning models in order to improve protein secondary-structure prediction accuracy. The second aim was to develop a customized set of forward- and backward-context attributes to leverage context information in order to predict when identical subsequences will have different structures. Since the ultimate intention was for these context attributes to comprise the predicted output classes of an instant subsequence’s immediate neighbors in a sliding-window scheme, using an iterative relaxation process in order to maximize prediction accuracy was included in this second aim.

1.2 Motivation for Investigating Features Based on Amino-Acid Properties

Measuring how similar two amino acids are to one another is deceptively difficult because there are hundreds of known properties [7] that can be compared; some may be similar to each other with regard to one property, but dissimilar to each other with regard to another property. While it is likely that many of these properties would not yield useful information for secondary structure prediction, it is difficult to define each property’s relevance *a priori*. Consequently, we planned to evaluate the relevance of each of these properties individually, if possible.

In addition to a metric that measures similarity between amino acids situated at identical positions, there should also be some metric that captures similarity between sequences. This would be useful because the test instances and training instances used by many secondary-structure prediction approaches consist of sliding windows applied across the linear sequence of amino acids that makes up a given protein’s primary structure. In these approaches, each input feature of a given instance is the single-letter representing an amino acid at a given

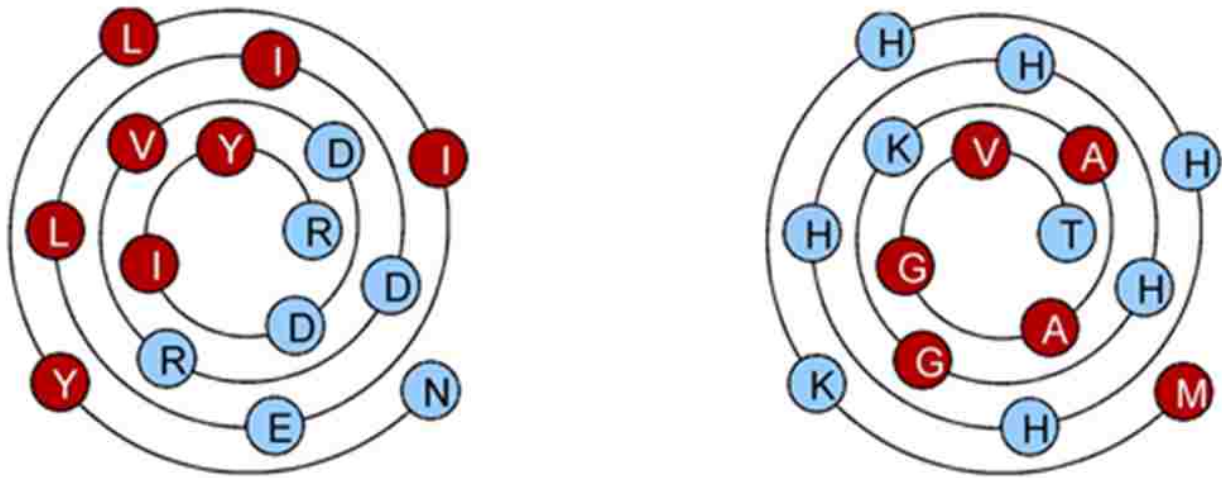


Figure 1.1: These are Wenxiang diagrams of a true alpha helix (left) and a region without secondary structure mapped as though it was an alpha helix (right). Hydrophobic residues are colored red.

position in the sliding window. To illustrate how this could be problematic, consider a sliding window of size k applied to a protein of total length n . Each instance would have k input features, so there would be a total of $n-k+1$ instances derived from the protein. Any two consecutive instances would be very similar because they would share a subsequence of length $k-1$. The values for the input features, however, would all be shifted over by one. Hence, a classifier that is only configured to compare input-feature values at identical positions would have no way of knowing that the two consecutive instances should actually be considered very similar.

Hydrophobic moment is an example of such a metric [8]. It is generally known that the interaction of amino-acid residues with water strongly influences the native structure of proteins [9].

Amphiphilic helices are often situated in proteins such that one side of the helix interacts with the hydrophobic interior of the protein and the other side interacts with the hydrophilic surrounding solution. As a result, hydrophobic and hydrophilic residues are generally distributed in a non-random pattern that isolates them on opposite sides of the

helix. The angle at which one residue is radially pointed outward from the center of an alpha helix is approximately 100 degrees greater than the angle of the previous residue in the helix. This principle is best illustrated with a Wenxiang diagram [9], a "conical projection of an α -helix onto a plane perpendicular to its axis" as shown in figure 1.1 [10].

Some researchers have quantified this property by using the hydrophobic moment [9]. To calculate the hydrophobic moment, a descriptive vector is created for each amino acid. The direction of the vector points outward from the center of the helical axis toward the residue, while the magnitude of the vector equals the hydrophobic magnitude of the residue (which is, of course, negative for hydrophilic residues). The hydrophobic moment of a sequence of amino acids is calculated by adding all the individual residue vectors. It has been shown to be a helpful metric for secondary-structure prediction [8].

In order to glean more information from the hydrophobicity patterns, though, a slightly modified approach was also used for this project. The cumulative moments of the hydrophilic and hydrophobic residues can be calculated separately and the angle between them can be determined. The inter-moment angle is a metric we invented independently and have not seen used in any of the literature, but it looked promising because data gleaned from `ss.txt` demonstrates that the distributions of inter-moment angles for alpha-helical regions and unstructured regions appear to be very distinguishable; that data is shown in the histograms found in figures 1.2 and 1.3.

1.3 Motivation for Investigating Contextual Features Comprising Predicted Output Classes of Neighboring Residues

When aiming to identify the secondary-structure label of any single amino acid in a sequence, it is important to remember that there is a high degree of dependence between its label and the labels of the amino acids immediately next to it. An amino acid that is part of an alpha helix, for example, is *always* next to at least one other amino acid that also has the same label because at least four consecutive amino acids are needed to form an alpha helix

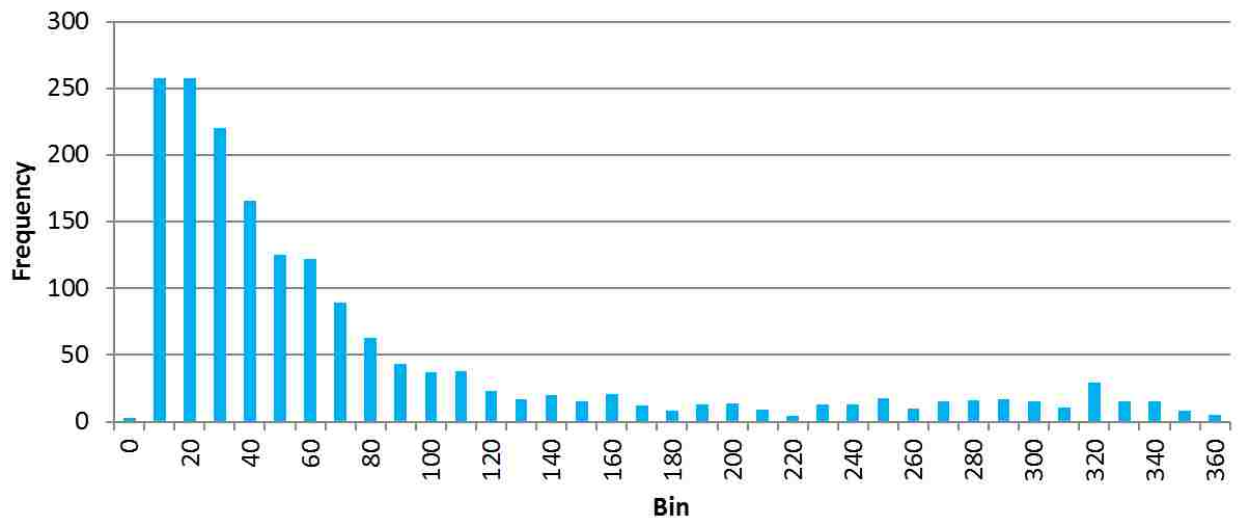


Figure 1.2: Distribution of Hydrophobic Inter-Moment Angles Found in α -Helices of Length 13aa in ss.txt

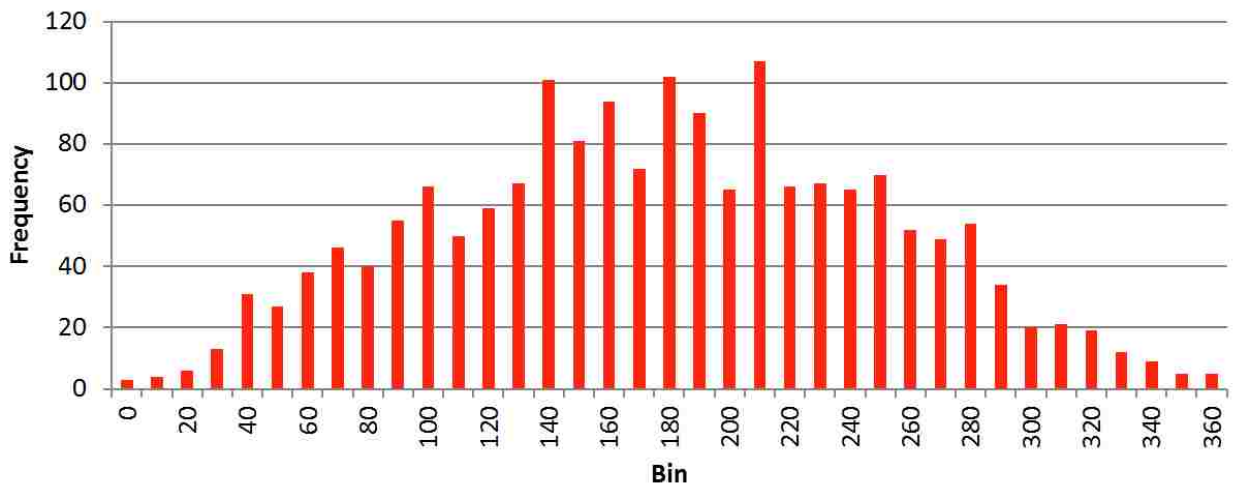


Figure 1.3: Distribution of Hydrophobic Inter-Moment Angles Found in Non-Structured Sequences of Length 13aa in ss.txt

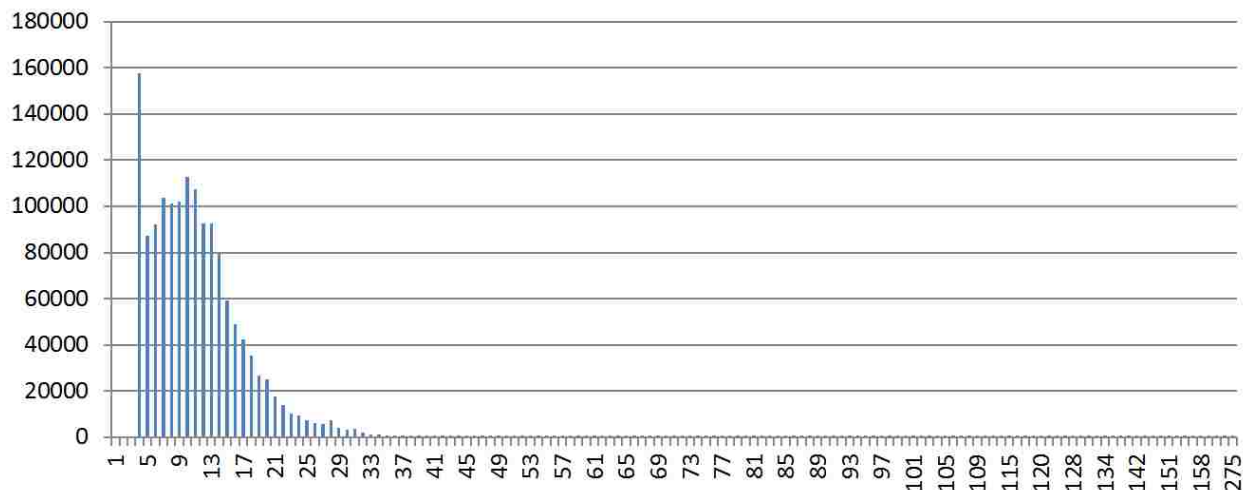


Figure 1.4: Distribution of Lengths of All Contiguous α -Helix Sequences Found in ss.txt (measured in amino acids)

structure [11]. The principle is also relevant to other types of secondary structures, since they are all formed as a result of bonding between the backbones of at least two amino acids. The data found in ss.txt, a file containing the known secondary structure labels for all RCSB Protein Data Bank files, is consistent with this principle. As an example, the distribution of lengths of all contiguous alpha helix structures found in ss.txt is shown in figure 1.4.

Given this high degree of dependence between the labels of successive amino acids, the best machine-learning models for secondary-structure prediction should incorporate some means for capturing the dependence information that is found in a given training set. Complex dependencies that cannot be captured by simply looking back one step clearly exist in this project's data set. For example, if the label N signifies no secondary structure and the label H signifies an alpha-helix structure, four amino acids with the labels NHHH *must* be followed by an amino acid with label H, but one amino acid with the label H may or may not be followed by an amino acid with the label H. As a result, it would be preferable to use predicted labels from at least four preceding instances as temporal backward-context attributes for a current instance.

1.4 A Note on Project Scope

DSSP is a database of secondary structure assignments for all entries in the RCSB Protein Data Bank [12]. DSSP also refers to the program that extrapolates secondary-structure assignment based on the three-dimensional coordinates available to a given protein in the RCSB Protein Data Bank [13]. Kabsch and Sander’s Dictionary of Secondary Structures of Proteins (the unabbreviated form of the acronym DSSP) defines eight possible secondary-structure labels: α -helix (H), residue in isolated beta bridge (B), extended beta strand (β -sheet) (E), 3_{10} helix (G), π -helix (I), hydrogen-bonded turn (T), bend (S), and none (). For the purposes of evaluating programs that predict secondary structure, however, Rost and Sander outlined the following convention: the three different types of helices are grouped together into once class (H), the extended beta strand remains a stand-alone class, and the remaining structures (including "none") are grouped together into the loop class (L, though we will call it N) [14]. Qian and Sejnowski also provided a concise explanation of a common metric used for measuring model performance on secondary-structure prediction:

$$Q_3 = \frac{P_\alpha + P_\beta + P_{coil}}{N} \quad (1.1)$$

where N is the total number of residues whose structures were predicted and P_α , P_β , and P_{coil} are the number of residues with each respective type of secondary structure that were predicted correctly [47].

For this project, we chose to evaluate our final methods using the three-class convention because some classes from the eight-class definitions are extremely rare [13]. Furthermore, most published studies on secondary-structure prediction have used this convention, so it will be easier to compare our results to those of other researchers if we use it.

As we mention at various points throughout this thesis, the current models that achieve the highest secondary-structure prediction accuracy are those that use information from multiple-sequence alignments. Aydin [4] refers to models that do not use information from

homologous proteins as single-sequence algorithms. Note that the *single-sequence* concept is more stringent than the *sequence-unique* concept (as used in CASP). The latter only requires that there be no significant similarities between proteins in the test set and proteins in the training set [4]. Unlike the single-sequence condition, however, it still allows sequence profile information to be used; this improves prediction accuracy by several percentage points [4]. The best current single-sequence approach, though, achieves a prediction accuracy below 70% [15].

We chose to focus on a single-sequence algorithm for several reasons: (1) multiple-sequence alignments are computationally expensive; (2) the accuracy of any algorithm applied in conjunction with a multiple-sequence alignment might be more dependent on the degree of homology between the aligned sequences and the test sequences than on the merits of the algorithm itself; (3) there are many proteins with no known homologues [4]; and (4) our method can be used in conjunction with methods that use multiple-sequence alignments in the future if we so desire.

1.5 Summary of Introduction and Thesis Statement

There is demonstrable evidence that information that can be gleaned from amino-acid properties and from predicted labels of neighboring amino-acids may help identify patterns that may ultimately prove useful for improving protein secondary-structure prediction, particularly for proteins that lack known homologues. In this project, a set of input features based on amino-acid properties is developed and shown to aid several machine-learning classifiers in achieving better Q_3 secondary-structure prediction accuracy under conditions where close homologues are not used in the training set. In addition, a set of inputs that harnesses contextual dependence information between the secondary structures of successive amino acids is also shown to aid a few machine-learning classifiers in achieving better Q_3 secondary-structure prediction accuracy in some limited circumstances.

Chapter 2

Related Work

Researchers have been focused on identifying and predicting protein secondary structure for over five decades. The following sections provide some concise chronological summaries of the development of the theory and the approaches that researchers have used for secondary-structure prediction. The focus of the last two subsections narrows in to describe approaches that have used some form of information derived from amino-acid properties and approaches that have used predicted labels of neighboring amino acids to provide context information. Since this project focuses on secondary-structure prediction for proteins that lack known homologues, the approaches that do not require homology information will ultimately provide the best apples-to-apples benchmark to which our approach can be compared.

2.1 Brief Explanation of the History of Secondary-Structure Prediction

2.1.1 The First Decade

In 1951, researchers first described the patterns we call secondary structure in proteins [16, 17]. A few years later (1954), researchers identified proline as an amino acid that strongly affected secondary-structure patterns [18, 19]. In that same decade (1958), x-ray analysis of proteins progressed to the point where it was finally possible to generate complete three-dimensional models of proteins [20].

2.1.2 The Hypothesis of Absolute Determinism

In 1964, Straub published a thorough article describing the "widely accepted hypothesis" that secondary and tertiary structure could be determined entirely based on sequence data [21]. The theory seemed attractive enough, especially given that some previous and subsequent studies demonstrated that many unfolded proteins can refold into their original conformations when placed in the proper environments [22, 23]. However, Straub wisely noted that some observations were "not in harmony with the theory of absolute determinism," thereby showing awareness of the problem's greater complexity [21].

Early methods for secondary-structure prediction continued to develop; in general, they were simple rule-based models based on statistical correlations between the presence of certain amino acids and helices. In 1965, Guzzo suggested that certain amino acids—specifically proline, aspartic acid, glutamic acid, and histidine—were needed for an alpha helix to form [24]. The following year, Prothero extended Guzzo's work by proposing that any region of five amino-acid residues would be helical if at least three of the five were Ala, Val, Leu, or Glu and that any region of seven amino-acid residues would be helical if at least three residues were Ala, Val, Leu, or Glu and at least one was Ile, Thr, or Gln [25]. Periti [26] and Ptitsyn [27] also used statistical analyses to generate simple predictive models. In addition, Schiffer (1967) observed that helical-wheel representations of α -helices in proteins like myoglobin and hemoglobin demonstrated distinctive hydrophobic arcs that could help distinguish helical regions from non-helical ones; this was perhaps one of the earliest examples of how a periodic pattern of an amino-acid property could aid in secondary-structure prediction [28].

2.1.3 Free Energy, Levinthal's Paradox, and Anfinsen's Dogma

By 1969, some had theorized that a protein would simply fold into the conformation corresponding to its globally lowest free energy; Cyrus Levinthal, however, presented the now-famous "Levinthal's paradox" in which he argued that a protein could not randomly move through all of its possible conformations quickly enough to find its global minimum in

time to explain the experimental refolding speeds of some proteins [29]. Several years later (1973), Anfinsen presented a postulate that is now known as "Anfinsen's Dogma": given a specific set of environmental conditions, a small globular protein's native conformation will be a unique, stable, and kinetically accessible structure—though it may only represent a *local* minimum relative to free energy [30]. Simon also published useful research detailing some structural features that contribute to refolding ability [23]. (As a side note, it is now known that there are exceptions to Anfinsen's dogma, such as intrinsically disordered proteins [31]).

2.1.4 Early Statistical Models

In the meantime, models for secondary-structure prediction continued to develop. In 1971, Robson and Pain used an information-theory approach to harness some known statistical information about single residues and pairwise residue combinations into a simple predictive model [32]. That same year, the Protein Data Bank was officially established [33]. Nagano [34], Garnier [35], and Chou & Fasman [36] all developed methods that harnessed correlations between amino acids and secondary structure. Lim [37] and Ptitsyn [38] also began considering the influence of physico-chemical properties on secondary structure. In 1983, Kabsch & Sander compared the methods of Chou & Fasman, Lim, and Garnier, respectively, and tested them with newly available data; they ultimately concluded that the best overall three-state prediction accuracy that these methods could consistently achieve was about 56% [39]. Cohen [40] developed a model that considered hydrophilicity (the inverse of hydrophobicity) spacing patterns. Since these models were not designed to use information from multiple-sequence alignments or other information that is dependent on homology, they can appropriately be compared to the models developed in this project.

2.1.5 Early Machine-learning Models and Multiple-sequence Alignment Information

As early as 1978, it had been suggested that information from multiple-sequence alignments would improve the results of secondary-structure prediction [35]. A number of different researchers aimed to harness this information throughout the 1970s and 1980s [41, 42, 43, 44, 45].

In 1988, both Bohr [46] and Qian & Sejnowski [47] applied neural-network approaches to the secondary-structure prediction problem [47]. The latter selected a set of 106 proteins with known structures, taking care to limit the number of sequences that were "almost identical" because their results were "highly sensitive to homologies between the testing and training sets" [47]. Each data instance was derived from a sliding window of 13 amino-acid residues; the amino-acid identities of the 13 residues comprised the input features, while the three-class secondary structure label for one of the amino acids in the window comprised the output class [47]. They also provided a concise explanation of the Q_3 metric used for measuring model performance on secondary-structure prediction (shown in equation 1.1)[47]. Qian and Sejnowski's method ultimately achieved a Q_3 prediction accuracy of 64.3%; they suggested that a theoretical limit of about 70% could be achieved using local methods [47]. Since Qian and Sejnowski's did not use homology information and were careful ensure there was minimal homology between proteins in the training set and the test set, their results probably provide the best apples-to-apples benchmark for the methods used in this project.

Other researchers quickly followed suit by applying neural networks to secondary-structure prediction [48, 49, 50, 51, 52]. In 1993, Rost and Sander were able to achieve 70.8% accuracy by using a neural-network approach that added information from multiple-sequence alignments; in the process, they compiled the data set of that is now commonly known as RS126 [14]. However, since Rost and Sander used multiple-sequence alignments (and therefore homology information), their results would not serve as a good benchmark for the results

achieved by models that do not use homology information (such as the ones developed in this project).

2.1.6 The Modern Era

In the past 20 years, researchers have continued to apply neural networks and other machine-learning methods to the secondary-structure prediction problem. Some additional models that have been used include support-vector machines [53, 54], recurrent neural networks [55], decision trees [56], Bayesian networks [57], nearest-neighbor algorithms [58, 59], and hidden Markov models [60]. In general, the methods that achieve the highest prediction accuracies use information from multiple-sequence alignments and position-specific scoring matrices [61]. Others have also shown that a protein’s family classification, which is another type of homology information, can also be used to increase prediction accuracy [15]. Berezovsky and Trifonov also presented evidence that that proteins fold into subunits of 25–30 amino acids in a local way [62]. One recent method that strategically used homology information even reported achieving prediction accuracy exceeding 90% [63].

There are many different methods available for secondary-structure prediction, but Pirovano and Heringa suggest that SSpro is ”among the leading secondary structure prediction algorithms in terms of accuracy” [64]. In addition, they identify Porter as the ”current top performer” out of all algorithms currently registered on the EVA (Evaluation of Automatic protein structure prediction) server—a web-based assessment tool for evaluating the accuracy of secondary-structure prediction methods [64]. They also mention that PSIPRED is relatively accurate, easy to use, and popular [64]. However, because SSpro, Porter, PSIPRED, and even Rost and Sander’s model all heavily rely on the use of homology information, they do not conform to the sequence-unique approach and are therefore not ideal models to which our sequence-unique model can be compared. As a result, the best models for apples-to-apples comparison include Qian and Sejnowski’s model and the single-sequence leaders described by Aydin.

2.1.7 Use of Amino-Acid Properties in Secondary-Structure Prediction

There are some key amino-acid properties that have been shown to aid in secondary-structure prediction in the absence of information from homologous proteins. The properties that are generally recognized as being most relevant include residue conformational propensities [9, 65, 8], hydrophobic moments [9, 8], sequence edge effects [8], and residue ratios [8]. Grantham polarity scales [66], molecular weight [52], pseudo amino acid composition [54], and pair-coupled amino acid composition [88] have also been used by different researchers to aid in secondary-structure prediction.

Amino-acid properties have also been frequently used for classifying proteins into families. Cai, for example, used properties such as hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, surface tension, and solvent accessibility to classify proteins into families [67]. Others have used different sets of properties to classify proteins into families [68]. Family classifications, in turn, have been shown to be helpful for secondary-structure prediction [15].

2.1.8 Use of Predicted Labels of Neighboring Amino Acids for Context in Secondary-Structure Prediction

A number of researchers have aimed to consider, in one form or another, the predicted labels of neighboring amino acids as context to aid in secondary-structure prediction. Petersen, for example, used a sliding window of 17 residues as input to a neural network that predicted the label of the middle amino acid and its immediate neighbors simultaneously such that the prediction for the central amino acid at position i was dependent on the predictions for the amino acids at positions $i + 1$ and $i - 1$ [69]. Lundegaard used a similar sliding-window approach that also predicted the labels of three consecutive amino acids simultaneously [70]. While both used a balloting process, there was no relaxation step after the balloting.

Nyugen and Rajapakse used two-stage multi-class support vector machines wherein the outputs of the first-stage SVM were used as inputs for the second-stage SVM in order to

leverage contextual information—like the fact that beta strands consist of at three consecutive residues and alpha helices consist of at least four [71]. Once the second-stage SVM reached its final predictions, though, there was no relaxation step.

Baldi and Pollastri used a bi-directional recurrent neural network wherein outputs from hidden layers on preceding and succeeding sliding windows serve as inputs to the output layer of an instant sliding window [72, 73]. Their approach also uses information from multiple sequence alignments and has ultimately been implemented in two of the most successful secondary-structure-prediction programs to date: SSPro and Porter [73].

Asai used a hidden Markov model that iteratively re-estimated parameters (e.g., transition probabilities) [60]; this might be considered a form of relaxation.

2.2 Brief Summary of the Approaches used in this Project that have not been used Previously

While a small number amino-acid properties have been applied in one way or another to secondary-structure prediction, our experiments in this project test the usefulness of over 500 different amino-acid properties in single-sequence secondary-structure prediction. In order to test these properties, we use some known attributes, such as the total hydrophobic moment and the letters for individual amino acids in a sliding window. We also devise several novel attributes that can be derived using a given amino-acid property, such as the inter-moment angle and a series of attributes that represents property moments across several sub-windows of the sliding window in order to capture information about how the moment is changing within the instance represented by the sliding window. We also demonstrate that helpful diversity can be created for a classifier set used to generate majority-vote ensembles with improved overall prediction accuracy for secondary-structure prediction by using three different approaches to create classifiers: (1) creating different types of classifiers on the using the same attributes sets, (2) creating classifiers using different attribute sets derived

from amino-acid different properties, and (3) creating cost-sensitive versions of classifiers (an approach that has not been used in in this way).

In addition, we also test the usefulness of attributes that represent both true and predicted output classes of neighboring instances. We also apply a multiple-round relaxation process in using the predicted output classes in order to test whether relaxation can be used to increase prediction accuracy.

Chapter 3

Preliminary Experimental Result

3.1 Design for the Proof-of-Concept Experiment

While we have provided some theoretical justifications for the proposal in the previous sections, we also saw the need to run a preliminary experiment for proof-of-concept purposes. This section describes this experiment.

Venkatarajan used multidimensional scaling to condense the information from 237 amino-acid properties into five quantitative descriptors [7]. It seemed prudent to use these descriptors for the preliminary proof-of-concept experiment, since they contained a great deal of information that we hoped might help a machine-learning model quantify amino-acid similarity.

In the first step of our proof-of-concept experiment, the protein-data-bank (PDB) file for ferritin from the pseudo-nitzschia series was chosen as the data set because ferritin is a large protein with intricate secondary-structure patterns. The PDB file was converted to an arff file using a Perl script; the resulting data set had thirteen attribute columns and one classification column. For every given instance, each of the attribute columns could have any single-letter value found in the set $\{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X\}$, where each letter represented its corresponding amino acid (or, in the case of X, an unknown amino acid; X values are occasionally found in PDB files). The classification column of each instance could have any single-letter value found in the set $\{H,B,E,G,I,T,S,N\}$, where H = helix, B = residue in isolated beta bridge, E = extended beta strand, G = 3_{10} helix, I = π -helix, T = hydrogen-bonded turn, S = bend, and N = nothing. The classification column

Algorithm	Prediction Accuracy on control arff file	Prediction accuracy on experimental arff file
Neural Network	69.05%	89.05%
Random Forest	65.23%	86.13%
Naïve Bayes	59.50%	71.96%
IBK (nearest neighbor)	66.73%	77.49%

Table 3.1: Prediction Accuracies of Several Algorithms on 3E6R Data (Ten-Fold Cross Validation)

represented the secondary-structure classification of the middle amino acid (i.e., the seventh) in the instance. The instances represented all successive subsequences of length 13aa (i.e., a sliding window of size 13 was used). This arff file was meant to serve as a control, since it used no property-based or temporal-context attributes.

Next, the first arff file was converted to a new arff file that replaced the original 13 amino-acid letter attributes with a set of amino-acid property attributes. This was done by exchanging each amino-acid letter for its five Venkatarajan quantifiers and its helical propensity (delimited by commas appropriately) for a total of 78 amino-acid property attributes. The three whole-subsequence hydrophobicity attributes (inter-moment angle, magnitude of positive moment, and magnitude of negative moment) were then added, followed by the output classes of the previous four instances as temporal backward-context attributes. Thus, each instance in the new arff file had a total of 85 attributes in all. This was also done with a Perl Script. Both arff files were then tested using several different machine-learning algorithms in Weka. The results are shown in table 3.1.

The sizable increase achieved in prediction accuracy when using the experimental attribute set suggested that the three-pronged approach of using temporal context attributes, individual amino-acid similarity attributes, and whole-subsequence similarity attributes was potentially more effective than the control approach.

While the results for the proof-of-concept experiments were encouraging, we recognized that there was a need to (1) test these feature sets on larger and commonly used data sets; (2) evaluate the amino-acid property features and the temporal context features separately;

(3) evaluate a larger number of properties; (4) use temporal context features that represented *predicted* classes of neighboring amino acids generated through a relaxation process rather than *known* ones.

3.2 Explanations Regarding Some Available Data Sets

We identified a number of data sets that have been used to benchmark different methods of secondary-structure prediction. For background purposes, a brief explanation of each follows. In general, data sets with higher resolution lead to better prediction accuracy [48].

3.2.1 The "Molecular Biology (Protein Secondary Structure) Data Set" [47]

This data set was originally compiled in 1988 by Ning Qian (Johns-Hopkins University) and Terry Sejnowski (UC-San Diego). They were the first researchers to use a neural-network model to approach secondary-structure prediction, though Robson, Garnier, and Chou & Fasman had all developed and applied different models to the same problem. This data, which was downloaded from the UCI Machine Learning repository, comprises a training set and a test set used in their 1988 paper [47]. They obtained a set of solved protein structures from the Brookhaven National Laboratory [74] (the predecessor to the RCSB protein data bank); a method developed by Kabsch and Sander had been used to assign three-class secondary structure (alpha helix, beta sheet, or coil) based on atomic coordinates found in each protein. Qian and Sejnowski noted that their results were "highly sensitive to homologies between proteins in the testing and training sets," so they divided the 106 proteins into a training set with 91 proteins and a test set with 15 proteins such that there was "no homology" between the training and test sets [47]. They noted—and much of the subsequent research cited above confirms—that much higher prediction accuracies can be achieved on test sets when models are trained with homologous data. Using 13 inputs (similar to our experimental setup), they achieved 64.3% Q_3 accuracy and suggested that "a theoretical limit of 70% [could] be obtained with local methods." To date, their paper has been cited over 1,000 times.

Table 3.2: List of Superseded Protein Structures and their Replacements for Data Set RS126

Superseded ID	New ID
3B5C	1CYO
2STV	2BUK
2GCR	1A45
1WSY	1BKS
3GAP	1G6N
2WRP	2OZ9
1FDX	1DUR
2FXB	1IQZ

3.2.2 The RS126 Data Set [14]

Rost and Sander compiled a set of 126 proteins known as the RS126 data set. The set comprises 126 globular- and 4 membrane-protein chains with less than 25% pairwise identity for lengths greater than 80aa. Subsequent research suggests, though, that pairwise identity is a poor method of measuring sequence similarity. They noted that "the most reliable prediction of the structure of new proteins is done by detection of significant similarities to proteins of known structures." [14 (citing 76)]. Using homology information derived from multiple-sequence alignments, they achieved an overall Q_3 accuracy of 70.8%.

3.2.3 The CB396, CB251, and CB513 Data Sets [61]

In 1999, Cuff and Barton re-iterated that most successful techniques for secondary-structure prediction rely on aligning test instances with homologues [61]. They emphasized that there should be "no detectable sequence similarity" between training and test sets [61]. They explained that up to four fifths of known homologues may be overlooked if only pairwise sequence-alignment methods are used to measure homology [61]. They therefore used more sensitive homology-detection methods to ensure that there was no homology in a set of 554 protein domains with resolutions ≤ 2.5 angstrom that they collected from the 3Dee database of structural domain definitions. Since they wanted to test some algorithms that had already been tested on RS126, they removed domains that had homologues in RS126 and domains that failed to meet some other more stringent requirements. This resulted in CB396. CB513

Table 3.3: List of Superseded Protein Structures and their Replacements for Data Set CB396

Superseded ID	New ID
1AMG	2AMG
1CHB	2CHB
1CTH	2CTH
1CXS	1EU1
1GEP	2GEP
1KIN	1KIM
1TSS	2TSS
2BLT	1XX2
3BCL	4BCL

Table 3.4: List of Superseded Protein Structures and their Replacements for Data Set PSS504

Superseded ID	New ID
1R5R	3BJH
1R0T	1Z7K

was made by adding RS126 to CB396 and removing 9 more domains based on more criteria. CB497 was made by removing the 16 domains in CB513 that are ≤ 30 aa in length.

3.2.4 The PSS504 Data Set [66]

In 2006, Gubbi et al. compiled the PSS504 data set using CATH, a hierarchical classification of protein domain structures published in 1997 by Orengo et al. [66] (The acronym CATH stands for categories used in the classification system: Class, Architecture, Topology, and Homologous superfamily [77]). The sequences included in PSS504 all have pairwise sequence identities (compared to all other respective sequences in the data set) of less than 20%. All of their respective PDB files have a resolution of at least 2 angstrom and are at least 40aa in length; has longer sequences and more residues than CB513.

3.2.5 The EVA6 Data Set [78]

EVA was a project started in 2001 for the purpose of benchmarking protein structure prediction [78]. Limited funding, however, caused the EVA project to be frozen in 2008 [87]. EVA was intended to address not only secondary structure prediction, but also the related

Table 3.5: List of Superseded Protein Structures and their Replacements for Data Set EVA6

Superseded ID	New ID
1KOM	1T23
1NNG	1YLI
1UW2	2VRD
1Z61	1ZAE

problems of comparative modeling, fold recognition/threading, and inter-residue contact predictions [78]. EVA6 is one of several different EVA sets that were compiled before the EVA project was frozen. It was generated by gleaning the latest (at the time) experimentally determined structures from the PDB website. The secondary-structure labels of each amino acid in each respective structure were determined using the DSSP program (which labels secondary structures based on the 3D atomic coordinates found in the PDB files). The extent to which any proteins in the EVA6 data set share homology with each other, though, is not immediately available (to our knowledge).

3.2.6 The PLP399, PLP364, and PLP273 Data Sets [79]

These relatively recent data sets were generated by Bent Petersen et al. in order to test their method for predicting beta-turns. They collected sequences from RCSB using the protein-culling server PISCES. They initially collected 3,572 protein chains with maximum pairwise sequence identities of $\leq 25\%$, resolutions of ≤ 2 angstrom, R-factors of ≤ 0.2 , and sequence lengths ranging from 25–10,000aa [79]. They reduced the initial set of protein chains to 399 (which make up PLP399) by using a Hobohm1 algorithm to ensure that there was minimal homology between all pairs of sequences [79]. As a note, no sequences in PLP399 have more than 25% sequence identity with any sequences in the BT426 data set [79]. PLP 364 consists of all protein chains in PLP399 that were deposited in RCSB between 2008 and 2010, inclusive [79]. PLP273 consists of all protein chains in PLP399 that were deposited in RCSB PDB between 2009 and 2010, inclusive [79].

Table 3.6: List of Superseded Protein Structures and their Replacements for Data Set BT426

Superseded ID	New ID
1GDO	1XFF
5ICB	1IG5
1ALO	1VLB
3B5C	1CYO

Table 3.7: List of Superseded Protein Structures and their Replacements for Data Set BT823

Superseded ID	New ID
1R5R	3BJH

3.2.7 The BT426 Data Set [80]

This data set was collected by Guruprasad and Rajkumar for the purpose of determining dependent positional preferences in beta and gamma turns [80]. They selected a set of 426 protein chains that all had at least one beta or gamma turn; there is $\leq 25\%$ pairwise sequence identity between all chains in the set and chains had a resolution of ≤ 2 angstrom. These protein chains were collected from the RCSB using the program PDB_SELECT.

3.2.8 The BT823 and BT547 Data Sets [81]

Fuchs and Alix compiled the BT547 and BT823 data sets for the purpose of testing their method of predicting beta turns [81]. They chose chains that had at least one beta turn and resolution ≤ 2 angstrom. The extent to which the chains have homology with each other is not listed.

3.2.9 The SPX Data Set [82]

Cheng et al. compiled the SPX data set for the purposes of testing their method of predicting disulfide bridges [82]. They assembled the set by first pulling all proteins having at least one

Table 3.8: List of Superseded Protein Structures and their Replacements for Data Set BT547

Superseded ID	New ID
1GDO	1XFF

intrachain disulfide bridge that were available in the RCSB PDB on May 17, 2004. They then used UniqueProt to exclude a number of chains such that there would be minimal homology in the remaining set. The end result was the set of 1,018 protein chains found in the SPX data set.

3.2.10 The TT1032 Data Set [69]

Thomas Nordahl Petersen et al. compiled the TT1032 by first pulling a large set of proteins available in the RCSB PDB as of August 1999. They excluded any chains that were less than 30aa in length and any chains that did not have ≤ 2.5 angstrom resolution. They then used the Hobohm algorithm to reduce intra-set homology between proteins and inter-set homology with the RS126 data set. They also manually removed transmembrane proteins. The result was the TT1032 data set.

3.3 Finding and Evaluating a Larger Set of Amino-Acid Properties

A large database of physicochemical and biochemical properties of amino acids has been compiled by Kawashima et al. [83]. This database actually has three sections: AAindex1 (individual amino-acid properties), AAindex2 (substitution matrix information), and AAindex3 (statistical protein contact potentials) [83]. For the purposes of this project, we restricted our focus to AAindex1—a compilation of 544 amino-acid properties.

Chapter 4

Primary Experimental Results

4.1 Evaluation of the Relevance of the Amino-Acid Properties

We initially opted to use PLP399 for the feature-selection process. The sequences from PLP399 with DSSP annotations were gleaned from the `ss.txt` file. We then wrote a Perl script to construct an arff file from those sequences. The instances in the arff file consisted of all successive sliding windows of 13 residues; we chose a sliding-window length of 13 residues because Hua identified 13 as the optimal sliding-window length [53] and Qian and Sejnowski also used a window length of 13 in their seminal paper. The letter of each amino acid at each position 1–13 served as an attribute value, so there were 13 attributes in all. The output class (i.e., label) for each instance was the three-class secondary structure label (as defined by Rost and Sander) of the middle residue. We also wanted to predict the structures of residues that were close to the ends of protein sequences. Since each sliding-window instance’s label represented the structure of the middle amino acid, there was a need to create a null category for attributes 1–6 and 8–13 as a space-filler at the edges of each protein. We used an asterisk to represent this null category. When this scheme was used, the PLP399 set produced an arff file with 71,098 instances. This file with only letter attributes was meant to serve as our control.

4.1.1 First Approach to Feature Selection

We downloaded `aaindex1.txt` (the text file containing the entries for Kawashima’s `AAindex1`) and wrote a script to generate an experimental arff file. The experimental file included 13

letter attributes corresponding to the 13 respective amino acids in each instance. For each letter attribute, the experimental arff file also added the 544 quantitative properties found in the aaindex1.txt file as new numeric attributes. Furthermore, for quantitative properties that had both positive and negative possible values, the magnitude of the alpha-helical positive moment, the magnitude of the negative alpha-helical moment, and the magnitude of the inter-moment angle between them for the sliding window were calculated and added as additional whole-subsequence attributes. This approach resulted in an arff file that had 71,098 instances and 7,597 attributes.

We initially attempted to perform feature selection using several different pairs of attribute evaluators and search methods in Weka [84]. However, this approach presented several problems. First, the arff file was so large that many attribute-evaluator/search-method pairs could not be tested because they exhausted all memory on the java heap—even when the heap size was increased to 10 gigabytes. Those that did execute successfully had inordinately long running times and produced results that were difficult to reconcile with each other. One evaluator, for example, would rank a large number of whole-subsequence attributes (i.e., moment magnitudes and inter-moment angles) before any single-position attributes, while another evaluator would rank over 100 single-position attributes before any whole-subsequence attributes.

4.1.2 Second Approach to Feature Selection

Since we believed that the unusually large number of attributes in the experimental arff file might be related to the drastically different results returned by the different attribute evaluators, we decided to generate a new set of 54 arff files wherein each file only contained attributes corresponding to ten amino acid properties (the last file only had attributes corresponding to four properties, since 544 is not evenly divisible by ten). We then ran Weka's ClassifierSubsetEval (using BayesNet as the classifier) on each of the 54 files using ten-fold cross validation. 235 attributes that were selected in at least nine folds of their respective

ten-property arff files were identified. A new 235-attribute arff file was then created and subjected to another round of Weka’s ClassifierSubsetEval (again using BayesNet as the classifier) using ten-fold cross-validation. After this second round, 78 attributes that were selected by at least nine folds were identified. A new 78-attribute arff file was then created and subjected to third round using ten-fold cross-validation. At this point, however, all 78 attributes were selected in at least seven folds. Those 78 attributes are shown in table B of the appendix. For convenience, table 4.1, an abbreviated version of table B that includes some of the more surprising and/or interesting attributes that were selected, is included here.

4.1.3 Third Approach to Feature Selection

Even though our second approach to feature selection did yield results that appeared more intelligible than the results from our first approach, there was lingering doubt about whether we had actually gathered enough information to ascertain the usefulness of the amino-acid properties considered—mainly because we had only used a single attribute evaluator and the results from our first approach had shown that different attribute evaluators often appraised the same attributes (and hence the properties from which those attributes were derived) very differently. Furthermore, each of the 54 arff files used in our second approach had been assigned ten properties based only on the order in which the properties appeared in the aaindex1.txt file. Hence, it was possible that attributes from some properties might have been overlooked because they happened to be grouped with attributes from properties that were even more relevant. In addition, a reference we uncovered in our ongoing research pointed out that using cross-fold validation can introduce sequence similarities between test and training sets when both sets contain non-identical instances that are nonetheless derived from the same protein [61] (e.g., when instances that represent successive sliding windows are put into both the training and test sets). As a result, there was a possibility that we may have unwittingly introduced a bias into our results by using cross-fold validation.

Short and medium range non-bonded energy per residue (Oobatake-Ooi, 1977) for amino acid #6
A parameter of charge transfer capability (Charton-Charton, 1983) for amino acid #7
Molecular weight (Fasman, 1976) for amino acid #7
Magnitude of negative moment for 14 Å contact number (Nishikawa-Ooi, 1986)
Magnitude of positive moment for ALTLS index (Cornette et al., 1987)
Magnitude of positive moment for Apparent partition energies calculated from Wertz-Scheraga index (Guy, 1985)
Magnitude of positive moment for HPLC parameter (Parker et al., 1986)
Magnitude of positive moment for Hydration free energy (Robson-Osguthorpe, 1979)
Magnitude of positive moment for Hydration potential (Wolfenden et al., 1981)
Magnitude of negative moment for Mean polarity (Radzicka-Wolfenden, 1988)
Magnitude of positive moment for Normalized composition from animal (Nakashima et al., 1990)
Angle between moments for Normalized composition from fungi and plant (Nakashima et al., 1990)
Angle between moments for Normalized composition of membrane proteins (Nakashima et al., 1990)
Magnitude of positive moment for Normalized composition of mt-proteins (Nakashima et al., 1990)
Magnitude of positive moment for Principal property value z1 (Wold et al., 1987)
Magnitude of positive moment for Relative partition energies derived by the Bethe approximation (Miyazawa-Jernigan, 1999)
Magnitude of positive moment for Retention coefficient in HFBA (Browne et al., 1982)
Angle between moments for Side chain hydrophathy, uncorrected for solvation (Roseman, 1988)
Angle between moments for Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988)

Table 4.1: These are 20 of the 78 attributes that were selected when the second approach to feature selection was used.

As a result, we decided that it would be necessary to (1) evaluate each property individually, (2) perform evaluations using several different types of dissimilar models, and (3) use separate training and test sets rather than cross-fold validation. We ultimately chose to use CB396 as a training set and RS126 as a test set because there is minimal homology between them and because RS126 provides a test set that is relatively large and commonly known.

Hence, we decided to make 544 individual arff files—one for each amino-acid property—and to test each property individually. We used the respective attributes for each property that had been used in stage 2, but we also added several more attributes for the following reasons. In the preceding experiments, we had only added attributes that represented what the moments for each instance would be if the instance was helical. We therefore decided to add an attribute to represent what the total moment of a property would be if the instance’s sequence was an extended beta strand. We also added an attribute to represent the total alpha-helical and beta-sheet moments of the entire subsequence that made up each instance.

In addition, we added six attributes that represented the alpha-helical moments and six attributes that represented the beta-sheet moments over all six subwindows of size 8 that could be extracted from the each instance’s larger sequence of 13. In this manner, we hoped to elucidate how each moment was changing over the course of the instance. Two instances might have identical total moments, for example, but one’s moment may show a trend of increasing over the course of the instance, while the other might show a trend of decreasing. The former might mean that a helix is starting, while the latter might mean that a helix is ending. At the edges of an alpha-helical or beta-sheet sequence, we believed that the difference could be informative. Like the inter-moment angles, these six attributes are novel contributions that have not been used in any of the literature (to our knowledge).

Ultimately, each of the 544 arff files had the 44 attributes shown in Table 4.2. We then used Weka [84] to generate several different machine-learning models on each arff file. We tried to select a variety of different models, such as a neural-network model (MultiLayerPerceptron),

a regression model (Logistic), a decision-tree model (J48 and DecisionStump), a nearest-neighbor model (IBK and IB1), a Bayesian model (BayesNet), a rule-based model (DTNB), and a homogenous boosting/ensemble model (RandomForest). As a side note, the size of the input files made it impractical to use some models. A support-vector machine model, for example, took approximately ten hours to finish running on a single property’s arff file. Since the SVM model generated did not achieve a high Q_3 accuracy and it would have taken months to create an SVM model for each property using the resources we had at our immediate disposal, we decided not to generate any additional SVM models.

Even though we wrote a script to automate most of the process of generating these models, it took several weeks to generate them all. The overall Q_3 prediction accuracies that each model type achieved using the arff files generated with each property are shown in table C of the appendix. Conditional formatting has been applied in table C to each column so that values that are higher relative to other values in each respective column appear more red. For convenience, some of the properties that improved Q_3 prediction accuracy are shown below in table 4.3.

4.1.4 Conclusions Regarding the Use of Amino-Acid Properties for Secondary-Structure Prediction

Not all models types achieved the same gains in Q_3 accuracy when using the same property arff files. This was to be expected. Nevertheless, with the help of the conditional formatting feature in Excel, trends were clearly visible. The models that achieved the highest Q_3 accuracies, such as Logistic and RandomForest, tended to benefit when using the same property files; a good property file generally improved a good model’s accuracy by 2–3% over the control file. That being said, not all of the algorithms tested benefitted from the addition of the new property-based attributes. The RBFNetwork approach, for example, performed best when using letter attributes only. The NaiveBayes approach also did better with letter attributes than it did with 536 of the 544 property files.

Attribute	Possible values as specified in WEKA
Letter for amino acid #1	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #1	NUMERIC
Letter for amino acid #2	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #2	NUMERIC
Letter for amino acid #3	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #3	NUMERIC
Letter for amino acid #4	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #4	NUMERIC
Letter for amino acid #5	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #5	NUMERIC
Letter for amino acid #6	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #6	NUMERIC
Letter for amino acid #7	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #7	NUMERIC
Letter for amino acid #8	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #8	NUMERIC
Letter for amino acid #9	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #9	NUMERIC
Letter for amino acid #10	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #10	NUMERIC
Letter for amino acid #11	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #11	NUMERIC
Letter for amino acid #12	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #12	NUMERIC
Letter for amino acid #13	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #13	NUMERIC
Magnitude of positive alpha-helical moment for <property name>	NUMERIC
Magnitude of negative alpha-helical moment for <property name>	NUMERIC
Angle between positive and negative alpha-helical moments for <property name>	NUMERIC
Magnitude of total alpha-helical moment for <property name>	NUMERIC
Magnitude of total beta-sheet moment for <property name>	NUMERIC
Magnitude of total alpha-helical moment for <property name> over subwindow 1	NUMERIC
Magnitude of total alpha-helical moment for <property name> over subwindow 2	NUMERIC
Magnitude of total alpha-helical moment for <property name> over subwindow 3	NUMERIC
Magnitude of total alpha-helical moment for <property name> over subwindow 4	NUMERIC
Magnitude of total alpha-helical moment for <property name> over subwindow 5	NUMERIC
Magnitude of total alpha-helical moment for <property name> over subwindow 6	NUMERIC
Magnitude of total beta-sheet moment for <property name> over subwindow 1	NUMERIC
Magnitude of total beta-sheet moment for <property name> over subwindow 2	NUMERIC
Magnitude of total beta-sheet moment for <property name> over subwindow 3	NUMERIC
Magnitude of total beta-sheet moment for <property name> over subwindow 4	NUMERIC
Magnitude of total beta-sheet moment for <property name> over subwindow 5	NUMERIC
Magnitude of total beta-sheet moment for <property name> over subwindow 6	NUMERIC
Secondary Structure Label for middle amino acid	{H,E,N}

Table 4.2: These are the 44 attributes that were used in each individual-property arff file.

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2, 100, 3
Letters only (Control)	n/a	62.5021	60.9603	60.4232	59.9065	61.2514	59.2833	61.2473	60.7143	53.7395
14 A contact number (Nishikawa-Ooi, 1986)	0	65.4215	62.8875	63.5312	63.3877	61.6287	59.1315	58.9183	59.1151	58.8486
ALTFT index (Cornette et al., 1987)	18	65.0443	62.58	63.4492	63.0556	61.969	59.0823	59.0864	58.742	57.8235
Bitterness (Venanzi, 1984)	84	64.3718	62.2601	62.3954	60.4109	57.0239	55.6544	58.455	57.4832	56.5934
Buriability (Zhou-Zhou, 2004)	86	65.0976	62.4036	63.4943	62.6415	62.2355	58.2746	59.9516	58.2582	57.3069
Effective partition energy (Miyazawa-Jernigan, 1985)	109	65.4543	62.744	64.1258	63.3262	62.9121	59.7425	61.1284	59.8409	58.1926
HPLC parameter (Parker et al., 1986)	137	65.36	63.2401	63.5066	62.9613	60.6364	59.4268	59.193	59.2792	58.4755
Hydrophathies of amino acid side chains, pi-values in pH 7.0 (Roseman, 1988)	150	65.2329	62.4938	63.318	61.3621	60.8414	59.3858	58.7051	58.6518	56.9173
Hydrophobic parameter pi (Fauchere-Pliska, 1983)	162	65.2575	62.6292	63.6543	62.7235	61.928	60.8619	59.0085	59.1151	58.4509
Hydrophobicity index, 3.0 pH (Cowan-Whittaker, 1990)	175	65.2247	62.6251	63.2852	61.805	61.3375	58.3074	58.2705	57.9178	57.4258
Information value for accessibility; average fraction 35% (Biou et al., 1988)	194	65.565	62.8465	63.6871	63.0105	61.2637	60.5995	58.6887	59.0495	58.7092
Interactivity scale obtained from the contact matrix (Bastolla et al., 2005)	197	65.1714	62.9736	63.3672	63.1499	62.1699	57.803	59.2997	58.7543	58.2213
Linker index (Bae et al., 2005)	206	64.5686	62.4077	63.0966	61.7763	61.6287	58.824	58.8445	59.0905	57.7251
Long range non-bonded energy per atom (Oobatake-Ooi, 1977)	218	64.6876	62.3216	63.2647	62.5595	62.6169	59.4883	61.2145	58.7133	57.352
Mean polarity (Radzicka-Wolfenden, 1988)	222	65.3805	62.8752	63.7814	63.1212	61.8214	60.3822	59.2669	59.3243	58.3443
Optimized relative partition energies - method D (Miyazawa-Jernigan, 1999)	344	65.7372	62.5841	63.5969	63.3508	60.3945	60.4232	58.9347	59.0372	58.1557
PRILS index (Cornette et al., 1987)	348	65.2862	62.9121	63.4246	62.8629	61.7722	59.0372	59.2176	59.2997	59.2464
Partition energy (Guy, 1985)	352	65.0689	62.9244	63.6707	62.6907	61.2637	59.4391	58.824	59.6605	57.7702
Polarity (Grantham, 1974)	356	65.0484	62.338	63.3754	62.8219	62.8957	60.8414	61.3539	57.7825	58.5903
Principal property value z1 (Wold et al., 1987)	364	65.2657	63.1909	63.8183	62.8055	61.928	59.9352	59.0126	58.7789	58.3853
Relative partition energies derived by the Bethe approximation (Miyazawa-Jernigan, 1999)	383	65.6265	63.0023	63.8962	63.0597	61.9444	60.1935	59.4596	59.4965	59.152
Retention coefficient at pH 2 (Guo et al., 1986)	408	65.0115	63.1253	62.8219	62.5144	60.9603	59.0782	58.9429	59.0331	57.7989
Side chain interaction parameter (Krigbaum-Komoriya, 1979)	429	64.421	61.7722	63.0802	62.5677	62.2068	59.1192	60.5749	58.4755	57.7087
Solvation free energy (Eisenberg-McLachlan, 1986)	447	64.7081	62.7686	63.2893	62.3462	61.194	60.333	58.7789	59.1561	57.7866
TOTLS index (Cornette et al., 1987)	459	65.196	62.744	63.8716	63.2196	61.6451	58.3402	59.1069	59.5252	58.291

Table 4.3: These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table 4.2 on selected properties.

Most of the properties that yielded notable gains for the best models were related to hydrophobicity, hydrophilicity, hydropathy, polarity, buriability, partition energy, average number of surrounding residues (e.g., contact number), and structural propensity. These results are, at the very least, very consistent with the wealth of previous research that identifies hydrophobicity as a property that is useful for secondary-structure prediction. To our knowledge, however, a few of these properties, such as buriability and partition energy, have not been specifically used before to enhance secondary structure prediction. However, the extent to which synergistic benefit might result from using buriability and partition energy alongside some of the properties that are already known to improve secondary-structure prediction is unclear because buriability and partition energy are correlated to some extent with some of those known properties (e.g., hydrophobicity).

Thus, for the purpose of secondary-structure prediction, it appears that some properties definitely do matter, while others probably do not. It also appears that the feature set we developed, which included some novel features like the inter-moment angle and the moments over subwindows, succeeded to some extent in facilitating better comparisons between instances with dissimilar sequences.

4.2 Using Majority-Vote Ensembles to Raise Prediction Accuracy

At this point in our research, we decided to investigate whether heterogenous ensemble models could be used to achieve a better overall Q_3 accuracy. Ensemble models that combine classifiers can often improve prediction performance [85]. Researchers in machine learning generally agree that "[d]iversity is a crucial condition for obtaining accurate ensembles" [85]. Some researchers have successfully created diversity in the component classifiers of ensembles by training each classifier on a different feature set [85]. In light of these considerations, we recognized that we had a unique opportunity to experiment with ensemble creation because the process of evaluating each property individually with several different machine-learning

algorithms had produced several thousand models that were trained with different feature sets using different types of classifiers.

4.2.1 First round of Majority-Vote Ensembles

We wrote a Perl program that determines the prediction accuracies (Q_3 , P_α , P_β , P_{coil}) of all majority-vote ensembles of an arbitrary number r of classifiers. These classifiers are selected from a total repository of n classifiers whose Weka [84] output buffers (including predictions for each instance) are stored in a given directory. Hence, the total number of non-redundant majority-vote ensembles of size r taken from a set of n classifiers is

$$\binom{n}{r} \tag{4.1}$$

Given that the number of ensembles therefore increases exponentially, we decided that it would be best to define a relatively small subset of the models generated for inclusion in our ensemble experiments.

4.2.1.1 Selecting a Set of Classifiers

Since raising Q_3 accuracy was our primary goal, we decided to add 23 of the most successful (i.e., having relatively high Q_3 accuracy) Logistic models to our set of classifiers. In addition, we added 6 of the most successful RandomForest models, 6 of the most successful BayesNet models, 2 of the most successful IBK models, and 1 successful DTNB model. Through some parameter modification and/or use of meta techniques available in Weka [84] (e.g., boosting, bagging, MultiClassClassifier, and CostSensistiveClassifier), we also teased out a number of other models with high Q_3 accuracies that were added to the classifier set.

In considering which models to include in our set of classifiers, we took note of the fact that there was a consistent imbalance between P_α , P_β , and P_{coil} regardless of the property considered and regardless of the model type used. P_β , in particular, was consistently about

10–20% lower than P_α and P_α was consistently about 10% lower than P_{coil} . Hence, in order to address the imbalance issue, we also included several cost-sensitive versions of the best models included in the set of classifiers. Each cost-sensitive model was designed to elevate exactly one of P_α , P_β , or P_{coil} at the expense of the other two.

4.2.1.2 Summary of Approaches used to Create and Verify Diversity

Summarily, then, there were three approaches we wanted to use to create diversity: (1) using different types of models, (2) using models that were trained using different properties, and (3) using models that were trained using cost-sensitivity. Ultimately, the classifier set included 66 models.

Before proceeding, we wanted to apply some method to verify that the three approaches we had used to create diversity had been effectual to at least some degree. Yule’s Q statistic for two classifiers, D_i and D_k , is defined as

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \quad (4.2)$$

where N^{11} is the number of instances correctly classified by both D_i and D_k , N^{00} is the number of instances incorrectly classified by both D_i and D_k , N^{10} is the number of instances correctly classified by D_i and incorrectly classified by D_k , and N^{01} is the number of instances correctly classified by D_k and incorrectly classified by D_i [86]. The expected value of $Q_{i,k}$ is zero for classifiers that are uncorrelated (i.e., independent) [86]. $Q_{i,k}$ can vary between -1 and 1; classifiers that generally classify the same objects correctly will have positive values of $Q_{i,k}$, while classifiers that generally commit errors on different objects tend to have negative values of $Q_{i,k}$ [86].

To visualize the pattern of diversity in the classifier set, we calculated Yule’s Q statistic for all combinations of two classifiers selected from the 66 models in the classifier set. After inspecting the results, as shown in Table D of the appendix, we were satisfied that all three

approaches for creating diversity had been effectual to some degree. For convenience, an exemplary portion of table D is shown in table 4.4.

4.2.1.3 Results for First Round of Majority-Vote Ensembles

We then used our Perl program to determine the prediction accuracies of all majority-vote ensembles consisting of combinations of 3, 5, and 7 classifiers selected from the classifier set. Since the best individual models included in the classifier set achieved Q_3 accuracies of up to 65%, we configured the program to identify any ensembles that achieved a threshold value of 66% Q_3 accuracy.

There were 4 ensembles of 3 classifiers (i.e., 0.00874% of the total number of ensembles of 3) that achieved 66% Q_3 accuracy. There were 254 ensembles of 5 (i.e., 0.00284% of the total number of ensembles of 5) and 5,673 ensembles of 7 (i.e., 0.000728% of the total number of ensembles of 7) that achieved 66% Q_3 accuracy. Hence, the number of ensembles achieving greater than 66% Q_3 accuracy does increase as the ensemble size increases, but at a rate that is smaller than the exponential rate at which the search space of possible ensembles increases.

Table 4.5 shows the number of times each classifier was used in ensembles that achieved 66% Q_3 accuracy.

We observed an interesting phenomenon in ensembles of size 7: a large number of the ensembles of size 7 that achieved the threshold accuracy used one or more of the cost-sensitive models. Furthermore, a number of models that had been used a moderate number of times in ensembles of size 5 were not used at all in ensembles of size 7 that achieved 66% Q_3 accuracy. Intrigued, we decided to further explore the influence and relevance of cost-sensitive models on majority-vote ensembles by performing a second round of ensemble creation as explained below.

	AttributeSelected_Bagging_15_RBF_7_56.txt	BayesNet_109.txt	BayesNet_351.txt	BayesNet_356.txt	BayesNet_53.txt
AttributeSelected_Bagging_15_RBF_7_56.txt	1				
BayesNet_109.txt	0.85366	1			
BayesNet_351.txt	0.86217	0.96421	1		
BayesNet_356.txt	0.85857	0.95471	0.97313	1	
BayesNet_53.txt	0.85383	0.96951	0.95382	0.96738	1
BayesNet_56.txt	0.88148	0.96612	0.95498	0.95948	0.95871
BayesNet_57.txt	0.88465	0.97309	0.96885	0.96094	0.96018
CostSensitive(E_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0.80295	0.94299	0.97938	0.9497	0.9332
CostSensitive(E_2.0)_Logistic_344_noInd.txt	0.83331	0.95526	0.94343	0.94238	0.94726
CostSensitive(E_2.0)_Logistic_86.txt	0.81715	0.94825	0.94711	0.93594	0.9359
CostSensitive(E_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0.7888	0.98051	0.93976	0.92495	0.94269
CostSensitive(E_2.0)_RF_225_5_383.txt	0.80924	0.888	0.87889	0.87066	0.87015
CostSensitive(H_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0.86053	0.931	0.96817	0.94328	0.92583
CostSensitive(H_2.0)_Logistic_344_noInd.txt	0.86527	0.91219	0.90852	0.90547	0.91392
CostSensitive(H_2.0)_Logistic_86.txt	0.85722	0.91045	0.91282	0.89969	0.9038
CostSensitive(H_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0.85942	0.96642	0.94083	0.92666	0.94482
CostSensitive(N_1.2)_Logistic_344_noInd.txt	0.92591	0.93407	0.93157	0.93034	0.93286
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0.90779	0.95816	0.91269	0.89562	0.91488
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0.9099	0.90542	0.95928	0.92116	0.89946
CostSensitive(N_2.0)_Logistic_86.txt	0.90092	0.80513	0.82594	0.81941	0.80324
DTNB_109.txt	0.91072	0.97042	0.95044	0.93948	0.94616
IBK_60_w_173.txt	0.88786	0.79201	0.78874	0.78493	0.7975
IBK_60_w_344.txt	0.88468	0.79218	0.79019	0.79835	0.80282
Logistic_0.txt	0.91813	0.94949	0.94866	0.94048	0.94718
Logistic_1.txt	0.9116	0.94069	0.93888	0.93927	0.95268
Logistic_109.txt	0.91855	0.95161	0.95415	0.93995	0.9379
Logistic_137.txt	0.91234	0.94475	0.95383	0.94323	0.93759
Logistic_150.txt	0.9127	0.94452	0.9504	0.94697	0.94218
Logistic_18.txt	0.91135	0.94156	0.94795	0.93293	0.93292
Logistic_194.txt	0.92293	0.9519	0.95013	0.94359	0.94702
Logistic_195.txt	0.91414	0.94116	0.94645	0.94031	0.94232
Logistic_196.txt	0.90331	0.9423	0.94849	0.93568	0.93583
Logistic_197.txt	0.90817	0.94019	0.94815	0.93909	0.93911
Logistic_222.txt	0.92137	0.95017	0.95187	0.94188	0.94461
Logistic_344.txt	0.92043	0.9495	0.94622	0.94487	0.94874
Logistic_347.txt	0.91248	0.94381	0.94453	0.93619	0.94144
Logistic_348.txt	0.91136	0.94267	0.94538	0.93655	0.93923
Logistic_352.txt	0.91736	0.94703	0.9487	0.93625	0.9344
Logistic_356.txt	0.91201	0.94073	0.94515	0.95122	0.94468
Logistic_364.txt	0.91068	0.94386	0.95623	0.9467	0.93544

Table 4.4: These are exemplary pairwise Yule’s Q statistics for combinations of two classifiers selected from the 66 models in the classifier set for the first round of majority-vote ensembles.

Classifier	Times used in ensembles of 3	Times used in ensembles of 5	Times used in ensembles of 7
AttributeSelected_Bagging_15_RBF_7_56.txt	0	91	459
BayesNet_109.txt	0	163	681
BayesNet_351.txt	0	0	1896
BayesNet_356.txt	0	0	1259
BayesNet_53.txt	0	0	494
BayesNet_56.txt	0	0	3454
BayesNet_57.txt	0	0	371
CostSensitive(E_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0	44	688
CostSensitive(E_2.0)_Logistic_344_noInd.txt	0	14	635
CostSensitive(E_2.0)_Logistic_86.txt	0	0	386
CostSensitive(E_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0	137	2903
CostSensitive(E_2.0)_RF_225_5_383.txt	0	0	1099
CostSensitive(H_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0	0	1512
CostSensitive(H_2.0)_Logistic_344_noInd.txt	0	0	2712
CostSensitive(H_2.0)_Logistic_86.txt	0	0	456
CostSensitive(H_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	4	83	4691
CostSensitive(N_1.2)_Logistic_344_noInd.txt	0	0	2457
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0	0	2164
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0	8	1820
CostSensitive(N_2.0)_Logistic_86.txt	1	58	1789
DTNB_109.txt	3	190	3725
IBK_60_w_173.txt	0	6	681
IBK_60_w_344.txt	0	5	3379
Logistic_0.txt	0	13	0
Logistic_1.txt	1	25	0
Logistic_109.txt	0	2	0
Logistic_137.txt	0	4	0
Logistic_150.txt	0	9	0
Logistic_18.txt	0	3	0
Logistic_194.txt	0	16	0
Logistic_195.txt	0	9	0
Logistic_196.txt	0	14	0
Logistic_197.txt	0	43	0
Logistic_222.txt	0	8	0
Logistic_344.txt	0	8	0
Logistic_347.txt	1	2	0
Logistic_348.txt	0	1	0
Logistic_352.txt	2	12	0
Logistic_356.txt	0	17	0
Logistic_364.txt	0	7	0
Logistic_383.txt	0	6	0
Logistic_408.txt	0	7	0
Logistic_458.txt	0	14	0
Logistic_459.txt	0	0	0
Logistic_86.txt	0	2	0
Logistic_98.txt	0	14	0
LogitBoost_285_DecisionStump_18.txt	0	0	0
LogitBoost_285_DecisionStump_344.txt	0	0	0
MAX_RF_225_5(63.5148).txt	0	0	0
MLP_H62_53.txt	0	0	0
MultiBoost_10_BayesNet_351.txt	0	0	0
MultiBoost_10_MLP_H62_56.txt	0	0	0
MultiBoost_15_BayesNet_356.txt	0	4	0
MultiClassClassifier_BayesNet_351.txt	0	58	0
MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0	3	0
MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0	5	0
MultiClassClassifier_MultiBoost_BayesNet_351.txt	0	44	0
RF_225_10_137.txt	0	9	0
RF_225_5_137.txt	0	23	0
RF_225_5_173.txt	0	78	0
RF_225_5_173noInd.txt	0	11	0
RF_225_5_364.txt	0	0	0
RF_225_5_383.txt	0	0	0
RF_225_5_408.txt	0	0	0
RF_225_5_419.txt	0	0	0
ZeroR173.txt	0	0	0

Table 4.5: These are the number of times each classifier was used in ensembles that achieved at least 66% Q_3 accuracy in the first round of majority-vote ensembles.

4.2.2 Second Round of Majority-Vote Ensembles

After seeing the results of the first round of ensemble generation, we wanted to investigate whether including additional cost-sensitive models in the set of classifiers could help create additional diversity that would lead to more majority-vote ensembles with higher Q_3 accuracy. Machine-learning literature suggests that helpful diversity can be created by varying model types, feature sets, and general input parameters (*see* [85]). In addition, it stands to reason that ensembles of cost-sensitive models can be expected to improve recognition of a *minority* class in imbalanced data sets. However, we have not yet come across any literature that suggests that including cost-sensitive models and non-cost-sensitive models together in set of classifiers can lead to ensembles that have greater *overall* prediction accuracy. Hence, we felt it was worth doing a second round of ensemble creation with a modified classifier set that included more cost-sensitive model variations in order to explore this possibility.

4.2.2.1 Selecting a Set of Classifiers Including More Cost-sensitive Models

First, we selected four base models that had achieved relatively high Q_3 accuracies: RandomForest, BayesNet (paired with MultiBoost), DecisionStump (paired with LogitBoost), and Logistic (paired with MultiBoost). We added the best versions of these models (e.g., those achieving highest Q_3 accuracies) to the classifier set. In addition, we derived seven cost-sensitive models from each base model: three models in which a single class’s prediction accuracy was elevated (i.e., a model with elevated P_α , a model with elevated P_β , and a model with elevated P_{coil}), three models in which two of the three classes’ prediction accuracies were elevated (i.e., a model with elevated P_α and P_β , a model with elevated P_α and P_{coil} , and a model with elevated P_β and P_{coil}), and a model in which P_α , P_β , and P_{coil} were constrained to all be within 2% of each other. In the models that had a single elevated class, we tuned the cost-sensitivity parameters so that the prediction accuracies for two non-elevated classes were within 2% of each other. In the models that had two elevated classes, we tuned the cost-sensitivity parameters so that the prediction accuracies for the two elevated classes were

within 2% of each other. These cost sensitive models were also added to the classifier set. Finally, we added an RBFNetwork model, a MultilayerPerceptron Model, an IBK (nearest neighbor) model, a DTNB (rule-based) model, and an additional boosted DecisionStump model to the classifier set. Thus, the classifier set included the 37 models in total. The Yule's Q statistics for all pairs of classifiers in this second round's classifier set are shown in table F of the appendix. For convenience, an exemplary portion of table F is shown in table 4.6.

4.2.2.2 Results for Second Round of Majority-Vote Ensembles

We then used our Perl program to determine the prediction accuracies of all majority-vote ensembles consisting of combinations of 3, 5, and 7 classifiers selected from the new classifier set. We again configured the program to identify any ensembles that achieved a threshold value of 66% Q_3 accuracy. As was the case with the first round, the number of ensembles achieving greater than 66% Q_3 accuracy increased as the ensemble size increased, but at a rate that was smaller than the exponential rate at which the search space of possible ensembles increased.

There were 6 ensembles of 3 classifiers (i.e., 0.0773% of the total number of ensembles of 3) that achieved 66% Q_3 accuracy. There were 300 ensembles of 5 (i.e., 0.0688% of the total number of ensembles of 5) and 5,576 ensembles of 7 (i.e., 0.0542% of the total number of ensembles of 7) that achieved 66% Q_3 accuracy. Table 4.7 shows the number of times each model type was used in ensembles that achieved 66% Q_3 accuracy.

Again, cost-sensitive models were used much more frequently in ensembles of size 7 than in ensembles of 5 or 3. In ensembles of 5, however, at least one cost-sensitive model in which P_α , P_β , and P_{coil} were constrained to all be within 2% of each other (i.e., an "EVEN" model) was used in 279 of the 300 ensembles that achieved 66% Q_3 accuracy. In ensembles of 7, at least one cost-sensitive EVEN model was used in 4,549 of the 5,576 ensembles that achieved 66% Q_3 accuracy, while at least one cost-sensitive MAX model was used in 4,827 of the 5,576 ensembles.

	AttributeSelected_Bagging_15_RBF_7_56.txt	DTNB_109.txt	EVEN_LogitBoost_285_DecisionStump_173_(63.3877).txt	EVEN_MultiBoost_13_BayesNet_(61.5).txt	EVEN_MultiBoost_13_Logistic_(63.3).txt	EVEN_RF_225_5(63.0023).txt
AttributeSelected_Bagging_15_RBF_7_56.txt	1					
DTNB_109.txt	0.91072	1				
EVEN_LogitBoost_285_DecisionStump_173_(63.3877).txt	0.84788	0.93009	1			
EVEN_MultiBoost_13_BayesNet_(61.5).txt	0.80013	0.90961	0.95008	1		
EVEN_MultiBoost_13_Logistic_(63.3).txt	0.83138	0.92795	0.98985	0.95367	1	
EVEN_RF_225_5(63.0023).txt	0.83409	0.88769	0.92539	0.92027	0.93105	1
IBK_60_w_344.txt	0.88468	0.87504	0.76627	0.70324	0.75272	0.77359
LogitBoost_285_DecisionStump_173.txt	0.92599	0.97005	0.96477	0.90413	0.94973	0.89099
LogitBoost_285_DecisionStump_344.txt	0.92095	0.97038	0.95734	0.90219	0.95646	0.89855
MAX_Logistic_(65.8767).txt	0.92598	0.96812	0.92632	0.88052	0.93002	0.87617
MAX_MultiBoost_13_BayesNet_(64.3021).txt	0.89438	0.96792	0.93228	0.96531	0.92785	0.90005
MAX_RF_225_5(65.0689).txt	0.91569	0.92167	0.86466	0.85518	0.86674	0.92602
MLP_H62_53.txt	0.86771	0.86709	0.82096	0.7636	0.79745	0.78982
RAISE_EN_LogitBoost_285_DecisionStump_173_(70.9).txt	0.73627	0.79959	0.92631	0.87797	0.92864	0.85985
RAISE_EN_MultiBoost_13_BayesNet_(71.3).txt	0.5533	0.70567	0.74073	0.80871	0.75692	0.715
RAISE_EN_MultiBoost_13_Logistic_(70.2).txt	0.7758	0.8556	0.95023	0.91756	0.9708	0.89689
RAISE_EN_RF_225_5(70.2).txt	0.7407	0.77863	0.85232	0.85119	0.86451	0.9135
RAISE_E_LogitBoost_285_DecisionStump_173_(78.3).txt	0.67164	0.79454	0.96035	0.92031	0.9597	0.89197
RAISE_E_MultiBoost_13_BayesNet_(84.4).txt	0.49187	0.65826	0.8594	0.92082	0.87373	0.84228
RAISE_E_MultiBoost_13_Logistic_(70.9).txt	0.77077	0.88894	0.98188	0.94886	0.99617	0.9239
RAISE_E_RF_225_5(86.9).txt	0.45283	0.57248	0.81443	0.82149	0.8337	0.86636
RAISE_HE_LogitBoost_285_DecisionStump_173_(73).txt	0.51367	0.68962	0.90988	0.85524	0.89178	0.81634
RAISE_HE_MultiBoost_13_BayesNet_(71.3).txt	0.49276	0.69374	0.85676	0.91728	0.86391	0.81661
RAISE_HE_MultiBoost_13_Logistic_(73.6).txt	0.49444	0.68698	0.88236	0.85456	0.90797	0.8159
RAISE_HE_RF_225_5(74.4).txt	0.4082	0.56902	0.78655	0.78448	0.79464	0.82197
RAISE_HN_LogitBoost_285_DecisionStump_173_(72.1).txt	0.93087	0.95765	0.92598	0.84088	0.89445	0.83228
RAISE_HN_MultiBoost_13_BayesNet_(71.3).txt	0.90153	0.93643	0.83036	0.87345	0.81571	0.79908
RAISE_HN_MultiBoost_13_Logistic_(70.4).txt	0.91333	0.96682	0.94664	0.88922	0.94798	0.87804
RAISE_HN_RF_225_5(74.4).txt	0.89006	0.87361	0.73029	0.6983	0.71432	0.78028
RAISE_H_LogitBoost_285_DecisionStump_173_(78.1).txt	0.7982	0.89743	0.96397	0.90772	0.94972	0.8796
RAISE_H_MultiBoost_13_BayesNet_(84.4).txt	0.6967	0.81835	0.84558	0.87085	0.83605	0.79527
RAISE_H_MultiBoost_13_Logistic_(86.1).txt	0.72401	0.83236	0.88011	0.827	0.86891	0.79658
RAISE_H_RF_225_5(91.6).txt	0.58116	0.66832	0.71299	0.68691	0.69168	0.70391
RAISE_N_LogitBoost_285_DecisionStump_173_(90.7).txt	0.85958	0.84845	0.71163	0.67334	0.69519	0.6851
RAISE_N_MultiBoost_13_BayesNet_(85).txt	0.86574	0.9014	0.79667	0.8337	0.79861	0.80081
RAISE_N_MultiBoost_13_Logistic_(88.9).txt	0.8897	0.90147	0.77434	0.75102	0.78575	0.75928
RAISE_N_RF_225_5(86).txt	0.87912	0.86652	0.77264	0.76532	0.77307	0.83268

Table 4.6: These are exemplary Yule’s Q statistics for combinations of two classifiers selected from the 37 models in the classifier set for the second round of majority-vote ensembles.

Classifier	Times used in ensembles of 3	Times used in ensembles of 5	Times used in ensembles of 7
AttributeSelected_Bagging_15_RBF_7_56.txt	0	39	2866
DTNB_109.txt	1	56	1172
EVEN_LogitBoost_285_DecisionStump_173_(63.3877).txt	0	21	385
EVEN_MultiBoost_13_BayesNet_(61.5).txt	0	99	1209
EVEN_MultiBoost_13_Logistic_(63.3).txt	0	220	2848
EVEN_RF_225_5(63.0023).txt	1	137	2606
IBK_60_w_344.txt	5	241	4691
LogitBoost_285_DecisionStump_173.txt	0	64	1176
LogitBoost_285_DecisionStump_344.txt	1	128	2027
MAX_Logistic_(65.8767).txt	6	149	4068
MAX_MultiBoost_13_BayesNet_(64.3021).txt	1	20	591
MAX_RF_225_5(65.0689).txt	2	49	1989
MLP_H62_53.txt	1	7	1914
RAISE_EN_LogitBoost_285_DecisionStump_173_(70.9).txt	0	12	883
RAISE_EN_MultiBoost_13_BayesNet_(71.3).txt	0	11	295
RAISE_EN_MultiBoost_13_Logistic_(70.2).txt	0	26	646
RAISE_EN_RF_225_5(70.2).txt	0	10	579
RAISE_E_LogitBoost_285_DecisionStump_173_(78.3).txt	0	18	438
RAISE_E_MultiBoost_13_BayesNet_(84.4).txt	0	12	214
RAISE_E_MultiBoost_13_Logistic_(70.9).txt	0	14	667
RAISE_E_RF_225_5(86.9).txt	0	46	279
RAISE_HE_LogitBoost_285_DecisionStump_173_(73).txt	0	121	346
RAISE_HE_MultiBoost_13_BayesNet_(71.3).txt	0	0	407
RAISE_HE_MultiBoost_13_Logistic_(73.6).txt	0	0	484
RAISE_HE_RF_225_5(74.4).txt	0	0	1311
RAISE_HN_LogitBoost_285_DecisionStump_173_(72.1).txt	0	0	1432
RAISE_HN_MultiBoost_13_BayesNet_(71.3).txt	0	0	23
RAISE_HN_MultiBoost_13_Logistic_(70.4).txt	0	0	429
RAISE_HN_RF_225_5(74.4).txt	0	0	153
RAISE_H_LogitBoost_285_DecisionStump_173_(78.1).txt	0	0	233
RAISE_H_MultiBoost_13_BayesNet_(84.4).txt	0	0	41
RAISE_H_MultiBoost_13_Logistic_(86.1).txt	0	0	83
RAISE_H_RF_225_5(91.6).txt	0	0	128
RAISE_N_LogitBoost_285_DecisionStump_173_(90.7).txt	0	0	558
RAISE_N_MultiBoost_13_BayesNet_(85).txt	0	0	228
RAISE_N_MultiBoost_13_Logistic_(88.9).txt	0	0	588
RAISE_N_RF_225_5(86).txt	0	0	1045

Table 4.7: These are the number of times each classifier was used in ensembles that achieved at least 66% Q_3 accuracy in the second round of majority-Vote ensembles.

4.2.3 Conclusions and Possible Directions for Future Research Regarding Majority-Vote Ensembles that have Diversity Generated from the Three Approaches

Our two rounds of experiments with majority-vote ensembles answered some questions, but also engendered many new questions and directions for future research that could be pursued (though they would be beyond the scope of this project). We discuss these issues in turn.

First, both rounds of ensemble experiments seem to suggest that diversity that is helpful for increasing the overall prediction accuracy of majority-vote ensembles can indeed be created by using cost-sensitive versions of one or more classifiers. Cost-sensitive classifiers that are tuned to predict all output classes with similar accuracy seem to be particularly useful, at least in ensembles of the sizes considered in our experiments. In majority-vote ensembles using at least 7 classifiers, cost-sensitive classifiers that are tuned to only increase the prediction accuracies of one or two output classes may also be helpful as well. Hence, it appears that cost-sensitivity can be leveraged not only for increasing the prediction accuracy for a single output class in the context of a single classifier, but also for increasing *overall* prediction accuracy in the context of majority-vote ensembles.

Second, both rounds of ensemble experiments support the proposition that diversity can be generated by training classifiers on different feature sets and by using different classifier models. This is consistent with what was expected, since both of these two approaches are fairly commonly known methods for creating diversity.

There are, however, a number of questions that could be explored in further research. For example, though all three approaches succeeded in creating diversity, it is unclear how much benefit accrues from each approach individually and to what extent the different approaches have a cumulative synergistic effect. In addition, it would be useful to explore whether the most successful ensembles follow a pattern that might be exploited so that the search space of possible majority-vote ensembles can be explored more efficiently. Do most of the best ensembles, for example, consist of classifiers that meet a baseline overall accuracy? Does the distribution of pairwise Yule Q statistics between classifiers in the best ensembles

BayesNet	IBK	DTNB	RandomForest	NaiveBayes	J48	RBFNetwork	Multilayer Perceptron	Logistic
99.84%	99.01%	100%	100%	99.80%	100%	99.86%	100%	100%

Table 4.8: Q_3 Accuracies of Classifiers Using CB396 Training Set and RS126 Test Set with True Output Classes of 8 Neighboring Instances Used as Temporal Context Features

follow a specific pattern? Given an ensemble of size n , is there a way to select or generate an $(n + 1)^{th}$ classifier—perhaps using cost-sensitivity—that can be added to the ensemble (or swapped in) and predictably increase overall prediction accuracy? Can these approaches for creating diversity somehow be harnessed to create ensembles that achieve high prediction accuracy while using constituent classifiers that achieve relatively low accuracy? These are some of the questions that occurred to us. However, in order to avoid expanding the project scope unreasonably, we decided it was prudent to move forward and explore the relevance of temporal context nodes rather than drill deeper into the ensemble questions.

4.3 Evaluating the use of Temporal Context Nodes

4.3.1 Relaxation

For a first step, we decided to establish an upper bound of Q_3 accuracy that we might expect to achieve using temporal context nodes by creating test and training sets that included the true output classes of instances $n - 4$ through $n - 1$ and instances $n + 1$ through $n + 4$ as attributes for each instance n . In addition, each instance n had the original 13 amino-acid letter features. Using CB396 as a training set and RS126 as test set, we created several different classifiers. The Q_3 accuracies of those classifiers are shown in table 4.8.

Since the best models achieved up to 100% Q_3 accuracy, we were initially very optimistic. If 100% accuracy was possible when the true secondary structures of an amino acid’s neighboring amino acids were known, we reasoned that we might be able to achieve good prediction results by (1) predicting the output classes for the instances in the test set in a first iteration without using temporal context features, (2) using the predictions from

Classifier	Iteration				
	0	1	2	3	4
BayesNet	61.2414	61.2227	61.2268	61.2309	61.2309
DTNB	60.4232	60.4232	60.4232	60.4232	60.4232
IBK	60.333	60.5011	60.5298	60.5339	60.5339
J48	53.7395	53.8913	53.8913	53.8913	53.8913
Logistic	62.5021	62.5021	62.5021	62.5021	62.5021
NaiveBayes	61.2473	61.1899	61.1817	61.1858	61.1858
RBFNetwork	60.7143	61.0300	61.0177	61.0300	61.0300
RandomForest	60.9603	61.2514	61.2514	61.2514	61.2514

Table 4.9: Q_3 Accuracies Achieved in Successive Iterations Using the Relaxation Process

the first iteration as temporal context features for a second iteration, and (3) continuing to use predictions from previous iterations in successive iterations until the Q_3 accuracy relaxed into an asymptotic value. We believed that such a process would likely yield at least some increase in Q_3 because some errors that might occur in the first iterations, such as predicted alpha-helical sequences interrupted by single-amino-acid beta sheets, would likely be corrected by a model that considered the structural context provided by temporal context inputs.

We therefore implemented the relaxation process, as explained above, using several different model types that were iteratively generated using Weka [84]. The results are shown in table 4.9.

While the relaxation process resulted in some very small accuracy increases for some model types, such as IBK and RandomForest, these accuracy increases were an order of magnitude less than what we had hoped; the relaxation process never succeeded in raising the Q_3 accuracy more than three tenths of one percent. Upon examining the predictions from the zeroth iteration (i.e., the iteration in which only letter attributes were used), we noted that both correct predictions and incorrect predictions tended to appear in sequences. Some clusters of consecutive instances in a protein chain would be correctly predicted to be alpha helices, for example, while other clusters of consecutive instances would be incorrectly predicted to be alpha helices when they were actually beta sheets. In hindsight, it seemed

reasonable that clusters of incorrect predictions would limit the usefulness of the relaxation process because the incorrect predictions would provide an incorrect context. As Minor demonstrated in 1996, a sequence of up to eleven amino acids can fold into an alpha helix or a beta sheet depending on context [6]. Hence, a classifier given incorrect context for an instance might actually be making a prediction that *would be* correct *if* that instance was actually surrounded by the predicted context rather than the true context.

We therefore decided to explore the possibility of whether Q_3 accuracy improvement could be achieved in a scheme that only provided a smaller number of context values—specifically, context values that could be predicted with a higher degree of confidence. We initially tried to build a prediction-confidence classifier that could predict whether or not a prediction was correct based on the confidence probabilities provided in Weka output buffers for some of the secondary-structure-prediction classifiers we had used. However, we quickly discovered that the prediction-confidence classifier was only able to identify when a secondary-structure-prediction classifier was making an error with about 60% accuracy. As a result, we decided to apply a different approach, as follows.

4.3.2 Collaborative Model Using Three High-Precision Classifiers

We generated three different cost-sensitive Logistic classifiers, each tuned to have very high precision for a single one of the three output classes (at the expense of recall). We then wrote a script that compared each high-precision classifier’s predictions for each instance in the test set (the RS126 data set). For each instance, if all three classifiers agreed, the consensus label was assigned as the predicted label for that instance. If the three classifiers disagreed, but only one classifier voted for its high-precision label, the high-precision label would be assigned as the predicted label for that instance. Any other instance on which the classifiers disagreed was assigned a label of unknown. Using these rules, 33.49% of the instances were assigned predicted labels, while the remaining instances were given unknown labels. We noted that the predicted labels were 79.75% accurate. Hence, at the very least, the approach with the

Training Set	Test Set	Classifier	
		Logistic	RandomForest
CB396 ($\approx 40\%$ unknown)	RS126 ($\approx 40\%$ unknown)	78.21%	79.68%
CB396 (none unknown)	RS126 ($\approx 40\%$ unknown)	75.23%	78.99%

Table 4.10: Q_3 Accuracies Achieved Using Training Sets having Different Percentages Temporal Context Features Unknown

three high-specificity models had succeeded in raising Q_3 accuracy on the 33% of instances whose labels were actually predicted.

The assigned labels were then used to generate an arff file that included the predicted output labels (including the unknown label, where applicable) of instances $n - 4$ through $n - 1$ and instances $n + 1$ through $n + 4$ as attributes for each instance n . In addition, each instance n had the original 13 amino-acid letter features. We were unsure of whether it would be best to train a model using a training file (CB396) wherein all context labels were known, since about 66% of the labels used as context attributes were unknowns. As a result, we decided to generate a training file with a large number of unknown values for context attributes in the following manner. First, we used a Logistic classifier in Weka using the standard 13-attribute CB396 file as both the training set and the test set. We then wrote a script that generated a new CB396 file with the temporal context attributes. Any instance that was incorrectly predicted was assigned a context label (i.e., for the purposes of the context attributes only) of unknown, while instances that were correctly predicted were assigned their true labels. This resulted in a training file wherein just under 40% of the context attributes had unknown values. We then used the same process to generate an RS126 file wherein about the same percentage of context labels were unknown. We then trained (1) a first set of logistic and RandomForest classifiers using the CB396 training set wherein the values for all context attributes were known and (2) a second set of classifiers using the CB396 training set wherein there were unknown values for some context attributes. Each classifier in each set was then run on the RS126 test set wherein there were unknown values for some context attributes. The results are shown in table 4.10.

Since both types of classifiers performed better when using the training set that had unknown values for some context attributes, we chose to use this training set to generate a model on the test set that had been created using the three high-precision classifiers (the final test set). We were still cautiously optimistic, since the results suggested that Q_3 accuracy could still be increased if a large percentage of context attributes had unknown values. However, to our disappointment, both a logistic model and a RandomForest model used on the final test set actually achieved *lower* Q_3 accuracies—57.22% and 60.49%, respectively. Hence, it appeared that the negative effect that incorrect context values caused may have been amplified when fewer context values were known, even when a larger percentage of known context values were correct. We considered trying to repeat the three-classifier approach using models with even higher precision. In making preparations to do so, we discovered that we had to push the recall for the N label all the way down to 7% to achieve precision of 92% using the cost-sensitive approach with a Logistic classifier. With our previous attempt, our efforts had achieved the best precision with the least impact on recall using the N label. Hence, if the N label’s precision and recall were to be considered upper bounds for the H and E labels, and if it would be necessary to push the precision for all labels up to 100%, we realized we would end up with so few known context labels that a good return would be unlikely.

4.3.3 Conclusions Regarding Temporal Context Attributes and Directions for Future Research

Ultimately, the approach of using predicted labels for context attributes and trying to relax them yielded only a very small amount of benefit. However, where true labels are known for at least some instances (about 60%, at least), it appears that Q_3 accuracy of nearly 80% is very achievable with fairly standard models. Relaxation and the collaborative three-model high-precision approach do not appear to be effective ways to discern those true labels, but other methods beyond the scope of this project might be. In particular, a good multiple-

sequence alignment could be helpful. Suppose, for example, a newly sequenced protein aligns well enough with some homologues whose structures are known. If all homologues have identical secondary structures at 60% of the amino-acid positions in the alignment, then the newly sequenced protein could be presumed to have those labels at those positions. These labels could then be used as input for a classifier that uses them for the context features that we have defined in this project.

Another observation worth noting is that it appears that certain regions in protein chains tend to have much more predictable secondary structures than others. The results of our collaborative three-model high-precision approach suggest that about 33% of the instances in RS126 can be predicted with about 80% Q_3 accuracy without using any information about amino-acid properties or multiple-sequence alignments. Other instances in RS126 are much more difficult to predict. There are many possible reasons why this might be the case. These difficult instances might, for example, represent regions that truly could fold into more than one secondary-structure conformation very easily—and there could even be a possible biological and evolutionary advantage to such a phenomenon. A gene that can be alternatively spliced, for example, might be better able to produce different proteins if certain regions are amenable to folding into both alpha helices and beta sheets. It would also be very interesting to explore whether multiple chaperone proteins that could all alternatively operate on the same peptide chain could fold it into proteins with similar *primary* structures, but different *secondary* and *tertiary* structures (and hence different functions). If this were the case, given n protein chains and k chaperone proteins, n new proteins could be produced simply by adding one new chaperone protein and k new proteins could be produced by adding one new protein chain. This might lead to better efficiency with evolution in that a single mutation could produce many new proteins. That being said, the presence of regions with flexible secondary structure could also sometimes simply be a random phenomenon of evolution.

Chapter 5

Conclusion

We spent a great deal of time and effort hoping to find a "holy grail" that would allow us to exceed the theoretical limit of 70% Q_3 accuracy posited by Qian and Sejnowski for single-sequence secondary-structure prediction methods (e.g., methods not using homology information). While our efforts did not ultimately result in the discovery of a "holy grail," we did ultimately make a number of contributions to the field, as explained below.

5.1 Contributions to the Field of Study

First, we have shown that a number of amino-acid properties that have not been used in previous studies can be used to improve single-sequence Q_3 prediction accuracy. While some previous studies have used isolated properties, such as hydrophobicity, we have conducted a thorough set of experiments exploring the relevance of amino-acid properties to secondary-structure prediction by creating thousands of models using over 500 different amino-acid properties. Our experiments demonstrate that classifiers trained using attributes derived from some of these properties we have identified can increase Q_3 accuracy by several percentage points compared to controls, depending on the classifier type that is used.

Second, we have devised a number of novel ways to derive attributes from properties that can aid in secondary-structure prediction. Attributes such as the inter-moment angle and the moments over instance sub-windows have not been used in previous research. However, when derived and used in the manner developed for this project, these novel attribute types can form part of an attribute set that enables classifiers of several different types to achieve

improved single-sequence Q_3 prediction accuracy versus controls. Third, we have developed a novel way to create diversity in a classifier set from which majority-vote ensembles for single-sequence secondary-structure with improved Q_3 prediction accuracy can be assembled. Our results suggest that at least some synergistic effect can be harnessed by including classifiers trained using attributes derived from different properties. Furthermore, our results also suggest that *overall* prediction accuracy—not just prediction accuracy for a single output class—can be improved by including some cost-sensitive classifiers that have been tuned to achieve (1) relatively even prediction accuracies for all classes, (2) increased prediction accuracy for two out of the three output classes, and (3) increased prediction accuracy for one output class. The diversity created from using cost-sensitive classifiers, when combined with diversity created by training classifiers using different feature sets and with diversity created by using classifiers constructed using different algorithms, can help raise Q_3 accuracy by about one percentage point in majority-vote ensembles of 3, 5, or 7.

Fourth, we have shown that the three-class secondary structure of an amino-acid in a protein can be predicted with near-perfect accuracy, even with very simple models, when the true labels of the four upstream predecessors and the four downstream successors are known and used as temporal context attributes. While this observation is not especially useful for predicting the structures of proteins that lack homologues of known structure, it is actually very useful for predicting the structures of proteins whose sequences vary from those of known homologues only at individual positions (e.g., proteins that have single-nucleotide polymorphisms (SNPs)). Furthermore, we have shown that nearly 80% Q_3 accuracy can be achieved when only about 60% of the temporal context attributes are known for a test set. This shows that high Q_3 accuracy can be achieved using models that are simpler than previous models that can achieve comparable Q_3 accuracy using homology information if 60% of the true amino-acid labels for protein can be ascertained (e.g., by using a multiple sequence alignment wherein all homologues share a consensus label at 60% of the positions in the protein). Thus, while we deliberately excluded homology information in our experiments,

we ironically made a pair of discoveries that are, in this respect, more relevant to models that incorporate homology information.

Fifth, we have shown that relaxing temporal context attributes used in the manner we have described can raise Q_3 accuracy by up to three tenths of a percent, depending on the model used, in single-sequence prediction methods. While this improvement is an order of magnitude smaller than what we had initially hoped, it is an improvement nonetheless.

5.2 Possible Directions for Future Work

In our experiments, we generated ARFF files using the same set of attribute types for each property. However, the results shown in Appendix table A from our second round of feature selection illustrate that some attribute types may be more relevant for certain properties than for others. Future work could seek to define which specific attribute types work best with certain properties with finer granularity. In addition, future work could also explore whether the same pairs of properties and attribute-types are best for all different types of classifiers. This may also help boost the prediction accuracy of some of the model types that were used. The RBFNetwork classifier that was ultimately used in both rounds of our ensemble experiments, for example, benefitted when a Weka filter (AttributeSelectedClassifier) was used to exclude consideration of certain attributes. In addition, further work could explore which properties can yield the most synergistic improvements in Q_3 prediction accuracy when used together. It would be interesting to determine whether properties that yield their best results with dissimilar attribute types are more likely to synergize well with each other.

It would also be interesting to explore the phenomenon how to best leverage cost-sensitive classifiers in a classifier set in order to achieve further improvements in Q_3 prediction accuracy of majority-vote ensembles. In our experiments, we used a brute-force approach and were thus only able to test relatively small ensembles. However, more efficient searches of the space of possible ensembles could likely be developed by using the pairwise Q statistic. Individual classifiers could be added to an ensemble in a greedy fashion, for example, based

on some metric that takes their pairwise Q value with each classifier that is already in the ensemble and based on their own individual prediction accuracy. Furthermore, perhaps a cost-sensitive classifier could be custom-tuned to match an existing ensemble's needs and added to the ensemble. These are only a few possibilities that could be explored.

Another important direction for future work is to investigate why several more complicated models, such as the MultilayerPerceptron models and the RandomForest models, did not achieve accuracy comparable to that of the simpler Logistic models. While we were thorough in terms of how many properties we investigated, we did not focus on optimizing model parameters (e.g., learning rate, momentum, number of nodes in each layer, and number of epochs for the MultilayerPerceptron and number of trees, maximum tree depth, pruning techniques, etc. for RandomForest) for individual model types. In theory, with optimal parameters and optimal feature sets, it should be possible to generate versions of the complicated models that perform at least as well as—and most likely better than—the simpler Logistic model.

References

- [1] Rhodri Saunders & Charlotte M. Deane, *Synonymous Codon Usage Influences the Local Protein Structure Observed*, 38 NUCLEIC ACIDS RESEARCH 6719 (2010).
- [2] John-March Chandonia & Steven E. Brenner, *The Impact of Structural Genomics: Expectations and Outcomes*, LAWRENCE BERKELEY NATIONAL LABORATORY (2005), available at <http://escholarship.org/uc/item/5v5659sk#page-1>.
- [3] *Current Prices*, U.C. Davis Genome Center (2014), available at <http://msf.ucdavis.edu/current-prices/>.
- [4] Zafer Aydin et al., *Protein Secondary Structure Prediction for a Single-Sequence Using Hidden Semi-Markov Models*, 7 BMC BIOINFORMATICS 178 (2006).
- [5] Helen M. Berman et al., *The Protein Data Bank*, 28 NUCLEIC ACIDS RESEARCH 235 (2000).
- [6] Daniel L. Minor and Peter S. Kim, *Context-Dependent Secondary Structure Formation of a Designed Protein Sequence*, 380 NATURE 730 (1996).
- [7] Mathura S. Venkatarajan & Werner Braun, *New Quantitative Descriptors of Amino Acids Based on Multidimensional Scaling of a Large Number of Physical-Chemical Properties*, 7 JOURNAL OF MOLECULAR MODELING 445 (2001).
- [8] Ross D. King & Michael J.E. Sternberg, *Identification and Application of the Concepts Important for Accurate and Reliable Protein Secondary Structure Prediction*, 5 PROTEIN SCIENCE 2298 (1996).

- [9] Kuo-Chen Chou et al., *Disposition of Amphiphilic Helices in Heteropolar Environments*, 28 PROTEINS 99 (1997).
- [10] Kuo-Chen Chou et al., *Wenxiang: A Web Server for Drawing Wenxiang Diagrams*, 3 NATURAL SCIENCE 862 (2011).
- [11] David Eisenberg, *The Discovery of the α -helix and β -sheet, the Principal Structural Features of Proteins*, 100 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCE 11207 (2003).
- [12] Robbie P. Joosten et al., *A Series of PDB Related Databases for Everyday Needs*, 39 NUCLEIC ACIDS RESEARCH D411 (2011).
- [13] Wolfgang Kabsch & Christian Sander, *Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features*, 22 BIOPOLYMERS 2577 (1983).
- [14] Burkhard Rost & Chris Sander, *Prediction of Protein Structure at Better Than 70% Accuracy*, 232 JOURNAL OF MOLECULAR BIOLOGY 584 (1993).
- [15] Yong-Sheng Ding et al., *Using Maximum Entropy Model to Predict Secondary Structure with Single Sequence*, 16 PROTEIN & PEPTIDE LETTERS 552 (2009).
- [16] Linus Pauling et al., *The Structure of Proteins: Two Hydrogen-bonded Helical Configurations of the Polypeptide Chain*, 37 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 205 (1951).
- [17]] C.H. Bamford, *Molecular Configuration and Physical Properties of Polypeptides and Proteins*, 44 PROCEEDINGS OF THE ROYAL SOCIETY OF MEDICINE 393 (1951).
- [18] John T. Edsall, *Configuration of Certain Protein Molecules: An Inquiry Concerning the Present Status of our Knowledge*, 12 JOURNAL OF POLYMER SCIENCE 253 (1954).
- [19] Andrew G. Szent-Györgyi & Carolyn Cohen, *Role of Proline in Polypeptide Chain Configuration of Proteins*, 126 SCIENCE 697 (1957).

- [20] John C. Kendrew et al., *A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis*, 181 NATURE 662 (1958).
- [21] F. B. Straub, *Formation of the Secondary and Tertiary Structure of Enzymes*, 236 ADVANCES IN ENZYMOLOGY AND RELATED AREAS OF MOLECULAR BIOLOGY 89 (1964).
- [22] Christian B. Anfinsen et al., *The Kinetics of Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain*, 47 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 1309 (1961).
- [23] Iva Simon, *Investigation of Protein Refolding: A Special Feature of Refolding Responsible for Refolding Ability*, 113 JOURNAL OF THEORETICAL BIOLOGY 703 (1985).
- [24] Anthony V. Guzzo, *The Influence of Amino-Acid Sequence on Protein Structure*, 5 BIOPHYSICAL JOURNAL 809 (1965).
- [25] John W. Prothero, *Correlation between the Distribution of Amino Acids and Alpha Helices*, 6 BIOPHYSICAL JOURNAL 367 (1966).
- [26] P.F. Periti et al., *Recognition of α -Helical Segments in Proteins of Known Primary Structure*, 24 JOURNAL OF MOLECULAR BIOLOGY 313 (1967).
- [27] Oleg B. Ptitsyn, *Statistical Analysis of the Distribution of Amino Acid Residues among Helical and Non-Helical Regions in Globular Proteins*, 42 JOURNAL OF MOLECULAR BIOLOGY 501 (1969).
- [28] Marianne Schiffer & Allen B. Edmundson, *Use of Helical Wheels to Represent the Structures of Proteins and Identify Segments with Helical Potential*, 7 BIOPHYSICAL JOURNAL 121 (1967).
- [29] Cyrus Levinthal, *How to Fold Graciously*, in MÖSSBAUN SPECTROSCOPY IN BIOLOGICAL SYSTEMS PROCEEDINGS, 67 UNIVERSITY OF ILLINOIS BUL-

- LETIN 22 (1969), available at http://www.cc.gatech.edu/~turk/bio_sim/articles/proteins_levinthal_1969.pdf.
- [30] Christian B. Anfinsen, *Principles that Govern the Folding of Protein Chains*, 181 SCIENCE 223 (1973).
- [31] A. Keith Dunker et al., *Intrinsically Disordered Protein*, 19 JOURNAL OF MOLECULAR GRAPHICS AND MODELING 26 (2001).
- [32] Barry Robson & Roger H. Pain, *Analysis of the Code Relating Sequence to Conformation in Proteins: Possible Implications for the Mechanism of Formation of Helical Regions*, 58 JOURNAL OF MOLECULAR BIOLOGY 237 (1971).
- [33]] Helen M. Berman, *The Protein Data Bank: A Historical Perspective*, 64 ACTA CRYSTALLOGRAPHICA SECTION A: FOUNDATIONS OF CRYSTALLOGRAPHY 88 (2007).
- [34] Kozo Nagano, *Logical Analysis of the Mechanism of Protein Folding*, 75 JOURNAL OF MOLECULAR BIOLOGY 401 (1973).
- [35] J. Garnier et al., *Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins*, 120 JOURNAL OF MOLECULAR BIOLOGY 97 (1978).
- [36] Peter Y. Chou & Gerald D. Fasman, *Conformational Parameters for Amino Acids in Helical, β -Sheet, and Random Coil Regions Calculated from Proteins*, 13 BIOCHEMISTRY 211 (1974).
- [37] Valery I. Lim, *Structural Principles of the Globular Organization of Protein Chains. A Stereochemical Theory of Globular Protein Secondary Structure*, 88 JOURNAL OF MOLECULAR BIOLOGY 857 (1974).
- [38] Oleg B. Ptitsyn & Alexey V. Finkelstein, *Theory of Protein Secondary Structure and Algorithm of its Prediction*, 22 BIOPOLYMERS 15 (1983).

- [39] Wolfgang Kabsch & Christian Sander, *How Good are Predictions of Protein Secondary Structure?*, 155 FEBS LETTERS 179 (1983).
- [40] Fred E. Cohen et al., *Secondary Structure Assignment for α/β Proteins by a Combinatorial Approach*, 22 BIOCHEMISTRY 4894 (1983).
- [41] Marketa J. Zvelebil et al., *Prediction of Protein Secondary Structure and Active Sites Using the Alignment of Homologous Sequences*, 195 JOURNAL OF MOLECULAR BIOLOGY 957 (1987).
- [42] Irving P. Crawford et al., *Prediction of Secondary Structure by Evolutionary Comparison: Application to the Alpha Subunit of Tryptophan Synthase*, 2 PROTEINS 118 (1987).
- [43] Geoffrey J. Barton et al., *Amino Acid Sequence Analysis of the Annexin Super-gene Family of Proteins*, 198 EUROPEAN JOURNAL OF BIOCHEMISTRY 749 (1991).
- [44] Robert B. Russell, *Conservation Analysis and Structure Prediction of the SH2 Family of Phosphotyrosine Binding Domains*, 304 FEBS LETTERS 15 (1992).
- [45] Frederick R. Maxfield & Harold A. Scheraga, *Improvements in the Prediction of Protein Backbone Topography by Reduction of Statistical Errors*, 18 BIOCHEMISTRY 697 (1979).
- [46] Hans-Herbert Bohr et al., *Protein Secondary Structure and Homology by Neural Networks*, 241 FEBS LETTERS 223 (1988).
- [47] Ning Qian & Terrence J. Sejnowski, *Predicting the Secondary Structure of Globular Proteins Using Neural Network Models*, 202 JOURNAL OF MOLECULAR BIOLOGY 865 (1988).
- [48] L. Howard Holley & Martin Karplus, *Protein Secondary Structure Prediction with a Neural Network*, 86 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 152 (1989).

- [49] Stefano Pascarella & Francesco Bossa, *PRONET: A Microcomputer Program for Predicting the Secondary Structure of Proteins with a Neural Network*, 5 COMPUTER APPLICATIONS IN THE BIOSCIENCES 319 (1990).
- [50] D.G. Kneller et al., *Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network*, 214 JOURNAL OF MOLECULAR BIOLOGY 171 (1990).
- [51] Paul Stolorz et al., *Predicting Protein Secondary Structure Using Neural Net and Statistical Methods*, 225 JOURNAL OF MOLECULAR BIOLOGY 363 (1992).
- [52] Steven M. Muskal & Sung-Hou Kim, *Predicting Secondary Structure Content: A Tandem Neural Network Approach*, 225 JOURNAL OF MOLECULAR BIOLOGY 713 (1992).
- [53] Sujun Hua & Zhirong Sun, *A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach*, 308 JOURNAL OF MOLECULAR BIOLOGY 397 (2001).
- [54] Chao Chen et al., *Prediction of Protein Secondary Structure Content by Using the Concept of Chou's Pseudo Amino Acid Composition and Support Vector Machine*, 16 PROTEIN AND PEPTIDE LETTERS 27 (2009).
- [55] Gianluca Pollastri et al., *Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles*, 47 PROTEINS 228 (2002).
- [56] Joachim Selbig et al., *Decision Tree-Based Formation of Consensus Protein Secondary Structure Prediction*, 15 BIOINFORMATICS 1039 (1999).
- [57] Xin-Qiu Yao et al., *A Dynamic Bayesian Network Approach to Protein Secondary Structure Prediction*, 9 BMC BIOINFORMATICS 49 (2008).
- [58] Tau-Mu Yi & Eric Lander, *Protein Secondary Structure Prediction Using Nearest-Neighbor Methods*, 232 JOURNAL OF MOLECULAR BIOLOGY 1117 (1993).

- [59] Asaf Salamov & Victor Solovyev, *Prediction of Protein Secondary Structure by Combining Nearest-Neighbor Algorithms and Multiple-Sequence Alignments*, 247 JOURNAL OF MOLECULAR BIOLOGY 11 (1995).
- [60] Kiyoshi Asai et al., *Prediction of the Protein Secondary Structure by the Hidden Markov Model*, 9 COMPUTER APPLICATIONS IN THE BIOSCIENCES 141 (1993).
- [61] James A. Cuff & Geoffrey J. Barton, *Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction*, 34 PROTEINS 508 (1999).
- [62] Igor Berezovsky & Edward Trifonov, *Loop Fold Structure of Proteins: Resolution of Levinthal's Paradox*, 20 JOURNAL OF BIOMOLECULAR STRUCTURES 5 (2002).
- [63] Christophe N. Magnan & Pierre Baldi, *SSpro/ACCpro 5: Almost Perfect Prediction of Protein Secondary Structure and Relative Solvent Accessibility Using Profiles, Machine Learning and Structural Similarity*, 30 BIOINFORMATICS 2592 (2014).
- [64] Walter Pirovano & Jaap Heringa, *Protein Secondary Structure Prediction*, in DATA MINING FOR THE LIFE SCIENCES, 609 METHODS IN MOLECULAR BIOLOGY 327 (2010).
- [65] C. Nick Pace & J. Martin Scholtz, *A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins*, 75 BIOPHYSICAL JOURNAL 422 (1998).
- [66] Jayavardhana Gubbi et al., *Protein Secondary Structure Prediction Using Support Vector Machines and a New Feature Representation*, 6 INTERNATIONAL JOURNAL OF COMPUTATIONAL INTELLIGENCE AND APPLICATIONS 551 (2006).
- [67] Cong Z. Cai et al., *SVM-Prot: Web-based Support Vector Machine Software for Functional Classification of a Protein from its Primary Sequence*, 31 NUCLEIC ACIDS RESEARCH 3692 (2003).

- [68] Serene Ong et al., *Efficacy of Different Protein Descriptors in Predicting Protein Functional Families*, 8 BMC BIOINFORMATICS 300 (2007).
- [69] Thomas Nordahl Petersen et al., *Prediction of Protein Secondary Structure at 80% Accuracy*, 41 PROTEINS: STRUCTURE, FUNCTION, AND GENETICS 17 (2000).
- [70] Claus Lundegaard et al., *Prediction of Protein Secondary Structure at High Accuracy Using a Combination of Many Neural Networks*, in MATHEMATICAL METHODS FOR PROTEIN STRUCTURE ANALYSIS AND DESIGN 117, Springer Berlin Heidelberg (2003).
- [71] Minh H. Nguyen & Jagath C. Rajapakse, *Two-stage Multi-Class Support Vector Machines to Protein Secondary Structure Prediction*, GENOME INFORMATICS SERIES 218 (2003).
- [72] Pierre Baldi et al., *Exploiting the Past and the Future in Protein Secondary Structure Prediction*, 15 BIOINFORMATICS 937 (1999).
- [73] Gianluca Pollastri & Aoife McLysaght, *Porter: A New, Accurate Server for Protein Secondary Structure Prediction*, 21 BIOINFORMATICS 1719 (2004).
- [74] Frances C. Bernstein et al., *The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures*, 112 JOURNAL OF MOLECULAR BIOLOGY 535 (1977).
- [75] Moshe Lichman, UCI MACHINE LEARNING REPOSITORY [<http://archive.ics.uci.edu/ml>], Irvine, CA, University of California, School of Information and Computer Science (2013).
- [76] Peer Bork et al., *What's in a Genome?*, 358 NATURE (LONDON) 287 (1992).
- [77] Christine A. Orengo et al., *CATH—a Hierarchic Classification of Protein Domain Structures*, 5 STRUCTURE 1093 (1997).
- [78] Burkhard Rost & Volker A. Eyrich, *EVA: Large-Scale Analysis of Secondary Structure Prediction*, Suppl 5 PROTEINS 192 (2001).

- [79] Bent Petersen et al., *NetTurnp - Neural Network Prediction of Beta-turns by Use of Evolutionary Information and Predicted Protein Sequence Features*, 5 PLOS ONE e15079 (2010).
- [80] K. Guruprasad & S. Rajkumar, *Beta-and Gamma-turns in Proteins Revisited: A New Set of Amino Acid Turn-type Dependent Positional Preferences and Potentials*, 25 JOURNAL OF BIOSCIENCES 143 (2000).
- [81] Patrick Fuchs & Alain Alix, *High Accuracy Prediction of β -Turns and Their Types Using Propensities and Multiple Alignments*, 59 PROTEINS: STRUCTURE, FUNCTION, AND BIOINFORMATICS 828 (2005).
- [82] Jianlin Cheng et al., *Large-Scale Prediction of Disulphide Bridges Using Kernel Methods, Two-Dimensional Recursive Neural Networks, and Weighted Graph Matching*, 62 PROTEINS 617 (2006).
- [83] Shuichi Kawashima et al., *AAindex: Amino Acid Index Database, Progress Report 2008*, 36 NUCLEIC ACIDS RESEARCH D202 (2008) (index available at http://www.genome.jp/aaindex/AAindex/list_of_indices).
- [84] Mark Hall et al., *The WEKA Data Mining Software: An Update*, 11 SIGKDD EXPLORATIONS 10, (2009).
- [85] Lior Rokach, *Ensemble-based Classifiers*, 33 ARTIFICIAL INTELLIGENCE REVIEW 1 (2010).
- [86]] Ludmila Kuncheva & Christopher Whitaker, *Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy*, 51 MACHINE LEARNING 181 (2003).
- [87] <https://groups.google.com/forum/#!topic/predictprotein/73V-DNZQqhI>

[88] Kuo-Chen Chou, *Using Pair-coupled Amino Acid Composition to Predict Protein Secondary Structure Content*, 18 JOURNAL OF PROTEIN CHEMISTRY 473 (1999).

Appendix A

(Table A)

Letter for amino acid #1
Letter for amino acid #2
Letter for amino acid #3
Letter for amino acid #4
Letter for amino acid #5
Relative preference value at C-cap (Richardson-Richardson, 1988) for amino acid #5
Letter for amino acid #6
Helix initiation parameter at position $i,i+1,i+2$ (Finkelstein et al., 1991) for amino acid #6
Short and medium range non-bonded energy per residue (Oobatake-Ooi, 1977) for amino acid #6
Weights for coil at the window position of 1 (Qian-Sejnowski, 1988) for amino acid #6
A parameter of charge transfer capability (Charton-Charton, 1983) for amino acid #7
Conformational parameter of beta-structure (Beghin-Dirkx, 1975) for amino acid #7
Conformational parameter of beta-turn (Beghin-Dirkx, 1975) for amino acid #7
Conformational parameter of inner helix (Beghin-Dirkx, 1975) for amino acid #7
Helix initiation parameter at position $i,i+1,i+2$ (Finkelstein et al., 1991) for amino acid #7
Molecular weight (Fasman, 1976) for amino acid #7
Relative population of conformational state E (Vasquez et al., 1983) for amino acid #7
Weights for coil at the window position of 2 (Qian-Sejnowski, 1988) for amino acid #7
Delta G values for the peptides extrapolated to 0 M urea (O*Neil-DeGrado, 1990) for amino acid #8
Helix initiation parameter at position $i,i+1,i+2$ (Finkelstein et al., 1991) for amino acid #8
Helix-coil equilibrium constant (Ptitsyn-Finkelstein, 1983) for amino acid #8
N.m.r. chemical shift of alpha-carbon (Fauchere et al., 1988) for amino acid #8
Relative population of conformational state E (Vasquez et al., 1983) for amino acid #8
Weights for coil at the window position of -3 (Qian-Sejnowski, 1988) for amino acid #8
Letter for amino acid #9
Weights for beta-sheet at the window position of 4 (Qian-Sejnowski, 1988) for amino acid #9
Weights for coil at the window position of 1 (Qian-Sejnowski, 1988) for amino acid #9
Letter for amino acid #10
Weights for beta-sheet at the window position of 4 (Qian-Sejnowski, 1988) for amino acid #10
Letter for amino acid #11
Delta G values for the peptides extrapolated to 0 M urea (O*Neil-DeGrado, 1990) for amino acid #12
Relative preference value at N5 (Richardson-Richardson, 1988) for amino acid #13
Magnitude of negative moment for 14 A contact number (Nishikawa-Ooi, 1986)
Magnitude of positive moment for 8 A contact number (Nishikawa-Ooi, 1980)
Magnitude of negative moment for A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet (Charton-Charton, 1983)
Magnitude of positive moment for ALTLS index (Cornette et al., 1987)
Magnitude of positive moment for Apparent partition energies calculated from Wertz-Scheraga index (Guy, 1985)
Magnitude of negative moment for Apparent partition energies calculated from Wertz-Scheraga index (Guy, 1985)

Table A.1: (Page 1 of 3) These are the 78 attributes that were selected when the second approach to feature selection was used.

Magnitude of positive moment for HPLC parameter (Parker et al., 1986)
Magnitude of negative moment for HPLC parameter (Parker et al., 1986)
Magnitude of positive moment for Helix formation parameters (delta delta G) (O*Neil-DeGrado, 1990)
Magnitude of positive moment for Hydration free energy (Robson-Osguthorpe, 1979)
Magnitude of negative moment for Hydration free energy (Robson-Osguthorpe, 1979)
Magnitude of positive moment for Hydration potential (Wolfenden et al., 1981)
Magnitude of negative moment for Mean polarity (Radzicka-Wolfenden, 1988)
Magnitude of negative moment for Modified Kyte-Doolittle hydrophobicity scale (Juretic et al., 1998)
Magnitude of negative moment for Normalized average hydrophobicity scales (Cid et al., 1992)
Magnitude of positive moment for Normalized composition from animal (Nakashima et al., 1990)
Magnitude of positive moment for Normalized composition from fungi and plant (Nakashima et al., 1990)
Magnitude of negative moment for Normalized composition from fungi and plant (Nakashima et al., 1990)
Angle between moments for Normalized composition from fungi and plant (Nakashima et al., 1990)
Magnitude of negative moment for Normalized composition of membrane proteins (Nakashima et al., 1990)
Angle between moments for Normalized composition of membrane proteins (Nakashima et al., 1990)
Magnitude of positive moment for Normalized composition of mt-proteins (Nakashima et al., 1990)
Magnitude of negative moment for Normalized composition of mt-proteins (Nakashima et al., 1990)
Magnitude of positive moment for Optimal matching hydrophobicity (Sweet-Eisenberg, 1983)
Magnitude of negative moment for Optimal matching hydrophobicity (Sweet-Eisenberg, 1983)
Angle between moments for Optimal matching hydrophobicity (Sweet-Eisenberg, 1983)
Magnitude of positive moment for Principal property value z1 (Wold et al., 1987)
Magnitude of negative moment for Principal property value z1 (Wold et al., 1987)
Magnitude of positive moment for Relative partition energies derived by the Bethe approximation (Miyazawa-Jernigan, 1999)
Magnitude of negative moment for Relative partition energies derived by the Bethe approximation (Miyazawa-Jernigan, 1999)
Magnitude of positive moment for Retention coefficient in HFBA (Browne et al., 1982)
Magnitude of positive moment for Side chain hydrophathy, uncorrected for solvation (Roseman, 1988)
Angle between moments for Side chain hydrophathy, uncorrected for solvation (Roseman, 1988)
Magnitude of positive moment for Solvation free energy (Eisenberg-McLachlan, 1986)
Magnitude of positive moment for TOTFT index (Cornette et al., 1987)

Table A.2: (Page 2 of 3) These are the 78 attributes that were selected when the second approach to feature selection was used.

Magnitude of positive moment for Weights for beta-sheet at the window position of 1 (Qian-Sejnowski, 1988)
Angle between moments for Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988)
Magnitude of positive moment for Weights for beta-sheet at the window position of 4 (Qian-Sejnowski, 1988)
Magnitude of negative moment for Weights for coil at the window position of -2 (Qian-Sejnowski, 1988)
Magnitude of positive moment for Weights for coil at the window position of -3 (Qian-Sejnowski, 1988)
Magnitude of negative moment for Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)
Magnitude of positive moment for Weights for coil at the window position of 1 (Qian-Sejnowski, 1988)
Magnitude of negative moment for Weights for coil at the window position of 2 (Qian-Sejnowski, 1988)
Magnitude of negative moment for Weights for coil at the window position of 3 (Qian-Sejnowski, 1988)
Magnitude of positive moment for Weights from the IFH scale (Jacobs-White, 1989)
Angle between moments for Weights from the IFH scale (Jacobs-White, 1989)

Table A.3: (Page 3 of 3) These are the 78 attributes that were selected when the second approach to feature selection was used.

Appendix B

(Table B)

Attribute	Possible values as specified in WEKA
Letter for amino acid #1	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #1	NUMERIC
Letter for amino acid #2	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #2	NUMERIC
Letter for amino acid #3	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #3	NUMERIC
Letter for amino acid #4	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #4	NUMERIC
Letter for amino acid #5	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #5	NUMERIC
Letter for amino acid #6	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #6	NUMERIC
Letter for amino acid #7	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #7	NUMERIC
Letter for amino acid #8	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #8	NUMERIC
Letter for amino acid #9	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #9	NUMERIC
Letter for amino acid #10	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #10	NUMERIC
Letter for amino acid #11	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #11	NUMERIC
Letter for amino acid #12	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #12	NUMERIC
Letter for amino acid #13	{A,R,D,N,C,E,O,G,H,I,L,K,M,F,P,Q,S,T,W,Y,V,X,Z,*}
<property name> for amino acid #13	NUMERIC
Magnitude of positive alpha-helical moment for <property name>	NUMERIC
Magnitude of negative alpha-helical moment for <property name>	NUMERIC
Angle between positive and negative alpha-helical moments for <property name>	NUMERIC
Magnitude of total alpha-helical moment for <property name>	NUMERIC
Magnitude of total beta-sheet moment for <property name>	NUMERIC
Magnitude of total alpha-helical moment for <property name> over subwindow 1	NUMERIC
Magnitude of total alpha-helical moment for <property name> over subwindow 2	NUMERIC
Magnitude of total alpha-helical moment for <property name> over subwindow 3	NUMERIC
Magnitude of total alpha-helical moment for <property name> over subwindow 4	NUMERIC
Magnitude of total alpha-helical moment for <property name> over subwindow 5	NUMERIC
Magnitude of total alpha-helical moment for <property name> over subwindow 6	NUMERIC
Magnitude of total beta-sheet moment for <property name> over subwindow 1	NUMERIC
Magnitude of total beta-sheet moment for <property name> over subwindow 2	NUMERIC
Magnitude of total beta-sheet moment for <property name> over subwindow 3	NUMERIC
Magnitude of total beta-sheet moment for <property name> over subwindow 4	NUMERIC
Magnitude of total beta-sheet moment for <property name> over subwindow 5	NUMERIC
Magnitude of total beta-sheet moment for <property name> over subwindow 6	NUMERIC
Secondary Structure Label for middle amino acid	{H,E,N}

Table B.1: (Page 1 of 1) These are the 44 attributes that were used in each individual-property arff file.

Appendix C

(Table C)

Property	Index of Property	logistic	Randomforest 225, 5	DTNB	BK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2, 100, 3
Letters only (Control)	n/a	62.5021	60.9603	60.4232	59.9065	61.2514	59.2833	61.2473	60.7143	53.7395
14 A contact number (Nishikawa-Ooi, 1986)	0	65.4215	62.8875	63.5312	63.3877	61.6287	59.1315	58.9183	59.1151	58.8486
8 A contact number (Nishikawa-Ooi, 1980)	1	65.196	62.5472	63.1212	62.5185	60.7225	58.5575	57.6636	58.1351	56.8394
A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet (Charton-Charton, 1983)	2	62.3462	59.8737	61.518	59.2135	59.4104	51.6689	56.2121	54.6621	54.6908
A parameter of charge transfer capability (Charton-Charton, 1983)	3	62.6579	60.5257	60.9562	56.4499	55.8471	54.3259	58.8568	54.7072	54.0061
A parameter of charge transfer donor capability (Charton-Charton, 1983)	4	62.4938	59.8778	61.1202	54.7605	59.7712	50.4592	60.6446	56.6221	53.9159
AA composition of CYT of multi-spanning proteins (Nakashima-Nishikawa, 1992)	5	62.4897	59.9229	61.6656	59.312	60.6569	55.2936	59.3366	56.1506	54.0963
AA composition of CYT of single-spanning proteins (Nakashima-Nishikawa, 1992)	6	62.3667	60.2673	61.5262	59.7261	60.8045	54.6252	58.8773	56.3064	54.0758
AA composition of CYT2 of single-spanning proteins (Nakashima-Nishikawa, 1992)	7	62.4938	61.0218	61.3621	59.3817	60.4273	56.1383	57.8809	56.2695	55.6503
AA composition of EXT of multi-spanning proteins (Nakashima-Nishikawa, 1992)	8	62.4979	60.2755	61.7763	58.9634	60.538	55.8512	59.9475	53.4525	53.8462
AA composition of EXT of single-spanning proteins (Nakashima-Nishikawa, 1992)	9	62.4405	60.2837	61.4606	59.2874	60.8537	56.5032	60.6118	55.3674	54.0471
AA composition of EXT2 of single-spanning proteins (Nakashima-Nishikawa, 1992)	10	62.4856	60.1566	61.5508	59.7056	60.6897	56.2367	58.3607	54.371	53.9036
AA composition of MEM of multi-spanning proteins (Nakashima-Nishikawa, 1992)	11	64.4374	62.0264	62.744	61.2596	61.2965	58.3648	58.4714	57.9916	57.0691
AA composition of MEM of single-spanning proteins (Nakashima-Nishikawa, 1992)	12	63.9823	61.6779	61.9362	60.4314	60.7225	57.6718	57.3233	56.6549	56.0481
AA composition of membrane proteins (Nakashima et al., 1990)	13	62.8711	60.3206	61.5836	59.5703	60.497	56.7287	58.2869	55.8512	54.1988
AA composition of mt-proteins (Nakashima et al., 1990)	14	63.2114	60.9111	61.6574	59.6236	59.722	55.1214	56.0932	53.8749	54.2849
AA composition of mt-proteins from animal (Nakashima et al., 1990)	15	63.2114	60.866	61.8173	59.4063	59.4186	56.044	56.0112	54.2193	54.2152
AA composition of mt-proteins from fungi and plant (Nakashima et al., 1990)	16	63.0269	60.5503	61.7189	60.009	60.2181	55.2526	56.7451	55.1829	54.8425
AA composition of total proteins (Nakashima et al., 1990)	17	62.539	60.1197	61.3498	59.3202	60.3781	53.7395	58.8363	55.7815	53.5632
ALTFT index (Cornette et al., 1987)	18	65.0443	62.58	63.4492	63.0556	61.969	59.0823	59.0864	58.742	57.8235
ALTLS index (Cornette et al., 1987)	19	64.9213	62.3175	63.6994	62.8424	61.6451	59.3448	59.1356	58.7953	58.7748
Absolute entropy (Hutchens, 1970)	20	62.5677	60.5503	61.5303	59.1233	60.8701	58.3156	60.2222	55.2936	56.1793

Table C.1: (Page 1 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Accessibility reduction ratio (Ponnuswamy et al., 1980)	21	64.7942	62.0961	62.9121	62.3011	62.3134	57.762	59.7384	58.1721	57.3233
Accessible surface area (Radzicka-Wolfenden, 1988)	22	62.4774	60.5421	61.4401	59.4637	59.2751	56.8558	56.1588	55.7036	56.5975
Activation Gibbs energy of unfolding, pH7.0 (Yutani et al., 1987)	23	62.5677	60.8947	61.7517	58.3976	58.2295	53.3869	51.1399	52.6119	54.4038
Activation Gibbs energy of unfolding, pH9.0 (Yutani et al., 1987)	24	62.5759	60.8988	61.6041	58.5165	58.537	53.3705	51.1563	52.001	55.9948
Alpha helix propensity of position 44 in T4 lysozyme (Blaber et al., 1993)	25	62.3954	60.4437	61.4319	59.0946	54.371	58.1392	47.8145	49.4711	56.454
Alpha-helix indices (Geisow-Roberts, 1980)	26	62.5677	60.5339	61.2719	59.3366	60.1115	57.6472	58.7543	55.9907	56.0522
Alpha-helix indices for alpha-proteins (Geisow-Roberts, 1980)	27	62.4692	60.5831	61.3539	59.4801	58.3484	56.1301	55.5478	54.74	56.003
Alpha-helix indices for alpha/beta-proteins (Geisow-Roberts, 1980)	28	62.5103	60.8168	61.6123	59.6318	60.5257	58.0367	58.3115	56.6139	56.1178
Alpha-helix indices for beta-proteins (Geisow-Roberts, 1980)	29	62.5226	60.0213	61.5508	59.1602	60.7389	55.9866	58.9347	55.7897	54.002
Alpha-helix propensity derived from designed sequences (Koehl-Levitt, 1999)	30	62.5021	60.3001	61.4852	58.1474	57.8891	54.3833	50.8201	53.2106	56.3064
Amino acid composition (Dayhoff et al., 1978a)	31	62.5431	60.1197	61.4647	59.0167	60.5954	56.0891	60.2181	55.4617	53.678
Amino acid distribution (Jukes et al., 1975)	32	62.6251	60.2222	61.5713	58.8896	60.661	54.1988	59.9106	55.0189	53.6493
Amphiphilicity index (Mitaku et al., 2002)	33	62.621	60.1033	61.3498	58.5534	54.6457	54.3054	57.5324	54.1332	53.7887
Aperiodic indices (Geisow-Roberts, 1980)	34	63.154	60.9808	61.641	60.8414	60.2181	60.7512	58.8773	57.1921	57.9137
Aperiodic indices for alpha-proteins (Geisow-Roberts, 1980)	35	62.4323	60.7594	61.3047	59.3817	59.6031	56.1342	56.946	54.8097	55.9784
Aperiodic indices for alpha/beta-proteins (Geisow-Roberts, 1980)	36	63.2278	61.2883	61.6205	61.2473	60.2345	61.071	59.0085	57.9506	58.0121
Aperiodic indices for beta-proteins (Geisow-Roberts, 1980)	37	63.4984	61.436	62.0592	61.4688	60.9931	59.1397	59.8573	56.3474	57.2372
Apparent partial specific volume (Bull-Breese, 1974)	38	63.1048	61.1202	62.1781	60.9931	61.1079	57.3438	59.3612	57.2905	55.6626
Apparent partition energies calculated from Chothia index (Guy, 1985)	39	64.4497	61.8337	62.6825	61.1366	61.1817	57.4873	57.6841	57.6062	56.8353
Apparent partition energies calculated from Janin index (Guy, 1985)	40	64.503	62.0797	62.6251	60.6692	60.0787	58.619	56.7533	56.3515	56.6713
Apparent partition energies calculated from Robson-Osguthorpe index (Guy, 1985)	41	64.4538	62.4569	62.8916	58.2172	61.2145	57.0978	57.5693	57.6882	56.1383
Apparent partition energies calculated from Wertz-Scheraga index (Guy, 1985)	42	64.8721	62.2888	62.6907	62.5226	60.0008	58.3853	58.0572	58.6559	57.1921
Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)	43	62.7645	60.2632	61.5713	58.4919	58.2787	57.721	54.3259	53.92	54.4284
Average accessible surface area (Janin et al., 1978)	44	63.888	61.2227	62.174	60.4437	61.0136	58.4796	58.0572	56.1547	55.9291

Table C.2: (Page 2 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)	45	63.0679	61.2145	62.0674	61.3621	60.9316	57.0649	60.1812	57.3971	56.3064
Average gain in surrounding hydrophobicity (Ponnuswamy et al., 1980)	46	64.7409	62.4856	63.564	62.7153	62.4036	59.1151	61.2309	57.9014	57.7579
Average gain ratio in surrounding hydrophobicity (Ponnuswamy et al., 1980)	47	64.6629	62.3544	62.7153	62.6702	61.9936	60.0131	61.0792	58.6395	58.0449
Average interactions per side chain atom (Warne-Morgan, 1978)	48	63.6502	61.7927	62.379	60.907	61.2268	58.5657	60.1771	56.3474	56.2203
Average internal preferences (Olsen, 1980)	49	64.421	62.1289	63.0187	61.6	61.8624	58.1639	58.3853	57.3192	57.2905
Average membrane preference: AMP07 (Degli Esposti et al., 1990)	50	64.5727	62.3298	62.7071	62.1986	61.6369	59.599	58.7379	58.3607	57.2413
Average non-bonded energy per atom (Oobatake-Ooi, 1977)	51	63.9905	61.6451	62.8793	61.4113	61.2063	58.7092	58.865	58.3443	56.0071
Average non-bonded energy per residue (Oobatake-Ooi, 1977)	52	62.7604	60.9152	61.6	59.8573	60.9029	57.5734	60.3494	55.4043	55.7487
Average number of surrounding residues (Ponnuswamy et al., 1980)	53	64.7819	62.2437	62.9408	62.379	62.5431	60.1402	61.2514	58.2049	57.3356
Average reduced distance for C-alpha (Meirovitch et al., 1980)	54	63.9126	62.1986	62.8055	62.7153	61.5016	60.3412	60.5667	58.9962	57.2495
Average reduced distance for C-alpha (Rackovsky-Scheraga, 1977)	55	64.0725	62.0797	62.7153	62.8219	61.2924	58.988	60.7963	59.4063	57.4176
Average reduced distance for side chain (Meirovitch et al., 1980)	56	64.4169	62.1166	62.5021	62.6251	62.42	60.8455	62.0715	59.4432	57.6472
Average reduced distance for side chain (Rackovsky-Scheraga, 1977)	57	64.421	61.7763	63.0269	62.8424	62.6989	61.6369	61.969	59.4924	57.4586
Average relative fractional occurrence in A0(i) (Rackovsky-Scheraga, 1982)	58	62.8752	60.8578	61.7476	60.3083	60.6897	56.2203	59.7138	55.638	54.8097
Average relative fractional occurrence in A0(i-1) (Rackovsky-Scheraga, 1982)	59	62.4815	60.0746	61.7681	59.5334	60.7389	55.5109	57.4955	54.9287	54.6293
Average relative fractional occurrence in AL(i) (Rackovsky-Scheraga, 1982)	60	62.9408	60.4601	61.5098	59.5908	59.9557	56.3638	59.07	56.9009	54.002
Average relative fractional occurrence in AL(i-1) (Rackovsky-Scheraga, 1982)	61	62.5677	60.0869	61.0587	57.6882	57.1552	54.8507	51.5827	53.1983	55.4207
Average relative fractional occurrence in AR(i) (Rackovsky-Scheraga, 1982)	62	62.5472	60.0992	61.518	58.8937	60.5995	57.598	58.4878	56.0194	53.8134
Average relative fractional occurrence in AR(i-1) (Rackovsky-Scheraga, 1982)	63	62.5923	60.3617	61.3908	59.64	59.9598	56.5237	57.0239	56.2408	55.8225
Average relative fractional occurrence in E0(i) (Rackovsky-Scheraga, 1982)	64	62.99	61.0054	61.5549	60.5913	60.9808	57.4545	59.6031	54.8302	55.5396
Average relative fractional occurrence in E0(i-1) (Rackovsky-Scheraga, 1982)	65	62.9859	61.0997	61.399	60.2837	60.8824	53.9323	59.927	56.8968	54.3464
Average relative fractional occurrence in EL(i) (Rackovsky-Scheraga, 1982)	66	62.5144	60.1935	61.5467	59.9434	60.6856	55.6134	57.8932	54.2029	55.6995

Table C.3: (Page 3 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	#48 0.2,100,3
Average relative fractional occurrence in EL(i-1) (Rackovsky-Scheraga, 1982)	67	62.3585	60.4314	61.3703	58.4345	58.1515	56.3105	54.5022	54.2849	55.351
Average relative fractional occurrence in ER(i) (Rackovsky-Scheraga, 1982)	68	62.461	60.3165	61.5221	57.8809	59.5047	55.064	53.063	52.3782	54.5145
Average relative fractional occurrence in ER(i-1) (Rackovsky-Scheraga, 1982)	69	62.5185	60.5954	61.7189	59.64	58.168	54.8507	52.4479	54.4817	54.2521
Average relative probability of beta-sheet (Kanehisa-Tsong, 1980)	70	63.7937	61.8788	62.3667	62.1576	61.2637	58.3525	59.8819	56.0522	55.6503
Average relative probability of helix (Kanehisa-Tsong, 1980)	71	62.539	60.4724	61.5549	60.2755	60.3247	57.6431	58.0736	57.3356	56.0604
Average relative probability of inner beta-sheet (Kanehisa-Tsong, 1980)	72	64.1832	62.0223	62.3708	61.6656	61.0628	57.5939	58.6559	56.4581	55.4289
Average relative probability of inner helix (Kanehisa-Tsong, 1980)	73	62.6702	60.743	61.6533	60.5093	60.3247	58.8814	58.1105	57.8399	56.5811
Average side chain orientation angle (Meirovitch et al., 1980)	74	64.8516	62.1863	62.7563	62.4118	61.6984	60.1771	61.0177	59.886	57.1552
Average surrounding hydrophobicity (Manavalan-Ponnuswamy, 1978)	75	64.6753	61.8911	63.4533	62.58	62.3257	58.3607	61.2842	58.2336	57.9752
Average volume of buried residue (Chothia, 1975)	76	62.6046	60.4519	61.6492	59.4514	60.9193	57.2044	60.173	54.5801	56.1957
Average volumes of residues (Pontius et al., 1996)	77	62.58	60.5257	61.4647	58.9839	60.7676	57.434	60.4068	54.9697	56.0973
Averaged turn propensities in a transmembrane helix (Monne et al., 1999)	78	64.1914	61.7968	62.5021	62.4897	61.4483	59.7753	60.1484	58.5903	56.7082
Beta-coil equilibrium constant (Ptitsyn-Finkelstein, 1983)	79	64.0561	61.7025	62.2683	61.6574	60.9644	60.0705	58.6108	57.2536	57.9342
Beta-sheet propensity derived from designed sequences (Koehl-Levitt, 1999)	80	63.1417	61.071	62.1453	60.1771	59.8245	53.6083	56.9419	56.7451	54.9328
Beta-strand indices (Geisow-Roberts, 1980)	81	63.2032	61.8255	61.8214	60.9767	61.112	57.1716	59.7753	56.4048	55.5888
Beta-strand indices for alpha/beta-proteins (Geisow-Roberts, 1980)	82	63.7076	61.9403	62.2519	61.7107	61.2186	59.4637	60.2468	58.0244	56.4581
Beta-strand indices for beta-proteins (Geisow-Roberts, 1980)	83	63.277	61.4975	61.6615	61.5836	60.7348	59.0413	59.9557	56.3392	56.3925
Bitterness (Venanzi, 1984)	84	64.3718	62.2601	62.3954	60.4109	57.0239	55.6544	58.455	57.4832	56.5934
Bulkiness (Zimmerman et al., 1968)	85	63.359	61.5057	62.0264	59.8819	61.1448	56.331	59.4637	58.1105	57.9137
Buriability (Zhou-Zhou, 2004)	86	65.0976	62.4036	63.4943	62.6415	62.2355	58.2746	59.9516	58.2582	57.3069
Composition (Grantham, 1974)	87	63.3918	61.3662	62.0264	60.4847	61.0587	57.3479	60.3494	58.0162	57.2208
Composition of amino acids in anchored proteins (percent) (Cedano et al., 1997)	88	62.5677	60.0869	61.5754	59.517	60.4109	56.0768	58.7871	55.0435	53.7518
Composition of amino acids in extracellular proteins (percent) (Cedano et al., 1997)	89	62.5513	60.173	61.4606	59.2299	59.8409	54.7482	58.9921	55.3018	53.9733
Composition of amino acids in intracellular proteins (percent) (Cedano et al., 1997)	90	62.5964	60.3986	61.3211	59.4555	60.255	55.1911	59.2299	55.0353	53.6862

Table C.4: (Page 4 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	#48 0.2,100,3
Composition of amino acids in membrane proteins (percent) (Cedano et al., 1997)	91	62.8301	60.3699	61.1981	59.3038	60.2058	54.3956	57.7087	54.8712	54.0676
Composition of amino acids in nuclear proteins (percent) (Cedano et al., 1997)	92	62.4692	60.1566	61.6	59.5867	60.1976	55.3223	59.5908	55.9332	53.961
Conformational parameter of beta-structure (Beghin-Dirkx, 1975)	93	63.8306	61.8829	62.2109	62.3134	61.0956	57.7661	60.0582	57.6513	56.4294
Conformational parameter of beta-turn (Beghin-Dirkx, 1975)	94	63.1458	60.6446	61.4647	60.4027	59.8983	59.1807	60.0951	56.5196	56.4991
Conformational parameter of inner helix (Beghin-Dirkx, 1975)	95	62.8834	61.0013	61.6041	61.2924	60.2755	59.1766	59.0208	57.9219	56.5401
Conformational preference for all beta-strands (Lifson-Sander, 1979)	96	63.9372	61.8583	62.6456	61.7722	61.2555	58.9716	59.8532	57.5406	57.3356
Conformational preference for antiparallel beta-strands (Lifson-Sander, 1979)	97	63.3959	61.4688	62.3175	61.0218	61.0218	58.3484	59.8655	57.8604	56.7738
Conformational preference for parallel beta-strands (Lifson-Sander, 1979)	98	64.2611	62.0633	62.3708	61.4196	60.6118	57.1593	57.5078	55.7815	57.7661
Consensus normalized hydrophobicity scale (Eisenberg, 1984)	99	65.032	62.1494	63.441	61.7107	61.5918	59.6072	58.6436	57.7825	57.557
Correlation coefficient in regression analysis (Prabhakaran-Ponnuswamy, 1982)	100	62.4118	60.579	61.3252	59.4145	60.5954	56.1834	59.9024	55.2895	54.3177
Delta G values for the peptides extrapolated to 0 M urea (O*Neil-DeGrado, 1990)	101	62.5226	60.3535	61.4031	60.0131	53.6862	57.9137	48.9708	49.7417	56.6385
Dependence of partition coefficient on ionic strength (Zaslavsky et al., 1982)	102	62.8793	61.2637	61.6041	60.1894	60.3863	56.044	57.6718	56.2367	55.9168
Direction of hydrophobic moment (Eisenberg-McLachlan, 1986)	103	64.8803	62.4118	63.0474	61.8747	59.9393	59.9598	58.2992	58.2541	57.5406
Distance between C-alpha and centroid of side chain (Levitt, 1976)	104	62.5882	60.2468	61.3826	58.4017	60.6528	56.2859	59.6318	55.2977	53.7969
Distribution of amino acid residues in the 18 non-redundant families of mesophilic proteins (Kumar et al., 2000)	105	62.6251	60.4888	61.4934	58.8199	60.4109	55.6011	60.4437	55.2649	53.9077
Distribution of amino acid residues in the 18 non-redundant families of thermophilic proteins (Kumar et al., 2000)	106	62.662	60.1976	61.3498	58.7338	60.9152	55.3387	60.3658	55.0353	53.9077
Distribution of amino acid residues in the alpha-helices in mesophilic proteins (Kumar et al., 2000)	107	62.539	60.132	61.3744	58.8978	58.8691	54.4899	55.1542	54.7605	54.7646
Distribution of amino acid residues in the alpha-helices in thermophilic proteins (Kumar et al., 2000)	108	62.4487	60.3781	61.6492	59.2997	59.3243	55.8963	53.8585	54.4817	54.8466
Effective partition energy (Miyazawa-Jernigan, 1985)	109	65.4543	62.744	64.1258	63.3262	62.9121	59.7425	61.1284	59.8409	58.1926
Electron-ion interaction potential (Veljkovic et al., 1985)	110	62.5677	60.2345	61.5262	58.3771	60.9111	56.495	59.9598	55.8717	53.6985

Table C.5: (Page 5 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Electron-ion interaction potential values (Cosic, 1994)	111	62.5636	60.3494	61.5918	58.3812	60.8578	56.3638	59.9516	55.8594	54.9328
Energy transfer from out to in(95%buried) (Radzicka-Wolfenden, 1988)	112	63.9987	61.4893	61.5016	60.6159	60.1361	56.3884	58.0818	56.413	55.3797
Entire chain compositino of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa, 2001)	113	62.5226	59.9311	61.4852	59.2915	60.4109	53.4402	58.0162	54.0922	54.084
Entire chain composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)	114	62.5964	60.538	61.6287	58.7379	60.8414	53.9733	60.8414	56.0809	53.4525
Entire chain composition of amino acids in intracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)	115	62.6046	60.0746	61.6574	58.9634	60.62	53.0138	59.3981	55.3141	54.4694
Entire chain composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa, 2001)	116	62.5349	60.1648	61.5262	59.0167	60.6774	53.6329	59.2915	55.2321	53.9569
Entropy of formation (Hutchens, 1970)	117	62.6005	60.2919	61.559	59.3612	60.784	57.0773	59.1028	55.8881	55.5109
Flexibility parameter for no rigid neighbors (Karplus-Schulz, 1985)	118	62.9613	61.6902	62.0264	60.9357	60.3412	58.9019	53.6944	55.5765	56.495
Flexibility parameter for one rigid neighbor (Karplus-Schulz, 1985)	119	63.2155	61.6697	62.174	62.1781	60.4437	57.7743	54.7195	58.0777	57.6431
Flexibility parameter for two rigid neighbors (Karplus-Schulz, 1985)	120	62.539	61.1366	61.3457	59.8163	59.1192	56.0317	43.7428	54.2603	56.5073
Fraction of site occupied by water (Krigbaum-Komoraya, 1979)	121	64.3103	61.9977	62.0592	61.7968	61.0833	59.152	59.7507	57.7456	56.8394
Free energies of transfer of AcWI-X-LL peptides from bilayer interface to water (Wimley-White, 1996)	122	63.5804	61.3703	62.2109	60.866	61.3457	59.029	59.1438	57.1675	55.1747
Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga, 1978)	123	62.5718	60.3822	61.4811	58.7133	59.9311	54.6129	56.2941	55.9291	53.8052
Free energy change of epsilon(i) to alpha(Rh) (Wertz-Scheraga, 1978)	124	63.1417	60.3042	61.8296	59.2956	60.6077	56.5606	57.7374	56.864	54.1168
Free energy change of epsilon(i) to epsilon(ex) (Wertz-Scheraga, 1978)	125	62.7194	60.2304	61.4319	59.3325	60.5421	54.1373	55.6093	54.6293	54.4981
Free energy in alpha-helical conformation (Munoz-Serrano, 1994)	126	62.42	60.8127	61.7804	60.1812	58.1639	58.3525	52.6939	53.8667	56.741
Free energy in alpha-helical region (Munoz-Serrano, 1994)	127	62.3749	60.4232	61.4729	59.7958	58.7625	57.2946	54.5391	55.3346	56.3474
Free energy in beta-strand conformation (Munoz-Serrano, 1994)	128	63.3631	61.6246	62.2888	61.4524	60.7635	59.1889	59.6646	57.1511	56.782
Free energy in beta-strand region (Munoz-Serrano, 1994) (1)	543	62.5062	60.8332	61.5385	58.988	56.4499	60.1525	49.6638	51.8247	56.6344
Free energy in beta-strand region (Munoz-Serrano, 1994) (2)	129	62.5144	60.5626	61.436	58.6108	56.0768	56.7164	49.2537	50.8652	56.6918
Free energy of solution in water, kcal/mole (Charton-Charton, 1982)	130	62.4938	60.6692	61.4975	59.6728	60.4478	57.352	55.6749	56.5606	55.105
Frequency of occurrence in beta-bends (Lewis et al., 1971)	131	62.6702	60.3617	61.3703	59.7589	59.9188	56.3351	58.6518	56.8681	55.3797

Table C.6: (Page 6 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Frequency of the 1st residue in turn (Chou-Fasman, 1978b)	132	62.5759	60.3658	61.1776	59.7671	60.4724	57.2782	58.701	56.9829	56.5606
Frequency of the 2nd residue in turn (Chou-Fasman, 1978b)	133	63.4328	61.3211	61.4401	60.5913	57.4504	57.7456	50.4633	52.8826	57.3356
Frequency of the 3rd residue in turn (Chou-Fasman, 1978b)	134	62.8178	60.8332	61.1571	59.111	60.4068	56.1793	57.9014	56.1342	55.6298
Frequency of the 4th residue in turn (Chou-Fasman, 1978b)	135	62.5513	60.3289	61.5877	58.4468	61.5649	54.9656	54.4325	52.9605	54.1742
Graph shape index (Fauchere et al., 1988)	136	63.1417	61.0874	61.5959	59.6113	60.5216	54.9738	59.0331	56.7779	55.4084
HPLC parameter (Parker et al., 1986)	137	65.36	63.2401	63.5066	62.9613	60.6364	59.4268	59.193	59.2792	58.4755
Heat capacity (Hutchens, 1970)	138	62.621	60.6856	61.559	59.7507	60.907	57.27	60.3494	55.31	55.9784
Helix formation parameters (delta delta G) (O*Neil-DeGrado, 1990)	139	62.4733	60.5298	61.7722	59.2135	53.4525	57.0732	47.4824	51.1522	56.5237
Helix initiation parameter at position i,j+1,i+2 (Finkelstein et al., 1991)	140	62.4774	60.4601	61.153	58.3976	52.2675	53.719	47.7612	48.2204	55.2444
Helix initiation parameter at position i-1 (Finkelstein et al., 1991)	141	62.5021	60.7717	61.3006	58.8896	60.1525	51.8534	57.6513	53.3582	54.453
Helix termination parameter at position j+1 (Finkelstein et al., 1991)	142	62.4159	60.5544	60.948	59.2258	58.0777	57.1839	55.9209	55.1583	54.658
Helix termination parameter at position j-2,j-1,j (Finkelstein et al., 1991)	143	62.4774	60.132	61.2227	59.1643	59.2956	55.3059	59.8532	54.1291	53.6821
Helix-coil equilibrium constant (Finkelstein-Ptitsyn, 1977)	144	62.5841	60.6364	61.358	60.0664	59.8163	59.2505	58.0654	56.5852	56.741
Helix-coil equilibrium constant (Ptitsyn-Finkelstein, 1983)	145	62.703	60.8906	61.1571	60.4806	58.5452	60.0254	55.0927	55.4986	57.1347
Hydration free energy (Robson-Osguthorpe, 1979)	146	64.4579	62.4077	62.785	61.9731	61.1571	57.4053	57.7497	57.2536	57.4258
Hydration number (Hopfinger, 1971), Cited by Charton-Charton (1982)	147	63.4779	61.2227	61.7558	59.6564	59.0823	56.7492	54.5842	55.1583	54.6006
Hydration potential (Wolfenden et al., 1981)	148	63.888	61.0915	62.5021	60.6979	60.3083	56.5278	57.0075	56.4376	55.802
Hydropathies of amino acid side chains, neutral form (Roseman, 1988)	149	64.4702	61.8091	63.2401	59.7671	60.7758	57.5816	58.4386	58.7338	56.2572
Hydropathies of amino acid side chains, pi-values in pH 7.0 (Roseman, 1988)	150	65.2329	62.4938	63.318	61.3621	60.8414	59.3858	58.7051	58.6518	56.9173
Hydropathy index (Kyte-Doolittle, 1982)	151	64.9295	62.3052	63.2647	62.0633	60.661	60.5626	58.4509	58.2992	57.0937
Hydropathy scale based on self-information values in the two-state model (16% accessibility) (Naderi-Manesh et al., 2001)	152	65.1796	62.4487	63.0474	62.5554	60.3124	59.681	57.8112	58.3197	58.6518
Hydropathy scale based on self-information values in the two-state model (20% accessibility) (Naderi-Manesh et al., 2001)	153	65.0156	62.6005	63.2401	62.3708	60.1894	59.2258	57.4258	57.9014	57.3274

Table C.7: (Page 7 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Hydrophathy scale based on self-information values in the two-state model (25% accessibility) (Naderi-Manesh et al., 2001)	154	64.6096	62.5103	62.8834	62.2806	60.0664	57.885	57.0691	57.8522	56.3146
Hydrophathy scale based on self-information values in the two-state model (36% accessibility) (Naderi-Manesh et al., 2001)	155	63.4984	61.6902	62.133	60.7553	61.0382	57.4135	57.967	58.4222	57.4463
Hydrophathy scale based on self-information values in the two-state model (5% accessibility) (Naderi-Manesh et al., 2001)	156	64.3882	62.1986	62.8834	62.174	60.1853	59.2505	57.8809	58.0367	57.5119
Hydrophathy scale based on self-information values in the two-state model (50% accessibility) (Naderi-Manesh et al., 2001)	157	62.5185	60.7676	61.5795	59.5252	59.9967	57.4053	56.8271	57.4996	56.5811
Hydrophathy scale based on self-information values in the two-state model (9% accessibility) (Naderi-Manesh et al., 2001)	158	65.1427	62.4241	63.2032	62.0305	60.374	59.886	57.9957	58.0654	57.9424
Hydrophilicity scale (Kuhn et al., 1995)	159	64.4046	61.5467	62.6415	61.1243	61.7599	58.7051	59.5785	56.7902	57.6431
Hydrophilicity value (Hopp-Woods, 1981)	160	64.7614	62.6292	62.6046	61.7927	60.6159	59.6277	58.0449	58.09	58.9839
Hydrophobic parameter (Levitt, 1976)	161	64.7696	62.1781	62.9326	61.4196	60.6856	58.5739	57.4381	57.7169	57.4709
Hydrophobic parameter pi (Fauchere-Pliska, 1983)	162	65.2575	62.6292	63.6543	62.7235	61.928	60.8619	59.0085	59.1151	58.4509
Hydrophobicity (Jones, 1975)	163	63.5066	60.9521	62.5677	60.3042	61.2637	58.1844	59.0946	58.0162	55.2444
Hydrophobicity (Prabhakaran, 1990)	164	64.5563	62.2478	63.1089	61.2473	60.4724	60.1279	57.3192	57.6759	58.1967
Hydrophobicity (Zimmerman et al., 1968)	165	63.4615	60.8291	62.2109	60.0951	61.1079	58.0121	58.8158	56.9337	55.0189
Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-ProH/MeCN/H2O (Wilce et al. 1995)	166	62.6784	60.4642	61.4893	59.4063	59.5662	54.1783	55.8922	54.2685	54.2152
Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)	167	64.5645	62.1412	62.5636	61.4811	60.7758	55.5314	57.7374	58.414	57.2208
Hydrophobicity coefficient in RP-HPLC, C4 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)	168	62.3462	60.0869	61.3744	59.6892	60.5462	55.4945	59.3571	55.9209	53.6739
Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)	169	63.8347	61.8706	62.6538	61.3416	60.7143	58.2869	57.7087	57.885	56.4048
Hydrophobicity factor (Goldsack-Chalifoux, 1973)	170	63.5107	61.3457	62.4651	60.5585	61.3949	56.3515	59.1233	56.6016	54.941
Hydrophobicity index (Argos et al., 1982)	171	63.5189	60.9726	62.5677	60.2632	61.2678	58.0654	59.0126	57.9629	55.1747
Hydrophobicity index (Engelman et al., 1986)	172	64.5563	62.5349	63.0966	61.2432	60.5339	60.1935	57.311	57.6759	58.4181
Hydrophobicity index (Fasman, 1989)	173	65.6921	62.5718	63.9126	62.9121	61.4565	60.0869	58.701	59.2423	58.3197
Hydrophobicity index (Wolfenden et al., 1979)	174	63.7363	61.2227	62.3708	60.4109	60.0992	55.4617	56.7943	56.0522	55.9825

Table C.8: (Page 8 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Hydrophobicity index, 3.0 pH (Cowan-Whittaker, 1990)	175	65.2247	62.6251	63.2852	61.805	61.3375	58.3074	58.2705	57.9178	57.4258
Hydrophobicity scale from native protein structures (Casari-Sippl, 1992)	176	64.8434	62.2765	62.9449	62.0756	60.2427	57.9916	57.2618	58.1474	57.1839
Hydrophobicity scales (Ponnuswamy, 1993)	177	65.1263	62.8014	63.1704	62.6825	61.5795	58.0818	58.6641	58.3115	57.8973
Hydrophobicity-related index (Kidera et al., 1985)	178	64.9131	61.7845	63.2114	61.8747	60.6282	59.599	57.2782	57.3438	55.6421
Hydrostatic pressure asymmetry index, PAI (Di Giulio, 2005)	179	62.6005	60.661	61.399	58.7953	60.3699	56.536	57.9342	56.6057	54.0594
Information measure for C-terminal helix (Robson-Suzuki, 1976)	180	62.6743	60.6487	61.6287	59.2751	56.1096	56.6057	49.8442	52.8908	56.1957
Information measure for C-terminal turn (Robson-Suzuki, 1976)	181	62.6046	60.2345	61.2596	58.5083	59.6728	55.8225	55.5232	55.2854	55.2772
Information measure for N-terminal helix (Robson-Suzuki, 1976)	182	62.7727	60.5011	61.3949	59.2792	59.6277	55.5888	55.8799	56.1875	53.6493
Information measure for N-terminal turn (Robson-Suzuki, 1976)	183	62.4692	60.538	61.1366	60.1607	59.3161	53.2721	55.8348	56.8312	54.535
Information measure for alpha-helix (Robson-Suzuki, 1976)	184	62.4241	60.6938	61.7353	59.9024	60.3986	56.6344	58.3115	58.1064	55.6421
Information measure for coil (Robson-Suzuki, 1976)	185	62.8465	60.702	61.8214	60.0172	59.9475	57.6513	57.8317	57.9793	56.8681
Information measure for extended (Robson-Suzuki, 1976)	186	63.8716	62.0797	62.4815	61.4647	60.8373	56.7615	57.475	56.6672	56.7123
Information measure for extended without H-bond (Robson-Suzuki, 1976)	187	62.6005	60.333	61.436	59.1807	59.2669	56.5319	54.7851	55.8553	54.2357
Information measure for loop (Robson-Suzuki, 1976)	188	63.195	60.7348	62.0756	61.0259	61.1694	57.6267	60.3165	58.8527	55.925
Information measure for middle helix (Robson-Suzuki, 1976)	189	62.8834	60.62	61.5016	60.5872	57.147	59.599	51.9272	55.7569	57.6636
Information measure for middle turn (Robson-Suzuki, 1976)	190	63.4902	60.7799	61.9977	60.6569	61.2432	58.9675	60.0254	58.7215	56.5729
Information measure for pleated-sheet (Robson-Suzuki, 1976)	191	63.9741	61.9239	62.3093	62.3544	60.8045	59.2053	58.2869	58.1269	56.495
Information measure for turn (Robson-Suzuki, 1976)	192	63.277	60.8414	61.6369	60.8168	61.3252	57.3356	60.1689	58.865	55.5519
Information value for accessibility; average fraction 23% (Biou et al., 1988)	193	65.4625	62.6702	63.3303	62.3257	61.0833	59.9475	58.4714	58.3771	58.3443
Information value for accessibility; average fraction 35% (Biou et al., 1988)	194	65.565	62.8465	63.6871	63.0105	61.2637	60.5995	58.6887	59.0495	58.7092
Interactivity scale obtained by maximizing the mean of correlation coefficient over pairs of sequences sharing the TIM barrel fold (Bastolla et al., 2005)	195	65.1345	62.5923	63.031	62.8342	61.7599	59.6154	59.271	58.8691	57.7415

Table C.9: (Page 9 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Interactivity scale obtained by maximizing the mean of correlation coefficient over single-domain globular proteins (Bastolla et al., 2005)	196	65.0238	62.5267	63.5558	62.5144	61.8296	60.2878	58.9921	58.7379	58.5206
Interactivity scale obtained from the contact matrix (Bastolla et al., 2005)	197	65.1714	62.9736	63.3672	63.1499	62.1699	57.803	59.2997	58.7543	58.2213
Intercept in regression analysis (Prabhakaran-Ponnuswamy, 1982)	198	62.6374	60.9152	61.4975	60.3042	60.6159	57.9465	59.3325	56.9501	56.4171
Interior composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)	199	63.236	60.5913	61.6697	59.4309	60.6241	55.0517	58.0449	55.7733	54.5473
Interior composition of amino acids in intracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)	200	63.3754	60.9931	61.7599	59.722	60.7143	54.9328	57.9998	56.6549	54.9492
Interior composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa, 2001)	201	63.4697	61.1284	61.8665	60.0746	60.7225	55.105	57.3684	56.1793	55.0927
Interior composition of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa, 2001)	202	63.8142	61.559	61.682	60.1648	60.2632	57.5652	57.2905	55.5191	56.1957
Isoelectric point (Zimmerman et al., 1968)	203	62.5636	60.4027	61.477	59.2176	60.3165	54.5924	56.1055	54.4571	53.879
Knowledge-based membrane-propensity scale from 1D_Helix in MPtopo databases (Punta-Maritan, 2003)	204	64.9336	62.6784	63.3139	62.6784	60.4355	59.5539	58.7502	59.1848	57.7005
Knowledge-based membrane-propensity scale from 3D_Helix in MPtopo databases (Punta-Maritan, 2003)	205	64.9459	62.785	62.9695	62.7317	61.0997	59.0782	58.9675	59.3161	57.9301
Linker index (Bae et al., 2005)	206	64.5686	62.4077	63.0966	61.7763	61.6287	58.824	58.8445	59.0905	57.7251
Linker propensity from 1-linker dataset (George-Heringa, 2003)	207	62.5431	60.5544	61.5836	59.3899	57.3356	55.8635	52.1732	52.8498	55.5806
Linker propensity from 2-linker dataset (George-Heringa, 2003)	208	62.4733	60.5462	61.4729	59.6523	58.824	57.7169	53.6042	54.863	56.0645
Linker propensity from 3-linker dataset (George-Heringa, 2003)	209	62.4569	60.3083	61.8829	59.7302	58.7953	57.6431	53.7682	54.2972	55.6954
Linker propensity from all dataset (George-Heringa, 2003)	210	62.5144	60.3289	61.5139	59.6113	57.6472	57.1429	52.3864	53.6944	55.0886
Linker propensity from helical (annotated by DSSP) dataset (George-Heringa, 2003)	211	62.539	60.5872	61.559	59.8696	60.0418	57.6472	57.9793	56.7164	56.1547
Linker propensity from long dataset (linker length is greater than 14 residues) (George-Heringa, 2003)	212	62.5144	60.7102	61.6041	60.0705	60.0049	57.393	55.679	55.187	55.2444
Linker propensity from medium dataset (linker length is between six and 14 residues) (George-Heringa, 2003)	213	62.4405	60.6569	61.3867	59.4555	58.2664	57.6595	53.7436	55.0066	57.4668

Table C.10: (Page 10 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	#48 0.2,100,3
Linker propensity from non-helical (annotated by DSSP) dataset (George-Heringa, 2003)	214	62.5677	60.2427	61.682	59.0208	55.0927	55.3387	48.9708	50.2296	55.7446
Linker propensity from small dataset (linker length is less than six residues) (George-Heringa, 2003)	215	63.7158	60.9275	62.0305	60.2755	61.5426	59.476	60.3412	57.1675	55.1542
Linker propensity index (Suyama-Ohara, 2003)	216	62.9203	60.62	61.3211	59.5047	58.8937	57.7251	54.0225	55.2034	54.863
Localized electrical effect (Fauchere et al., 1988)	217	62.9408	60.333	61.7353	59.4555	60.5257	54.2111	57.3971	55.7077	54.0061
Long range non-bonded energy per atom (Oobatake-Ooi, 1977)	218	64.6876	62.3216	63.2647	62.5595	62.6169	59.4883	61.2145	58.7133	57.352
Loss of Side chain hydrophathy by helix formation (Roseman, 1988)	219	62.5472	60.2755	61.3047	58.7707	60.1976	56.0768	59.7056	55.7405	54.4079
Mean area buried on transfer (Rose et al., 1985)	220	63.1212	61.3088	61.8788	61.1161	61.1407	58.1967	60.8496	56.3187	56.3228
Mean fractional area loss (Rose et al., 1985)	221	64.9418	62.3257	63.2565	62.8711	62.0838	60.5421	61.8255	58.9552	57.6226
Mean polarity (Radzicka-Wolfenden, 1988)	222	65.3805	62.8752	63.7814	63.1212	61.8214	60.3822	59.2669	59.3243	58.3443
Mean volumes of residues buried in protein interiors (Harpaz et al., 1994)	223	62.6989	60.4478	61.6492	59.2423	60.9111	56.1711	60.4601	54.8918	56.0727
Melting point (Fasman, 1976)	224	63.1827	60.6405	61.8952	60.9521	60.9316	58.09	59.5621	56.8148	54.494
Membrane preference for cytochrome b: MPH89 (Degli Esposti et al., 1990)	225	64.1053	62.2437	62.4446	61.9895	61.0259	56.8886	58.0285	57.6226	55.6954
Membrane-buried preference parameters (Argos et al., 1982)	226	64.3964	62.133	62.5144	61.2965	60.5257	60.2345	58.0859	57.8686	57.5611
Modified Kyte-Doolittle hydrophobicity scale (Juretic et al., 1998)	227	64.9664	62.6661	63.1704	62.2109	61.0382	59.7343	58.7543	58.4181	57.7907
Molecular weight (Fasman, 1976)	228	62.6333	60.6405	61.6984	59.1192	60.9808	56.8558	59.9229	54.6539	54.7236
N.m.r. chemical shift of alpha-carbon (Fauchere et al., 1988)	229	62.5718	60.62	61.5426	57.639	59.6933	56.6795	55.1583	54.8753	56.4745
NNEIG index (Cornette et al., 1987)	230	64.1627	62.5103	62.7604	62.3339	61.3949	59.3448	58.9388	58.3566	57.6759
Negative charge (Fauchere et al., 1988)	231	62.6087	60.2591	60.7184	56.9091	51.9682	51.3736	52.2429	53.0466	53.838
Net charge (Klein et al., 1984)	232	63.2196	60.091	61.276	59.2217	53.4361	53.0507	55.1296	55.7528	54.0922
Normalized average hydrophobicity scales (Cid et al., 1992)	233	64.8721	62.8178	63.1581	62.6784	61.9567	58.9019	58.7092	58.5206	58.1064
Normalized composition from animal (Nakashima et al., 1990)	234	63.5476	61.0751	62.2724	60.7307	60.0582	59.0085	58.0039	57.4996	55.8471
Normalized composition from fungi and plant (Nakashima et al., 1990)	235	63.277	60.9275	61.8583	60.702	60.2714	56.8148	56.6754	56.0891	54.6129
Normalized composition of membrane proteins (Nakashima et al., 1990)	236	64.544	61.7927	62.6702	61.4975	60.3945	56.5811	57.4258	57.0403	55.3961
Normalized composition of mt-proteins (Nakashima et al., 1990)	237	63.7117	61.3908	62.1986	61.0628	60.0705	58.6887	58.25	57.5898	55.4084

Table C.11: (Page 11 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Normalized flexibility parameters (B-values) for each residue surrounded by none rigid neighbours (Vihinen et al., 1994)	238	62.9531	62.0633	62.1863	61.8132	60.5544	58.5042	49.3029	56.9214	57.9055
Normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbours (Vihinen et al., 1994)	239	62.7276	61.9854	61.9075	62.2191	59.9434	59.5088	42.8653	53.2065	57.0732
Normalized flexibility parameters (B-values) for each residue surrounded by two rigid neighbours (Vihinen et al., 1994)	240	62.6538	61.2022	62.215	61.276	59.9229	57.8153	40.6757	50.5618	55.515
Normalized flexibility parameters (B-values), average (Vihinen et al., 1994)	241	63.3262	62.1822	62.4569	62.6497	61.358	59.8942	53.6288	57.9793	57.0855
Normalized frequency of C-terminal beta-sheet (Chou-Fasman, 1978b)	242	63.6502	61.8911	62.1043	60.9193	61.0792	58.6313	59.4883	57.4627	56.3556
Normalized frequency of C-terminal helix (Chou-Fasman, 1978b)	243	62.621	60.4437	61.5959	59.7343	59.6359	57.3643	57.0649	56.6877	55.8225
Normalized frequency of C-terminal non beta region (Chou-Fasman, 1978b)	244	62.9408	60.8168	61.6	60.0705	59.7712	58.9757	57.9711	56.741	58.4058
Normalized frequency of C-terminal non helical region (Chou-Fasman, 1978b)	245	62.6292	60.6692	61.4196	59.7507	60.2796	56.7082	58.3197	55.7815	56.1711
Normalized frequency of N-terminal beta-sheet (Chou-Fasman, 1978b)	246	63.3549	61.5344	61.8993	61.5918	61.3498	58.7174	59.2997	58.09	56.4663
Normalized frequency of N-terminal helix (Chou-Fasman, 1978b)	247	62.8055	60.7922	61.4606	59.9188	59.968	55.8963	54.7072	55.228	55.1501
Normalized frequency of N-terminal non beta region (Chou-Fasman, 1978b)	248	63.6092	61.2801	61.9649	61.6779	61.1038	59.7835	60.2427	58.8445	58.1269
Normalized frequency of N-terminal non helical region (Chou-Fasman, 1978b)	249	62.4815	60.6979	61.5426	59.4719	60.3658	56.8968	59.111	57.6595	55.7364
Normalized frequency of alpha region (Maxfield-Scheraga, 1976)	250	62.8383	60.6979	61.4196	59.5252	59.476	56.8435	55.843	55.105	56.5114
Normalized frequency of alpha-helix (Burgess et al., 1974)	251	62.5226	60.3206	61.641	59.9762	60.8332	56.5647	59.4186	57.2085	56.3228
Normalized frequency of alpha-helix (Chou-Fasman, 1978b)	252	62.6087	60.6651	61.4934	60.1648	60.2878	57.4012	58.4427	57.4094	56.2039
Normalized frequency of alpha-helix (Maxfield-Scheraga, 1976)	253	62.5554	60.6446	61.3047	60.2181	60.1279	58.4796	58.3402	57.2864	56.6549
Normalized frequency of alpha-helix (Nagano, 1973)	254	62.5677	60.6159	61.2514	59.8942	60.3206	56.8025	58.4386	55.7897	55.9825
Normalized frequency of alpha-helix (Tanaka-Scheraga, 1977)	255	62.5103	60.5462	61.3744	59.7261	60.8209	57.2823	58.8814	56.9009	56.0973
Normalized frequency of alpha-helix from CF (Palau et al., 1981)	256	62.5308	60.6405	61.5508	60.3371	60.3494	57.9506	58.2746	56.9132	56.5483
Normalized frequency of alpha-helix from LG (Palau et al., 1981)	257	62.6415	60.3699	61.2432	59.4924	60.4437	57.2864	58.6067	55.3469	55.7405
Normalized frequency of alpha-helix in all-alpha class (Palau et al., 1981)	258	62.5267	60.4642	61.518	59.3243	59.5334	56.1465	54.4202	54.74	54.8589

Table C.12: (Page 12 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Normalized frequency of alpha-helix in alpha+beta class (Palau et al., 1981)	259	62.4323	60.8045	61.4401	59.4063	60.5175	56.3269	58.008	55.9332	55.9045
Normalized frequency of alpha-helix in alpha/beta class (Palau et al., 1981)	260	62.4815	60.1197	60.9398	59.6318	60.4396	58.0039	58.0244	56.5114	55.5683
Normalized frequency of alpha-helix, unweighted (Levitt, 1978)	261	62.6333	60.6651	61.5508	59.6769	60.456	57.1675	58.2828	56.4089	56.0686
Normalized frequency of alpha-helix, with weights (Levitt, 1978)	262	62.58	60.5749	61.3662	59.8737	60.3986	57.3684	57.9096	56.4991	55.9701
Normalized frequency of beta-structure (Nagano, 1973)	263	63.6953	61.8952	62.1863	61.8296	61.3416	58.5493	59.7507	57.5652	56.9009
Normalized frequency of beta-sheet (Chou-Fasman, 1978b)	264	63.724	61.9649	62.1494	62.1412	61.5754	59.6687	60.2755	57.6595	56.0727
Normalized frequency of beta-sheet (Crawford et al., 1973)	265	63.3836	61.7558	62.2437	61.7107	61.2145	59.3858	60.0131	57.6021	56.9788
Normalized frequency of beta-sheet from CF (Palau et al., 1981)	266	63.7568	62.092	62.4897	61.887	61.2719	59.2792	60.0459	57.0239	56.2736
Normalized frequency of beta-sheet from LG (Palau et al., 1981)	267	63.2565	61.3129	61.764	61.0997	60.7758	57.6513	59.8122	56.7287	55.0681
Normalized frequency of beta-sheet in all-beta class (Palau et al., 1981)	268	62.99	61.2309	61.4852	60.5708	60.7225	57.229	58.5616	55.3469	56.208
Normalized frequency of beta-sheet in alpha+beta class (Palau et al., 1981)	269	62.9285	61.0013	61.9075	61.0669	60.6036	57.1429	59.3981	55.7569	54.1578
Normalized frequency of beta-sheet in alpha/beta class (Palau et al., 1981)	270	63.236	61.6943	61.8296	60.8619	60.6036	56.9132	59.1807	56.5729	54.863
Normalized frequency of beta-sheet, unweighted (Levitt, 1978)	271	63.3016	61.6861	61.8829	61.1858	60.8373	57.229	59.4924	56.495	55.4453
Normalized frequency of beta-sheet, with weights (Levitt, 1978)	272	63.2729	61.4934	61.764	61.0095	60.7102	57.3807	59.1397	56.3843	56.1916
Normalized frequency of beta-turn (Chou-Fasman, 1978a)	273	63.2852	60.9644	61.5508	61.1243	60.5216	60.3658	59.804	57.7292	58.4263
Normalized frequency of beta-turn (Chou-Fasman, 1978b)	274	63.6666	61.5631	61.7517	62.2232	61.2186	61.2309	60.8373	58.6149	58.578
Normalized frequency of chain reversal (Tanaka-Scheraga, 1977)	275	63.3959	61.1899	61.641	61.7189	60.7184	60.743	60.3124	57.9424	57.4094
Normalized frequency of chain reversal D (Tanaka-Scheraga, 1977)	276	63.1909	60.9398	62.0756	59.7548	60.9726	55.5232	58.7502	55.9332	54.7113
Normalized frequency of chain reversal R (Tanaka-Scheraga, 1977)	277	62.8957	61.6861	61.5016	59.9926	56.2654	57.2946	49.2291	53.3049	56.3515
Normalized frequency of chain reversal S (Tanaka-Scheraga, 1977)	278	62.5226	60.7348	61.7025	58.8281	60.2714	57.0855	56.0932	55.1747	56.4581
Normalized frequency of coil (Nagano, 1973)	279	62.6825	60.8824	61.1202	60.3617	60.6036	58.66	59.6154	57.0855	56.4745
Normalized frequency of coil (Tanaka-Scheraga, 1977)	280	62.6456	60.6569	61.5016	57.9096	58.6846	56.8763	50.5864	49.9877	54.4284
Normalized frequency of extended structure (Burgess et al., 1974)	281	63.2278	61.1694	61.7066	60.1156	61.0177	55.9332	59.5703	56.6713	55.4125
Normalized frequency of extended structure (Maxfield-Scheraga, 1976)	282	63.2155	61.1899	61.5508	60.1976	61.2063	57.6308	59.2956	56.3064	55.4002
Normalized frequency of extended structure (Tanaka-Scheraga, 1977)	283	63.2401	61.2145	61.8337	60.2017	60.9644	56.8558	60.3124	56.2531	54.7277

Table C.13: (Page 13 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Normalized frequency of isolated helix (Tanaka-Scheraga, 1977)	284	62.5185	60.0623	61.4565	58.0203	59.8983	56.823	55.8676	54.2685	54.1783
Normalized frequency of left-handed alpha-helix (Maxfield-Scheraga, 1976)	285	62.6087	60.0131	61.7763	57.147	57.2249	56.9091	49.6761	50.7709	54.3956
Normalized frequency of left-handed helix (Tanaka-Scheraga, 1977)	286	62.7768	60.1607	61.559	57.721	57.0773	56.577	52.7924	53.801	54.4243
Normalized frequency of middle helix (Crawford et al., 1973)	287	62.5308	60.5257	60.9603	60.3165	60.2181	57.4586	58.4919	55.7569	55.1501
Normalized frequency of reverse turn, unweighted (Levitt, 1978)	288	63.2811	61.5139	61.5877	61.0587	60.0049	60.7717	58.3115	56.577	58.0203
Normalized frequency of reverse turn, with weights (Levitt, 1978)	289	63.2893	61.1571	61.559	60.9398	60.0295	60.4806	57.8071	56.864	57.844
Normalized frequency of the 2nd and 3rd residues in turn (Chou-Fasman, 1978b)	290	63.6051	61.2719	61.5918	61.5672	60.8168	60.579	59.8286	58.6764	58.4427
Normalized frequency of turn (Crawford et al., 1973)	291	62.7112	60.8045	61.6082	60.702	60.4273	57.7948	59.5375	56.5852	55.9045
Normalized frequency of turn from CF (Palau et al., 1981)	292	63.0802	61.0464	61.4975	60.9357	60.7184	59.8983	59.7179	57.7538	57.5365
Normalized frequency of turn from LG (Palau et al., 1981)	293	63.1335	60.7594	61.4483	60.5995	60.3453	58.8773	58.7174	57.1183	56.9706
Normalized frequency of turn in all-alpha class (Palau et al., 1981)	294	62.6169	60.2878	61.4893	58.7748	60.2058	54.7072	57.9588	55.6954	55.802
Normalized frequency of turn in all-beta class (Palau et al., 1981)	295	62.8014	60.9439	61.6902	60.5708	60.4068	58.4222	59.2751	56.8025	56.1506
Normalized frequency of turn in alpha+beta class (Palau et al., 1981)	296	62.9039	60.6405	61.4606	59.7958	60.2017	56.8271	58.7994	57.3274	56.9583
Normalized frequency of turn in alpha/beta class (Palau et al., 1981)	297	62.621	60.4929	61.5385	59.8122	60.2427	57.3151	58.9183	56.9173	55.8963
Normalized frequency of zeta L (Maxfield-Scheraga, 1976)	298	62.5923	60.5134	61.3621	57.6554	55.0599	55.1583	49.3152	53.514	54.043
Normalized frequency of zeta R (Maxfield-Scheraga, 1976)	299	62.4692	60.4724	61.5262	58.6682	60.4314	53.2721	54.6826	55.0804	53.678
Normalized frequency of zeta R (Tanaka-Scheraga, 1977)	300	62.5062	60.1976	61.6041	58.7543	60.62	53.3582	59.2258	54.2398	54.4776
Normalized hydrophobicity scales for alpha+beta-proteins (Cid et al., 1992)	301	64.9746	62.7686	63.1007	62.58	61.4852	57.7497	58.3115	58.6928	57.5365
Normalized hydrophobicity scales for alpha-proteins (Cid et al., 1992)	302	64.4784	62.2806	62.7891	62.3831	61.4647	58.2131	58.4099	58.4427	57.1306
Normalized hydrophobicity scales for alpha/beta-proteins (Cid et al., 1992)	303	65.1591	62.7768	62.9695	62.4774	61.3867	58.4263	58.3935	58.5575	57.5775
Normalized hydrophobicity scales for beta-proteins (Cid et al., 1992)	304	64.6794	62.6046	63.0269	62.4282	61.358	59.0905	58.7133	58.6682	57.926
Normalized positional residue frequency at helix termini C ⁺ (Aurora-Rose, 1998)	305	62.5349	60.3001	61.2596	58.66	60.7553	56.3392	59.4473	55.0845	53.6985
Normalized positional residue frequency at helix termini C [*] (Aurora-Rose, 1998)	306	62.8055	61.0956	61.2555	60.2591	57.27	56.2941	52.657	53.6903	55.8553

Table C.14: (Page 14 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	#48 0.2,100,3
Normalized positional residue frequency at helix termini C* (Aurora-Rose, 1998)	307	62.5964	60.1566	61.4893	57.0732	55.3838	55.6298	49.5736	50.3813	54.1127
Normalized positional residue frequency at helix termini C1 (Aurora-Rose, 1998)	308	62.5144	60.6733	61.3334	59.2381	60.5831	57.8317	58.8937	57.7169	56.4991
Normalized positional residue frequency at helix termini C2 (Aurora-Rose, 1998)	309	62.5841	60.5257	61.4319	60.3247	60.4027	58.2951	59.1438	58.0162	56.7287
Normalized positional residue frequency at helix termini C3 (Aurora-Rose, 1998)	310	62.5513	60.7471	61.1612	60.6651	60.2304	58.8486	58.1926	57.844	57.7374
Normalized positional residue frequency at helix termini C4 (Aurora-Rose, 1998)	311	62.5267	60.4765	61.5467	60.2222	59.681	57.1388	57.967	57.2618	56.9952
Normalized positional residue frequency at helix termini C4* (Aurora-Rose, 1998)	312	62.6702	60.1771	60.9726	58.824	59.2423	52.6242	54.8466	54.3095	54.9205
Normalized positional residue frequency at helix termini C5 (Aurora-Rose, 1998)	313	62.5964	60.3904	61.5959	59.6933	60.5585	58.0326	58.6518	57.5652	56.5032
Normalized positional residue frequency at helix termini Cc (Aurora-Rose, 1998)	314	62.5841	60.3576	61.5098	59.353	60.4314	56.3064	57.1757	55.2936	55.9825
Normalized positional residue frequency at helix termini N" (Aurora-Rose, 1998)	315	62.5677	60.1853	61.2227	59.5744	59.111	56.0112	56.6795	54.6457	53.9979
Normalized positional residue frequency at helix termini N"* (Aurora-Rose, 1998)	316	62.6579	60.7102	61.2104	59.8901	59.312	56.6877	55.6544	54.8671	55.9989
Normalized positional residue frequency at helix termini N*(Aurora-Rose, 1998)	317	62.3216	60.6159	61.3252	59.9188	60.5052	56.1752	58.1392	55.6216	56.495
Normalized positional residue frequency at helix termini N1 (Aurora-Rose, 1998)	318	62.6497	60.2386	61.559	59.3202	60.0705	55.8225	52.8211	52.9482	54.576
Normalized positional residue frequency at helix termini N2 (Aurora-Rose, 1998)	319	62.6538	60.3453	61.6246	59.1397	59.7671	55.0845	55.8676	53.5099	54.1045
Normalized positional residue frequency at helix termini N3 (Aurora-Rose, 1998)	320	62.5554	60.7266	61.5959	61.0177	60.3863	58.0654	58.0531	57.6677	56.9665
Normalized positional residue frequency at helix termini N4 (Aurora-Rose, 1998)	321	62.703	60.5995	61.3703	60.8619	59.9147	58.5698	58.3443	58.1474	56.823
Normalized positional residue frequency at helix termini N4*(Aurora-Rose, 1998)	322	62.6333	60.6815	61.3334	59.517	60.5175	57.7866	58.2664	56.5032	55.4617
Normalized positional residue frequency at helix termini N5 (Aurora-Rose, 1998)	323	62.3175	60.2591	61.3703	60.0377	60.3247	57.3643	58.0982	57.5324	55.8512
Normalized positional residue frequency at helix termini Nc (Aurora-Rose, 1998)	324	62.7932	60.8578	61.3047	59.8163	60.2755	57.1757	58.25	55.9948	55.9989
Normalized relative frequency of alpha-helix (Isogai et al., 1980)	325	62.4856	60.6651	61.4319	60.2919	60.2673	57.5324	58.1639	57.9096	55.7405

Table C.15: (Page 15 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Normalized relative frequency of bend (Isogai et al., 1980)	326	63.3344	61.3867	61.6328	61.1899	60.6569	60.7266	59.6687	57.3848	58.291
Normalized relative frequency of bend R (Isogai et al., 1980)	327	63.1581	61.2063	61.4113	60.2755	57.0773	58.4591	49.9754	54.0635	58.4017
Normalized relative frequency of bend S (Isogai et al., 1980)	328	62.6005	60.4437	61.4524	58.6846	59.5867	56.4007	56.2777	54.74	54.8589
Normalized relative frequency of coil (Isogai et al., 1980)	329	62.5103	60.3822	61.4483	57.4176	58.0449	56.577	51.2711	50.857	54.9615
Normalized relative frequency of double bend (Isogai et al., 1980)	330	62.8137	60.5831	61.6205	59.0003	60.0705	54.0184	57.0157	55.1665	53.8913
Normalized relative frequency of extended structure (Isogai et al., 1980)	331	63.0556	61.2842	61.5344	59.8737	60.5872	56.0973	59.1971	55.7569	55.0312
Normalized relative frequency of helix end (Isogai et al., 1980)	332	62.4774	60.456	61.2924	58.8896	59.2176	56.3556	55.2116	54.1578	55.5232
Normalized van der Waals volume (Fauchere et al., 1988)	333	62.5513	60.4396	61.5754	59.0085	60.948	55.3756	60.5995	55.5314	56.085
Number of full nonbonding orbitals (Fauchere et al., 1988)	334	63.4615	60.8168	61.6	59.5211	57.6759	55.5683	57.0567	56.3761	54.4448
Number of hydrogen bond donors (Fauchere et al., 1988)	335	63.523	60.6774	61.9116	59.8081	57.9834	53.3008	57.5078	55.5642	54.4817
Optical rotation (Fasman, 1976)	336	62.5062	60.1976	61.7066	58.7584	56.4376	54.3792	49.9508	52.4848	54.7277
Optimal matching hydrophobicity (Sweet-Eisenberg, 1983)	337	65.0279	62.9613	63.3836	62.9244	61.4278	59.0946	59.0085	58.9265	57.6841
Optimized average non-bonded energy per atom (Oobatake et al., 1985)	338	62.5964	60.7963	61.6615	59.9762	60.4724	55.1624	58.4509	56.2777	54.9123
Optimized beta-structure-coil equilibrium constant (Oobatake et al., 1985)	339	63.4984	61.2063	62.0428	60.9234	61.1448	57.4955	59.0331	56.6467	56.4253
Optimized propensity to form reverse turn (Oobatake et al., 1985)	340	62.7276	60.9111	61.7312	59.6359	60.9439	55.7364	60.0541	56.6836	55.72
Optimized relative partition energies - method A (Miyazawa-Jernigan, 1999)	341	65.6101	62.9367	63.5845	63.1622	61.5713	60.1689	59.5703	59.5416	58.3607
Optimized relative partition energies - method B (Miyazawa-Jernigan, 1999)	342	65.5199	62.7563	63.5394	63.4615	61.2432	59.8491	58.9183	59.6482	58.7461
Optimized relative partition energies - method C (Miyazawa-Jernigan, 1999)	343	65.5978	62.8793	63.8552	63.2975	61.6205	60.0746	58.8158	59.1561	58.2131
Optimized relative partition energies - method D (Miyazawa-Jernigan, 1999)	344	65.7372	62.5841	63.5969	63.3508	60.3945	60.4232	58.9347	59.0372	58.1557
Optimized side chain interaction parameter (Oobatake et al., 1985)	345	62.7727	60.8988	61.5426	59.6605	61.1448	56.7246	59.2053	56.4991	54.9533
Optimized transfer energy parameter (Oobatake et al., 1985)	346	62.9818	60.6528	61.6369	59.5047	59.8778	53.4197	55.9004	55.6544	54.2603
PRIFT index (Cornette et al., 1987)	347	65.2083	62.6989	63.3221	62.6497	61.2801	59.5908	58.9962	58.9101	58.824
PRILS index (Cornette et al., 1987)	348	65.2862	62.9121	63.4246	62.8629	61.7722	59.0372	59.2176	59.2997	59.2464
Partial specific volume (Cohn-Edsall, 1943)	349	63.3098	61.0259	62.1166	60.8168	61.2924	56.8517	60.5667	57.1552	55.4781
Partition coefficient (Garel et al., 1973)	350	63.4533	61.5672	62.2724	59.6974	61.1858	55.5683	59.5293	54.4161	58.25
Partition coefficient (Pliska et al., 1981)	351	64.8844	62.1658	63.5353	62.4733	63.236	60.173	62.3339	58.5739	57.8153

Table C.16: (Page 16 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Partition energy (Guy, 1985)	352	65.0689	62.9244	63.6707	62.6907	61.2637	59.4391	58.824	59.6605	57.7702
Percentage of buried residues (Janin et al., 1978)	353	64.4538	62.0961	62.908	61.6287	60.9726	60.3617	58.4304	57.5037	55.6749
Percentage of exposed residues (Janin et al., 1978)	354	63.9577	61.0464	61.5057	60.8373	60.333	58.291	57.639	56.3474	55.3428
Polar requirement (Woese, 1973)	355	64.2529	62.1658	62.9203	61.7025	61.7681	60.3822	60.0336	56.8722	58.0326
Polarity (Grantham, 1974)	356	65.0484	62.338	63.3754	62.8219	62.8957	60.8414	61.3539	57.7825	58.5903
Polarity (Zimmerman et al., 1968)	357	63.5681	61.1448	61.9362	57.2864	59.0577	53.0835	56.7738	55.2198	54.2849
Polarizability parameter (Charton-Charton, 1982)	358	62.58	60.6077	61.7189	59.3489	60.8947	56.0809	60.825	56.0973	56.167
Positive charge (Fauchere et al., 1988)	359	62.5964	60.0541	61.071	55.7979	53.6985	51.308	56.9993	53.7805	53.7805
Principal component I (Sneath, 1966)	360	62.949	60.989	61.4893	60.173	60.2468	57.5898	57.8604	55.7774	56.1629
Principal component II (Sneath, 1966)	361	62.8916	60.8045	61.3703	59.2997	59.7179	57.1183	56.4827	56.9911	55.6503
Principal component III (Sneath, 1966)	362	62.6989	60.5339	61.6533	59.4473	60.4888	52.2552	57.6472	55.6667	54.1045
Principal component IV (Sneath, 1966)	363	62.5677	60.1484	61.5672	58.3484	60.6487	52.5627	58.3935	56.167	54.4448
Principal property value z1 (Wold et al., 1987)	364	65.2657	63.1909	63.8183	62.8055	61.928	59.9352	59.0126	58.7789	58.3853
Principal property value z2 (Wold et al., 1987)	365	62.5021	59.9926	61.5713	58.2787	59.6564	53.8913	53.92	55.0353	54.3218
Principal property value z3 (Wold et al., 1987)	366	62.4487	60.3412	61.5672	59.8901	60.7348	56.3597	57.6718	57.2331	55.2526
Propensity of amino acids within pi-helices (Fodje-Al-Karadaghi, 2002)	367	62.8916	60.8168	61.805	60.5257	60.4519	57.6062	57.6882	55.7692	56.3556
Propensity to be buried inside (Wertz-Scheraga, 1978)	368	64.9172	62.1453	63.277	62.8875	62.2888	58.5903	60.7676	58.3279	57.3274
Proportion of residues 100% buried (Chothia, 1976)	369	63.6174	60.9726	61.9854	60.0869	59.3817	56.6508	56.6959	56.1424	54.7605
Proportion of residues 95% buried (Chothia, 1976)	370	64.7327	62.1084	62.8465	61.6656	60.9931	59.7015	58.4427	58.0121	56.5483
RF rank (Zimmerman et al., 1968)	371	64.5891	61.8173	63.0187	62.297	62.1822	58.9593	59.722	58.5821	55.9948
RF value in high salt chromatography (Weber-Lacey, 1978)	372	62.6743	60.989	61.9731	60.2591	60.3822	56.3515	59.3981	54.5391	55.3387
Radius of gyration of side chain (Levitt, 1976)	373	62.6128	60.0336	61.3867	58.7092	60.8291	56.0973	59.9721	54.1291	55.5273
Ratio of average and computed composition (Nakashima et al., 1990)	374	62.5841	60.3371	61.4729	59.0167	58.5452	56.4212	51.3244	54.3054	55.1993
Ratio of buried and accessible molar fractions (Janin, 1979)	375	64.7614	62.1453	62.7604	61.3744	60.9808	59.6359	57.3848	56.4991	56.2244
Refractivity (McMeekin et al., 1964), Cited by Jones (1975)	376	62.5308	60.7348	61.5467	59.1479	60.7184	56.823	60.4314	57.4012	55.7241
Relative frequency in alpha-helix (Prabhakaran, 1990)	377	62.58	60.5749	61.3662	59.8737	60.3986	57.3684	57.9096	56.4991	55.9701
Relative frequency in beta-sheet (Prabhakaran, 1990)	378	63.2729	61.4934	61.764	61.0095	60.7102	57.3807	59.1397	56.3843	56.1916
Relative frequency in reverse-turn (Prabhakaran, 1990)	379	63.2729	61.3498	61.4113	60.9521	60.0582	60.7512	57.7743	56.8763	57.6513

Table C.17: (Page 17 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	#48 0.2,100,3
Relative frequency of occurrence (Jones et al., 1992)	380	62.5226	60.4109	61.5262	59.4309	60.4724	56.4212	59.1274	55.2813	53.5222
Relative mutability (Dayhoff et al., 1978b)	381	62.8055	60.5257	61.1817	59.4965	60.6897	55.8922	60.0131	56.1547	54.0225
Relative mutability (Jones et al., 1992)	382	62.539	60.3781	61.276	58.7912	60.7143	54.4817	58.8445	54.822	53.5468
Relative partition energies derived by the Bethe approximation (Miyazawa-Jernigan, 1999)	383	65.6265	63.0023	63.8962	63.0597	61.9444	60.1935	59.4596	59.4965	59.152
Relative population of conformational state A (Vasquez et al., 1983)	384	62.5513	60.4478	61.6205	58.2705	60.8537	53.8093	59.3612	55.5683	53.514
Relative population of conformational state C (Vasquez et al., 1983)	385	62.5882	60.4437	61.6041	58.6518	60.2263	56.5155	57.4422	55.2239	54.5022
Relative population of conformational state E (Vasquez et al., 1983)	386	62.5513	60.497	61.4688	59.2381	59.7507	57.0444	56.4048	54.7031	54.3833
Relative preference value at C" (Richardson-Richardson, 1988)	387	62.6661	60.3781	61.3908	59.9229	60.5872	54.0676	58.4591	55.5478	54.4612
Relative preference value at C* (Richardson-Richardson, 1988)	388	62.7153	60.9316	61.3129	59.6318	58.6518	56.2654	55.2936	55.3182	55.0312
Relative preference value at C-cap (Richardson-Richardson, 1988)	389	62.4897	60.1771	61.2842	56.8927	53.5263	52.0912	48.897	50.3731	54.7195
Relative preference value at C1 (Richardson-Richardson, 1988)	390	62.5759	60.3371	61.2309	59.1438	59.8614	56.782	56.7738	55.597	56.3884
Relative preference value at C2 (Richardson-Richardson, 1988)	391	62.7194	60.6774	61.7476	59.1766	60.2632	56.9173	58.7748	56.9009	55.8594
Relative preference value at C3 (Richardson-Richardson, 1988)	392	62.4815	60.7143	61.5385	59.8737	60.4683	56.9665	59.4186	56.987	55.7569
Relative preference value at C4 (Richardson-Richardson, 1988)	393	63.3426	61.4524	62.0469	61.1161	60.8701	58.8978	59.9967	57.3356	57.6185
Relative preference value at C5 (Richardson-Richardson, 1988)	394	62.4815	60.1607	61.436	59.5703	60.9398	57.311	60.214	57.1347	56.5237
Relative preference value at Mid (Richardson-Richardson, 1988)	395	62.5431	61.1612	61.4031	60.0008	60.2509	57.1593	57.9383	57.4504	56.7533
Relative preference value at N" (Richardson-Richardson, 1988)	396	62.5349	60.3494	61.4934	58.66	58.9224	55.4043	58.008	54.982	54.2562
Relative preference value at N* (Richardson-Richardson, 1988)	397	62.5349	60.3494	61.4934	58.66	58.9224	55.4043	58.008	54.982	54.2562
Relative preference value at N-cap (Richardson-Richardson, 1988)	398	62.7891	60.6241	61.4524	59.5252	60.5421	57.4217	58.1844	57.2905	57.3561
Relative preference value at N1 (Richardson-Richardson, 1988)	399	62.5554	60.6118	61.3621	59.0331	58.8773	55.7118	54.1332	53.7805	55.023
Relative preference value at N2 (Richardson-Richardson, 1988)	400	62.7768	60.3453	61.4113	59.4391	58.8978	55.5601	54.5924	55.1419	54.1619
Relative preference value at N3 (Richardson-Richardson, 1988)	401	62.7276	60.2509	61.317	58.9265	56.6262	56.5278	51.267	51.1276	54.0471
Relative preference value at N4 (Richardson-Richardson, 1988)	402	63.359	61.5385	61.8296	61.4647	60.0623	59.1561	58.6231	56.1137	56.6262
Relative preference value at N5 (Richardson-Richardson, 1988)	403	62.6743	60.4314	61.4729	59.0249	59.9352	55.7282	58.9634	56.4335	54.9
Residue accessible surface area in folded protein (Chothia, 1976)	404	63.7814	61.5631	61.846	60.3617	60.5708	54.0635	58.0572	56.2449	55.4863
Residue accessible surface area in tripeptide (Chothia, 1976)	405	62.5841	60.4601	61.4072	58.8322	60.415	56.0071	58.9716	54.863	55.7569

Table C.18: (Page 18 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Residue volume (Bigelow, 1967)	406	62.6661	60.7594	61.7107	59.5703	60.9111	56.7287	60.3535	54.412	55.9373
Residue volume (Goldsack-Chalifoux, 1973)	407	62.6374	60.6323	61.7476	59.4473	60.6774	57.1101	60.0951	54.6211	56.0973
Retention coefficient at pH 2 (Guo et al., 1986)	408	65.0115	63.1253	62.8219	62.5144	60.9603	59.0782	58.9429	59.0331	57.7989
Retention coefficient in HFBA (Browne et al., 1982)	409	63.8675	61.6	62.3421	61.0915	60.4355	57.6062	57.6144	57.4586	55.2813
Retention coefficient in HPLC, pH2.1 (Meek, 1980)	410	64.4046	61.7394	62.8178	61.5508	61.3088	57.9096	57.7415	57.7415	55.4781
Retention coefficient in HPLC, pH7.4 (Meek, 1980)	411	63.7773	61.2309	61.9772	60.4191	59.3612	57.7538	56.9829	57.2126	55.0189
Retention coefficient in NaClO4 (Meek-Rossetti, 1981)	412	64.8393	62.4118	63.2237	61.8993	61.7025	57.7497	58.049	57.8071	58.2664
Retention coefficient in NaH2PO4 (Meek-Rossetti, 1981)	413	64.6999	62.3585	63.0187	61.7271	61.317	56.7082	57.8194	57.7989	57.5734
Retention coefficient in TFA (Browne et al., 1982)	414	64.1504	61.5385	62.4446	61.03	60.4355	55.6462	57.4955	56.6303	56.6959
SD of AA composition of total proteins (Nakashima et al., 1990)	415	62.5841	60.0131	61.3211	59.3612	60.6036	55.802	59.7097	55.6626	53.9979
STERIMOL length of the side chain (Fauchere et al., 1988)	416	62.5841	60.1771	61.7148	59.1807	60.3781	58.2049	59.2258	55.3346	54.0553
STERIMOL maximum width of the side chain (Fauchere et al., 1988)	417	62.5718	60.1484	61.5713	58.9183	60.4724	55.556	60.2058	54.6457	53.7436
STERIMOL minimum width of the side chain (Fauchere et al., 1988)	418	62.6415	61.03	61.1612	58.6026	51.3613	53.514	51.9354	52.9441	55.0025
SWEIG index (Cornette et al., 1987)	419	64.9992	63.0064	63.4328	62.8957	61.6246	59.3571	58.9142	58.9839	57.7128
Scaled side chain hydrophobicity values (Black-Mould, 1991)	420	64.8885	62.6333	63.1704	62.2683	62.1986	59.8122	59.2258	58.5534	57.7087
Screening coefficients gamma, local (Avbelj, 2000)	421	63.1089	60.8209	61.3211	54.9861	60.05	56.9542	57.9342	56.1342	55.6626
Screening coefficients gamma, non-local (Avbelj, 2000)	422	62.7686	60.8004	61.0915	54.1537	58.5288	57.3479	52.2921	52.4561	55.0312
Sequence frequency (Jungck, 1978)	423	62.58	60.4683	61.682	58.9716	61.0341	56.4376	60.5995	54.5022	53.6985
Short and medium range non-bonded energy per atom (Oobatake-Ooi, 1977)	424	62.6005	59.9803	61.2596	59.5006	59.9639	55.5683	55.4986	55.7405	53.8462
Short and medium range non-bonded energy per residue (Oobatake-Ooi, 1977)	425	62.6046	60.4437	61.4975	59.3366	60.7512	57.2782	58.5247	55.7118	55.5396
Side chain angle theta(AAR) (Levitt, 1976)	426	62.4897	59.845	59.6195	56.5606	57.0485	53.3623	49.8483	49.8483	55.9989
Side chain hydropathy, corrected for solvation (Roseman, 1988)	427	64.8475	62.4241	63.1376	61.6861	61.0382	59.1151	58.2377	58.3689	57.352
Side chain hydropathy, uncorrected for solvation (Roseman, 1988)	428	64.8475	62.256	62.9695	61.7066	61.0997	58.3689	57.8932	58.0244	57.6923
Side chain interaction parameter (Krigbaum-Komoriya, 1979)	429	64.421	61.7722	63.0802	62.5677	62.2068	59.1192	60.5749	58.4755	57.7087
Side chain interaction parameter (Krigbaum-Rubin, 1971)	430	63.9249	61.9608	62.4815	61.436	62.1002	58.4345	60.9316	58.049	57.8358
Side chain orientational preference (Rackovsky-Scheraga, 1977)	431	64.2324	62.1986	62.2929	61.4934	60.9029	60.009	57.9752	57.3561	57.4627
Side chain torsion angle phi(AAAR) (Levitt, 1976)	432	62.4405	60.296	61.6	58.4304	57.7989	56.8763	50.1722	51.2342	54.2562
Side chain volume (Krigbaum-Komoriya, 1979)	433	62.5923	60.7676	61.2186	59.5252	60.866	56.7369	60.3699	56.3187	56.0194

Table C.19: (Page 19 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Side-chain conformation by gaussian evolutionary method (Yang et al., 2002)	434	62.5308	60.8783	61.6656	59.4801	60.4642	57.0773	59.8819	55.2239	55.0804
Side-chain contribution to protein stability (kJ/mol) (Takano-Yutani, 2001)	435	63.765	61.8296	62.6497	60.7348	61.4606	58.5288	59.2874	57.4914	57.4709
Signal sequence helical potential (Argos et al., 1982)	436	64.3226	61.8173	61.9075	60.8168	60.0377	58.1433	57.3274	56.4909	57.0896
Size (Dawson, 1972)	437	62.6456	60.4519	61.2022	59.1192	59.0987	55.3551	58.6887	54.5514	54.0389
Slope in regression analysis x 1.0E1 (Prabhakaran-Ponnuswamy, 1982)	438	62.5185	60.5339	61.7107	59.4186	60.6446	56.5237	59.6974	53.5304	55.4207
Slopes decapeptide, FDPB VFF neutral (Avbelj, 2000)	439	62.5718	60.3289	61.2924	55.6339	59.6482	55.9332	55.1009	55.1747	54.8384
Slopes proteins, FDPB VFF neutral (Avbelj, 2000)	440	62.5923	60.3453	61.1612	55.9332	59.111	56.8558	54.3013	54.6785	54.863
Slopes tripeptide FDPB PARSE neutral (Avbelj, 2000)	441	62.5882	60.5913	60.8947	53.7231	59.0905	56.9255	55.4945	54.2603	54.5842
Slopes tripeptide FDPB VFF all (Avbelj, 2000)	442	62.461	60.4888	60.8332	53.2557	59.6933	56.4663	55.1747	55.1091	54.0307
Slopes tripeptide, FDPB VFF neutral (Avbelj, 2000)	443	62.4159	60.5011	61.5508	53.4074	58.4304	57.4176	54.9984	55.0599	54.8671
Slopes tripeptide, FDPB VFF noside (Avbelj, 2000)	444	62.4323	60.4806	60.8127	53.2147	58.8486	57.0485	54.1947	54.4202	54.74
Slopes tripeptides, LD VFF neutral (Avbelj, 2000)	445	62.4118	60.8332	61.0341	57.5037	55.8389	56.9706	52.2101	52.9031	55.6339
Smoothed epsilon steric parameter (Fauchere et al., 1988)	446	62.5718	60.6815	61.4811	59.3202	59.4678	58.9429	55.7733	54.9082	56.905
Solvation free energy (Eisenberg-McLachlan, 1986)	447	64.7081	62.7686	63.2893	62.3462	61.194	60.333	58.7789	59.1561	57.7866
Spin-spin coupling constants 3JH α -NH (Bundi-Wuthrich, 1979)	448	62.5308	60.1115	61.2186	59.1192	55.3551	55.597	50.5371	50.4592	55.4289
Steric parameter (Charton, 1981)	449	62.7153	60.8045	61.2883	58.8855	59.6605	58.3771	55.5314	54.6252	56.4294
Surface composition of amino acids in extracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)	450	63.9249	61.5385	62.6374	61.3088	62.0018	58.4427	60.1279	58.25	56.9009
Surface composition of amino acids in intracellular proteins of mesophiles (percent) (Fukuchi-Nishikawa, 2001)	451	63.7609	61.7107	61.8419	60.8865	60.9398	56.1055	57.803	55.6421	56.4786
Surface composition of amino acids in intracellular proteins of thermophiles (percent) (Fukuchi-Nishikawa, 2001)	452	63.1417	60.6364	61.8296	59.5252	59.3858	55.8758	54.2685	54.3915	54.6047
Surface composition of amino acids in nuclear proteins (percent) (Fukuchi-Nishikawa, 2001)	453	63.9864	61.8747	61.8009	60.661	60.8578	57.0567	57.3479	56.2982	56.9665
Surrounding hydrophobicity in alpha-helix (Ponnuswamy et al., 1980)	454	62.9777	60.6897	61.6205	60.0869	60.2099	58.7051	55.3018	56.0891	55.1173
Surrounding hydrophobicity in beta-sheet (Ponnuswamy et al., 1980)	455	63.2934	60.7758	61.4442	60.9439	59.3899	54.2234	56.823	56.2244	54.9205

Table C.20: (Page 20 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Surrounding hydrophobicity in folded form (Ponnuswamy et al., 1980)	456	64.503	62.1412	63.3057	62.5759	61.6246	60.1156	60.9644	57.885	57.5775
Surrounding hydrophobicity in turn (Ponnuswamy et al., 1980)	457	63.7691	61.4113	62.8629	61.6041	61.0464	59.599	60.1566	56.823	56.7697
TOTFT index (Cornette et al., 1987)	458	65.196	62.6948	63.8429	63.072	61.7107	57.9547	59.1356	59.3776	58.5452
TOTLS index (Cornette et al., 1987)	459	65.196	62.744	63.8716	63.2196	61.6451	58.3402	59.1069	59.5252	58.291
The Chou-Fasman parameter of the coil conformation (Charton-Charton, 1983)	460	63.113	61.0054	61.477	61.0177	60.2673	60.3822	58.8691	57.188	57.8932
The Kerr-constant increments (Khanarian-Moore, 1980)	461	62.5144	60.2304	61.5467	59.0126	60.2263	54.289	57.7251	54.8302	53.8872
The number of atoms in the side chain labelled 1+1 (Charton-Charton, 1983)	462	62.4897	60.8291	61.2063	58.6108	56.4909	55.4781	59.2669	55.3756	53.9364
The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)	463	62.4938	60.3904	61.2104	58.8117	57.4012	57.7333	58.2008	54.4776	53.5181
The number of atoms in the side chain labelled 3+1 (Charton-Charton, 1983)	464	62.4897	60.0664	61.1735	58.1762	55.1665	54.8466	58.0326	56.0194	53.6944
The number of bonds in the longest chain (Charton-Charton, 1983)	465	62.5513	60.3289	61.1981	58.496	59.8778	56.3474	60.05	56.8189	53.8339
The relative stability scale extracted from mutation experiments (Zhou-Zhou, 2004)	466	63.7035	61.6984	62.1248	61.3744	61.4196	57.2372	60.6651	56.3638	57.028
The stability scale from the knowledge-based atom-atom potential (Zhou-Zhou, 2004)	467	64.1299	61.9567	62.8383	61.8255	61.9608	58.2623	59.8122	57.475	57.8932
Thermodynamic beta sheet propensity (Kim-Berg, 1993)	468	62.6415	61.0136	61.682	58.3853	58.6641	58.2131	52.5217	52.9892	57.1306
Transfer energy, organic solvent/water (Nozaki-Tanford, 1971)	469	64.421	62.7071	62.3544	61.8173	60.1525	57.2167	58.5329	57.7579	57.7948
Transfer free energy (Janin, 1979)	470	64.6137	62.0346	62.297	60.7922	59.193	58.6395	56.8517	56.5606	56.5155
Transfer free energy (Simon, 1976), Cited by Charton-Charton (1982)	471	63.5845	61.071	62.4159	60.2755	61.2678	58.1105	58.8035	57.3192	55.0148
Transfer free energy from chx to oct (Radzicka-Wolfenden, 1988)	472	64.2406	61.3088	61.9403	60.7512	59.804	58.7256	56.7861	56.29	56.0932
Transfer free energy from chx to wat (Radzicka-Wolfenden, 1988)	473	64.8967	62.2601	62.6948	61.8009	61.1489	59.9557	58.168	57.8973	58.5206
Transfer free energy from oct to wat (Radzicka-Wolfenden, 1988)	474	64.3759	62.4528	62.5062	62.2314	60.9234	58.3238	58.3935	58.5247	58.1844
Transfer free energy from vap to chx (Radzicka-Wolfenden, 1988)	475	62.6251	60.3001	61.3334	58.8978	60.4314	56.2285	57.5447	56.1219	53.9774
Transfer free energy from vap to oct (Radzicka-Wolfenden, 1988)	476	63.0556	60.5503	61.3498	59.7507	59.8942	57.0362	57.2372	55.6995	55.269
Transfer free energy to lipophilic phase (von Heijne-Blomberg, 1979)	477	64.4087	62.2232	62.9572	61.5754	60.7553	57.4299	58.1351	57.885	57.0691
Transfer free energy to surface (Bull-Breese, 1974)	478	64.6711	62.7686	62.5964	62.133	61.2432	57.8399	58.4755	58.0449	57.5693
Transfer free energy, CHP/water (Lawson et al., 1984)	479	63.5928	61.3498	61.5959	61.1899	60.255	59.2381	55.1542	57.2495	56.1711
Transmembrane regions of mt-proteins (Nakashima et al., 1990)	480	63.683	60.8209	61.723	59.8942	59.4555	59.0413	56.2982	54.9574	55.1747

Table C.21: (Page 21 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Transmembrane regions of non-mt-proteins (Nakashima et al., 1990)	481	63.9454	61.7476	62.2642	60.6364	60.6733	56.7123	58.3525	57.0485	56.0153
Turn propensity scale for transmembrane helices (Monne et al., 1999)	482	63.9946	61.5672	62.4446	61.4524	59.8327	57.1183	59.3243	57.3971	55.679
Unfolding Gibbs energy in water, pH7.0 (Yutani et al., 1987)	483	63.683	61.9157	62.3831	61.2719	60.9152	57.7743	58.6682	56.9091	56.0481
Unfolding Gibbs energy in water, pH9.0 (Yutani et al., 1987)	484	63.3508	61.0464	61.5877	60.3822	60.456	57.9014	57.7784	55.8471	55.2157
Value of theta(i) (Rackovsky-Scheraga, 1982)	485	62.5308	60.5175	61.7599	58.9265	60.0295	56.5565	54.7769	53.555	55.9619
Value of theta(i-1) (Rackovsky-Scheraga, 1982)	486	62.379	60.4273	61.5016	58.9019	58.0203	55.6831	53.0507	53.7354	56.2654
Volume (Grantham, 1974)	487	62.5841	60.7512	61.5139	59.804	60.743	56.4909	60.1443	56.5975	56.0891
Volumes including the crystallographic waters using the ProtOr (Tsai et al., 1999)	488	62.6743	60.6323	61.7066	59.5129	60.8414	56.7738	60.4355	54.8589	56.0563
Volumes not including the crystallographic waters using the ProtOr (Tsai et al., 1999)	489	62.6292	60.7594	61.7435	59.5457	60.8742	56.7041	60.4478	54.7933	56.1096
Weights for alpha-helix at the window position of -1 (Qian-Sejnowski, 1988)	490	62.5513	60.4601	61.3498	59.7753	59.1274	57.6349	56.2203	56.0973	56.1916
Weights for alpha-helix at the window position of -2 (Qian-Sejnowski, 1988)	491	62.5103	60.7348	61.6451	60.1607	60.1525	57.2003	57.5406	57.3684	56.2572
Weights for alpha-helix at the window position of -3 (Qian-Sejnowski, 1988)	492	62.5636	60.6487	61.4565	59.5539	60.497	56.1055	57.7046	56.7082	56.331
Weights for alpha-helix at the window position of -4 (Qian-Sejnowski, 1988)	493	62.379	60.4314	61.5344	58.4796	59.3899	56.2326	55.802	55.7405	54.9328
Weights for alpha-helix at the window position of -5 (Qian-Sejnowski, 1988)	494	62.5759	60.3001	61.5426	58.3484	60.009	52.9974	57.7866	55.7897	53.4033
Weights for alpha-helix at the window position of -6 (Qian-Sejnowski, 1988)	495	62.4815	60.3781	61.4688	59.111	60.2509	55.3141	58.0654	56.9378	53.9569
Weights for alpha-helix at the window position of 0 (Qian-Sejnowski, 1988)	496	62.6292	60.7061	61.723	60.7758	60.0664	58.291	57.5611	58.3443	56.9173
Weights for alpha-helix at the window position of 1 (Qian-Sejnowski, 1988)	497	62.6825	60.5421	61.4113	60.0295	56.1834	58.1926	50.5577	53.7518	55.966
Weights for alpha-helix at the window position of 2 (Qian-Sejnowski, 1988)	498	62.5595	60.5462	61.6328	60.132	57.4791	57.7661	52.8457	53.719	56.1014
Weights for alpha-helix at the window position of 3 (Qian-Sejnowski, 1988)	499	62.5759	60.3945	61.477	60.2427	59.0126	57.0198	55.4617	56.5565	56.249
Weights for alpha-helix at the window position of 4 (Qian-Sejnowski, 1988)	500	62.826	60.9521	61.7148	61.1366	59.5375	56.7041	55.2321	56.6467	56.5811
Weights for alpha-helix at the window position of 5 (Qian-Sejnowski, 1988)	501	62.6538	60.8168	61.436	60.5872	59.2464	57.3889	56.5401	57.2413	55.6052

Table C.22: (Page 22 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	#48 0.2,100,3
Weights for alpha-helix at the window position of 6 (Qian-Sejnowski, 1988)	502	62.5923	60.4396	61.5467	59.4801	60.4519	55.4576	59.5375	56.0358	54.8425
Weights for beta-sheet at the window position of -1 (Qian-Sejnowski, 1988)	503	63.7691	61.7558	62.6579	61.2596	61.235	57.4627	59.4883	58.9429	56.6631
Weights for beta-sheet at the window position of -2 (Qian-Sejnowski, 1988)	504	63.1212	61.2719	61.9075	60.3535	60.1648	56.8353	57.3889	56.1137	54.9082
Weights for beta-sheet at the window position of -3 (Qian-Sejnowski, 1988)	505	62.539	60.1361	61.5303	58.9757	60.5421	55.0845	58.1474	55.9127	54.5268
Weights for beta-sheet at the window position of -4 (Qian-Sejnowski, 1988)	506	63.3877	60.9644	62.0387	60.2837	60.3986	57.0773	59.0331	57.4627	55.3223
Weights for beta-sheet at the window position of -5 (Qian-Sejnowski, 1988)	507	63.6707	60.9644	62.5062	60.5667	61.0464	59.394	60.2222	58.9962	56.618
Weights for beta-sheet at the window position of -6 (Qian-Sejnowski, 1988)	508	63.5107	60.9726	62.1166	60.0705	61.071	58.1433	59.0905	57.9793	55.2649
Weights for beta-sheet at the window position of 0 (Qian-Sejnowski, 1988)	509	63.8429	61.5467	62.7809	61.5754	61.8501	57.9711	60.0664	59.2628	57.6964
Weights for beta-sheet at the window position of 1 (Qian-Sejnowski, 1988)	510	63.7035	61.7353	62.4282	61.2063	61.6697	59.0331	59.6605	59.3448	56.9296
Weights for beta-sheet at the window position of 2 (Qian-Sejnowski, 1988)	511	64.0971	61.9198	62.539	61.1858	60.7922	57.8686	57.5898	58.0408	57.4299
Weights for beta-sheet at the window position of 3 (Qian-Sejnowski, 1988)	512	62.6538	60.4437	61.4237	58.9921	60.6651	56.6426	57.5734	55.5765	54.8671
Weights for beta-sheet at the window position of 4 (Qian-Sejnowski, 1988)	513	62.5308	60.3083	61.7476	58.6723	60.4724	57.147	59.0249	57.4381	54.0471
Weights for beta-sheet at the window position of 5 (Qian-Sejnowski, 1988)	514	62.7645	60.5831	61.4196	58.8117	60.1525	53.3541	57.8276	57.2167	54.3792
Weights for beta-sheet at the window position of 6 (Qian-Sejnowski, 1988)	515	62.7891	60.5749	61.6779	59.8286	60.8168	56.8107	59.8081	57.028	56.1752
Weights for coil at the window position of -1 (Qian-Sejnowski, 1988)	516	63.7486	61.477	62.7358	60.989	60.5708	60.0336	57.9916	58.2582	58.2869
Weights for coil at the window position of -2 (Qian-Sejnowski, 1988)	517	64.6014	62.0346	62.785	62.3339	61.4155	59.3407	59.6195	59.8327	58.5288
Weights for coil at the window position of -3 (Qian-Sejnowski, 1988)	518	62.8752	60.7225	61.3293	59.9598	60.5052	56.413	59.3694	58.2582	57.2331
Weights for coil at the window position of -4 (Qian-Sejnowski, 1988)	519	62.5226	60.2263	61.5631	59.3653	60.5052	56.1629	59.1602	56.7205	54.1209

Table C.23: (Page 23 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

Property	Index of Property	logistic	RandomForest 225, 5	DTNB	IBK 60, 1/	BayesNet	MLP 82, 0.1	NaiveBayes	RBFNetwork	48 0.2,100,3
Weights for coil at the window position of -5 (Qian-Sejnowski, 1988)	520	62.9367	60.5421	61.3047	59.4924	59.5949	54.3792	57.0855	55.3223	54.1168
Weights for coil at the window position of -6 (Qian-Sejnowski, 1988)	521	62.6661	60.5749	61.3129	59.029	60.9603	57.9178	59.3735	56.9419	54.6211
Weights for coil at the window position of 0 (Qian-Sejnowski, 1988)	522	63.9905	61.1161	62.5677	60.9849	61.2268	60.8127	59.0208	59.1971	58.0613
Weights for coil at the window position of 1 (Qian-Sejnowski, 1988)	523	64.6629	61.8419	62.9531	61.0956	59.7917	61.0956	54.6457	56.2203	59.2956
Weights for coil at the window position of 2 (Qian-Sejnowski, 1988)	524	64.4087	61.7066	62.5472	61.2104	60.415	59.9434	56.1588	57.8358	58.7051
Weights for coil at the window position of 3 (Qian-Sejnowski, 1988)	525	62.6743	60.702	61.5221	60.5298	58.1269	57.2044	53.8585	55.7651	55.3961
Weights for coil at the window position of 4 (Qian-Sejnowski, 1988)	526	62.5513	60.5544	61.6656	59.9516	57.3151	55.1501	53.0794	55.4125	54.9082
Weights for coil at the window position of 5 (Qian-Sejnowski, 1988)	527	62.3708	60.1935	61.682	59.2464	59.886	56.946	56.126	56.1875	54.3669
Weights for coil at the window position of 6 (Qian-Sejnowski, 1988)	528	62.5554	60.4396	61.4031	59.2094	60.2017	53.9979	58.7133	57.0321	53.9159
Weights from the IFH scale (Jacobs-White, 1989)	529	64.5973	62.0223	62.8055	60.6077	60.8988	59.2381	57.3438	56.9706	57.5406
Zimm-Bragg parameter s at 20 C (Sueki et al., 1984)	530	62.8793	61.1243	61.4811	61.2965	60.3001	61.6287	58.7338	57.7128	57.6308
Zimm-Bragg parameter $\sigma \times 1.0E4$ (Sueki et al., 1984)	531	62.6169	60.009	61.3785	58.5452	60.009	55.064	60.0008	54.9984	54.0512
alpha-CH chemical shifts (Andersen et al., 1992)	532	62.4897	60.8004	61.4196	58.7174	58.5862	56.29	47.5726	53.1081	55.1706
alpha-CH chemical shifts (Bundi-Wuthrich, 1979)	533	62.4528	60.3904	61.7599	58.2213	58.25	55.7323	45.7192	53.2311	54.9123
alpha-NH chemical shifts (Bundi-Wuthrich, 1979)	534	62.5472	60.4068	61.6205	58.4673	53.1245	54.7974	47.7448	48.5977	55.31
p-Values of mesophilic proteins based on the distributions of B values (Parthasarathy-Murthy, 2000)	535	63.8634	62.0223	62.662	62.1822	62.1125	60.3494	61.3006	59.111	58.1844
p-Values of thermophilic proteins based on the distributions of B values (Parthasarathy-Murthy, 2000)	536	63.2032	61.4811	61.5508	60.9398	60.6241	59.6769	58.3238	54.8835	56.7123
pK (-COOH) (Jones, 1975)	537	62.5718	60.173	61.0587	59.5211	60.0459	57.0732	57.8481	54.2562	54.3095
pK-C (Fasman, 1976)	538	62.8793	60.7471	61.6082	60.3863	60.2919	57.393	53.838	55.5888	54.3751
pK-N (Fasman, 1976)	539	62.58	60.6323	61.559	59.8409	56.3351	58.0039	48.3557	53.0179	55.7528
pK-a(RCOOH) (Fauchere et al., 1988)	540	62.5677	60.4888	61.6328	59.5006	56.6754	57.0198	50.3403	51.2629	56.2203
van der Waals parameter R_0 (Levitt, 1976)	541	63.1294	61.3375	62.0633	60.6323	60.9644	57.7702	59.4719	56.4335	58.9757
van der Waals parameter epsilon (Levitt, 1976)	542	62.8137	60.5134	61.1653	59.8696	61.235	54.289	59.5375	57.1716	54.5596

Table C.24: (Page 24 of 24) These are Q_3 accuracies achieved by various machine-learning models using the arff files generated using the attributes shown in table B (and in table 4.2) on all properties tested, as explained in section 4.1.4

	CostSensitive(H_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	CostSensitive(H_2.0)_Logistic_344_noInd.txt	CostSensitive(H_2.0)_Logistic_86.txt	CostSensitive(H_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	CostSensitive(N_1.2)_Logistic_344_noInd.txt	CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	CostSensitive(N_2.0)_Logistic_86.txt	D'NB_109.txt	IBK_60_w_173.txt	IBK_60_w_344.txt	Logistic_0.txt
AttributeSelected_Bagging_15_RBF_7_56.txt												
BayesNet_109.txt												
BayesNet_351.txt												
BayesNet_356.txt												
BayesNet_53.txt												
BayesNet_56.txt												
BayesNet_57.txt												
CostSensitive(E_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt												
CostSensitive(E_2.0)_Logistic_344_noInd.txt												
CostSensitive(E_2.0)_Logistic_86.txt												
CostSensitive(E_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt												
CostSensitive(E_2.0)_RF_225_5_383.txt												
CostSensitive(H_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	1											
CostSensitive(H_2.0)_Logistic_344_noInd.txt	0.95646	1										
CostSensitive(H_2.0)_Logistic_86.txt	0.96177	0.99769	1									
CostSensitive(H_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0.96986	0.96579	0.96424	1								
CostSensitive(N_1.2)_Logistic_344_noInd.txt	0.91136	0.92804	0.92366	0.91641	1							
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0.89577	0.88265	0.87861	0.93615	0.95935	1						
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0.93921	0.87609	0.87743	0.89485	0.95707	0.96654	1					
CostSensitive(N_2.0)_Logistic_86.txt	0.77594	0.76057	0.76365	0.75875	0.96006	0.93259	0.94432	1				
DTNB_109.txt	0.93593	0.93237	0.9297	0.95215	0.96977	0.97007	0.95351	0.90897	1			
IBK_60_w_173.txt	0.79344	0.82193	0.80617	0.80416	0.89859	0.88581	0.88205	0.89929	0.87573	1		
IBK_60_w_344.txt	0.7988	0.83269	0.8162	0.80553	0.90766	0.88591	0.87837	0.89943	0.87504	0.98177	1	
Logistic_0.txt	0.93938	0.9571	0.95518	0.94486	0.99355	0.95469	0.95221	0.924	0.97258	0.87777	0.88435	1
Logistic_1.txt	0.92954	0.95397	0.94828	0.93828	0.99135	0.9461	0.94279	0.92186	0.96587	0.86894	0.87738	0.99621
Logistic_109.txt	0.94317	0.95365	0.95313	0.94476	0.99084	0.95645	0.95644	0.92308	0.97324	0.87103	0.87353	0.99598
Logistic_137.txt	0.94446	0.95204	0.95284	0.94188	0.98792	0.94989	0.9555	0.92131	0.97005	0.86301	0.86998	0.9954
Logistic_150.txt	0.93888	0.94978	0.94725	0.93721	0.98972	0.94916	0.95037	0.92471	0.96912	0.86238	0.87199	0.99049
Logistic_18.txt	0.93914	0.95427	0.95056	0.93974	0.98715	0.94744	0.95033	0.91875	0.96831	0.86636	0.87084	0.99528
Logistic_194.txt	0.9401	0.9565	0.95325	0.9453	0.99435	0.9552	0.95238	0.92325	0.97318	0.88113	0.88396	0.99716
Logistic_195.txt	0.93747	0.95043	0.95006	0.93572	0.99043	0.94775	0.95204	0.92797	0.96768	0.87313	0.88005	0.99504
Logistic_196.txt	0.93997	0.95189	0.95413	0.93885	0.98816	0.94711	0.95254	0.92593	0.96751	0.8627	0.8691	0.99484
Logistic_197.txt	0.94062	0.95168	0.95313	0.93858	0.98834	0.94646	0.9532	0.92594	0.96688	0.86538	0.87511	0.99534
Logistic_222.txt	0.94111	0.95409	0.95162	0.94251	0.99174	0.95454	0.95541	0.9224	0.97276	0.88098	0.88019	0.9966
Logistic_344.txt	0.93769	0.95968	0.95332	0.94373	0.99641	0.95352	0.94933	0.92388	0.97303	0.87724	0.8881	0.99719
Logistic_347.txt	0.93548	0.95561	0.94849	0.94039	0.99043	0.94872	0.94657	0.91895	0.96917	0.87199	0.87769	0.99452
Logistic_348.txt	0.93701	0.95515	0.9489	0.93924	0.9903	0.94805	0.94891	0.91949	0.96847	0.87161	0.87675	0.99518
Logistic_352.txt	0.93786	0.95235	0.94839	0.94107	0.98834	0.95085	0.95048	0.92171	0.97097	0.86885	0.86759	0.9928
Logistic_356.txt	0.935	0.94725	0.94371	0.93436	0.98802	0.94667	0.94724	0.92232	0.96629	0.86243	0.87386	0.98997
Logistic_364.txt	0.94619	0.94805	0.9506	0.93942	0.98617	0.9466	0.95806	0.92257	0.96881	0.86036	0.86611	0.99152
Logistic_383.txt	0.94422	0.9561	0.95384	0.94676	0.99285	0.9554	0.95553	0.92153	0.97264	0.87686	0.88294	0.99768
Logistic_408.txt	0.94367	0.94865	0.95043	0.93855	0.98583	0.94759	0.95721	0.92386	0.96863	0.85623	0.86067	0.99062
Logistic_458.txt	0.93894	0.9563	0.9516	0.94289	0.99031	0.95006	0.94997	0.91803	0.97103	0.87243	0.878	0.99701
Logistic_459.txt	0.9396	0.95617	0.95203	0.94266	0.9894	0.94948	0.95068	0.91773	0.9702	0.87024	0.87499	0.99667
Logistic_86.txt	0.93976	0.95221	0.95774	0.93912	0.9897	0.9496	0.95379	0.93075	0.96932	0.86391	0.87155	0.99475
Logistic_98.txt	0.92655	0.94125	0.94057	0.92857	0.97652	0.935	0.93696	0.91667	0.95863	0.84252	0.84164	0.98305
LogitBoost_285_DecisionStump_18.txt	0.93363	0.94604	0.94309	0.9374	0.98399	0.95085	0.94933	0.9246	0.96744	0.87119	0.87607	0.99058
LogitBoost_285_DecisionStump_344.txt	0.92936	0.94942	0.94458	0.93728	0.99298	0.95267	0.94616	0.92546	0.97038	0.88231	0.89625	0.99233
MAX_RF_225_5(63.5148).txt	0.9355	0.92856	0.92759	0.93315	0.91278	0.89221	0.89365	0.80351	0.91971	0.85536	0.86419	0.93133
MLP_H62_53.txt	0.8123	0.81735	0.80941	0.82129	0.8924	0.86793	0.86341	0.86506	0.86709	0.84504	0.84035	0.88379
MultiBoost_10_BayesNet_351.txt	0.96996	0.9124	0.91708	0.9447	0.93515	0.91793	0.96049	0.82893	0.95311	0.79485	0.79508	0.95235
MultiBoost_10_MLP_H62_56.txt	0.77525	0.78448	0.772	0.78984	0.88359	0.86505	0.85502	0.86513	0.85341	0.86045	0.85558	0.86649
MultiBoost_15_BayesNet_356.txt	0.94568	0.91018	0.9042	0.93284	0.93422	0.90168	0.92396	0.82067	0.94262	0.78815	0.79873	0.94544
MultiClassClassifier_BayesNet_351.txt	0.976	0.92525	0.92752	0.94691	0.94232	0.93046	0.97125	0.85294	0.95912	0.81754	0.8203	0.95695
MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0.95244	0.94399	0.94215	0.98888	0.9566	0.97735	0.9385	0.85257	0.97283	0.83317	0.83669	0.97034
MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0.98553	0.93192	0.93593	0.95156	0.95611	0.944	0.98138	0.87659	0.96213	0.833	0.83355	0.96338
MultiClassClassifier_MultiBoost_BayesNet_351.txt	0.98939	0.93102	0.93526	0.95064	0.95636	0.94397	0.98225	0.87929	0.96175	0.87323	0.83944	0.96753
RF_225_10_137.txt	0.83313	0.82964	0.82959	0.816	0.91792	0.90132	0.91557	0.93039	0.88832	0.90597	0.90628	0.89999
RF_225_5_137.txt	0.79978	0.7999	0.79736	0.78711	0.91203	0.9099	0.91702	0.94059	0.88278	0.93437	0.93507	0.88576
RF_225_5_173.txt	0.82815	0.83738	0.83268	0.82275	0.92327	0.90683	0.9116	0.92898	0.89363	0.92092	0.91625	0.90172
RF_225_5_173noInd.txt	0.79667	0.81228	0.80498	0.79427	0.91798	0.91263	0.90728	0.93577	0.88538	0.94041	0.93343	0.89009
RF_225_5_364.txt	0.80081	0.79165	0.79162	0.78243	0.9108	0.90729	0.91957	0.94509	0.88007	0.92927	0.93053	0.88114
RF_225_5_383.txt	0.79608	0.80361	0.79783	0.79404	0.9186	0.91745	0.91457	0.94305	0.88955	0.94435	0.94401	0.89096
RF_225_5_408.txt	0.80449	0.799	0.80027	0.7905	0.9185	0.91505	0.92125	0.94953	0.88901	0.93231	0.93014	0.88652
RF_225_5_419.txt	0.7948	0.80187	0.79724	0.78166	0.91577	0.90423	0.91193	0.94255	0.87971	0.92594	0.93018	0.88735
ZeroR173.txt	0.01426	-0.02077	-0.03008	-0.02408	0.50884	0.54544	0.57443	0.82786	0.36253	0.63011	0.61432	0.35506

Table D.2: (Page 2 of 6) These are the pairwise Yule's Q statistics for all combinations of two classifiers selected from the 66 models in the classifier set for the first round of majority-vote ensembles, as explained in section 4.2.1.2.

	Logistic_1.txt	Logistic_109.txt	Logistic_137.txt	Logistic_150.txt	Logistic_18.txt	Logistic_194.txt	Logistic_195.txt	Logistic_196.txt	Logistic_197.txt	Logistic_222.txt	Logistic_344.txt	Logistic_347.txt
AttributeSelected_Bagging_15_RBF_7_56.txt												
BayesNet_109.txt												
BayesNet_351.txt												
BayesNet_356.txt												
BayesNet_53.txt												
BayesNet_56.txt												
BayesNet_57.txt												
CostSensitive(E_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt												
CostSensitive(E_2.0)_Logistic_344_nolnd.txt												
CostSensitive(E_2.0)_Logistic_86.txt												
CostSensitive(E_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt												
CostSensitive(E_2.0)_RF_225_5_383.txt												
CostSensitive(H_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt												
CostSensitive(H_2.0)_Logistic_344_nolnd.txt												
CostSensitive(H_2.0)_Logistic_86.txt												
CostSensitive(H_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt												
CostSensitive(N_1.2)_Logistic_344_nolnd.txt												
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt												
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt												
CostSensitive(N_2.0)_Logistic_86.txt												
DTNB_109.txt												
IBK_60_w_173.txt												
IBK_60_w_344.txt												
Logistic_0.txt												
Logistic_1.txt	1											
Logistic_109.txt	0.99024	1										
Logistic_137.txt	0.98912	0.99346	1									
Logistic_150.txt	0.98851	0.98966	0.9931	1								
Logistic_18.txt	0.99048	0.99331	0.99209	0.98731	1							
Logistic_194.txt	0.99421	0.99544	0.99312	0.99288	0.99116	1						
Logistic_195.txt	0.99162	0.99267	0.99234	0.99062	0.99025	0.9936	1					
Logistic_196.txt	0.989	0.99366	0.99406	0.98811	0.99154	0.99097	0.99354	1				
Logistic_197.txt	0.99156	0.99284	0.9948	0.98948	0.99347	0.9906	0.9964	0.99669	1			
Logistic_222.txt	0.99209	0.99509	0.99297	0.99084	0.99155	0.99828	0.99469	0.99011	0.99102	1		
Logistic_344.txt	0.995	0.99512	0.99236	0.99384	0.99164	0.99817	0.99351	0.99228	0.99227	0.99568	1	
Logistic_347.txt	0.99211	0.99279	0.99046	0.98573	0.99331	0.99116	0.99029	0.99104	0.99207	0.99157	0.99406	1
Logistic_348.txt	0.99175	0.99321	0.99262	0.98634	0.99414	0.99103	0.99103	0.99163	0.99358	0.99227	0.99371	0.99936
Logistic_352.txt	0.98696	0.99328	0.98716	0.98946	0.98964	0.99446	0.99134	0.98771	0.98804	0.99636	0.99223	0.9892
Logistic_356.txt	0.99005	0.98781	0.99127	0.99273	0.9847	0.99058	0.98987	0.98553	0.98849	0.98982	0.99152	0.98621
Logistic_364.txt	0.98601	0.99187	0.99714	0.99396	0.98834	0.99159	0.9928	0.99057	0.99285	0.99245	0.99006	0.98569
Logistic_383.txt	0.99269	0.99663	0.99604	0.99199	0.99404	0.99684	0.99513	0.99395	0.99528	0.99599	0.99687	0.99354
Logistic_408.txt	0.98456	0.99445	0.99501	0.99017	0.98924	0.98995	0.99014	0.99383	0.99223	0.99029	0.9901	0.98897
Logistic_458.txt	0.99356	0.99474	0.99286	0.98787	0.99794	0.99297	0.99132	0.99248	0.99409	0.99331	0.99461	0.99824
Logistic_459.txt	0.99289	0.99468	0.99302	0.98717	0.99832	0.99236	0.99113	0.9929	0.99446	0.9931	0.99387	0.99743
Logistic_86.txt	0.99021	0.99547	0.99351	0.99036	0.99045	0.9938	0.99491	0.99649	0.99579	0.99256	0.99403	0.98879
Logistic_98.txt	0.98572	0.98422	0.98108	0.97923	0.98321	0.9799	0.98526	0.9816	0.98615	0.98111	0.98004	0.98618
LogitBoost_285_DecisionStump_18.txt	0.9857	0.98912	0.98712	0.98225	0.99543	0.98642	0.98579	0.98642	0.98862	0.98645	0.98702	0.98851
LogitBoost_285_DecisionStump_344.txt	0.98985	0.9904	0.98679	0.98771	0.98569	0.99281	0.98792	0.98684	0.98618	0.99023	0.99553	0.98828
MAX_RF_225_5(63.5148).txt	0.91973	0.92734	0.9347	0.92625	0.92306	0.92745	0.92539	0.92577	0.93065	0.92539	0.92782	0.92523
MLP_H62_53.txt	0.87635	0.88516	0.87974	0.87994	0.87793	0.88726	0.88426	0.87609	0.87708	0.88844	0.88211	0.87452
MultiBoost_10_BayesNet_351.txt	0.94179	0.95753	0.95775	0.95318	0.9508	0.95397	0.95008	0.95175	0.9517	0.95601	0.95018	0.94794
MultiBoost_10_MLP_H62_56.txt	0.85417	0.86865	0.85476	0.86063	0.85577	0.87431	0.86229	0.84936	0.85373	0.87362	0.86723	0.85016
MultiBoost_15_BayesNet_356.txt	0.94447	0.94442	0.94695	0.95147	0.93657	0.94745	0.94408	0.94043	0.94259	0.94588	0.94874	0.9406
MultiClassClassifier_BayesNet_351.txt	0.94884	0.96712	0.96189	0.95805	0.95602	0.95815	0.95554	0.9567	0.9562	0.96051	0.9545	0.95676
MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0.96265	0.97262	0.96702	0.96394	0.96527	0.96966	0.96319	0.96454	0.96389	0.96898	0.96886	0.96516
MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0.95921	0.97271	0.97212	0.96611	0.96603	0.96773	0.96601	0.96806	0.96806	0.96987	0.96459	0.96331
MultiClassClassifier_MultiBoost_BayesNet_351.txt	0.95587	0.97203	0.97122	0.9652	0.96563	0.96653	0.96573	0.96733	0.96785	0.96904	0.96387	0.96295
RF_225_10_137.txt	0.89305	0.89957	0.90469	0.90054	0.89532	0.89725	0.89977	0.89683	0.90001	0.89617	0.89895	0.89742
RF_225_5_137.txt	0.87328	0.87992	0.88937	0.887	0.87588	0.87889	0.88135	0.8767	0.88018	0.87878	0.88143	0.88
RF_225_5_408.txt	0.89549	0.90076	0.89628	0.89591	0.89409	0.90332	0.89965	0.89184	0.89456	0.90169	0.9025	0.89947
RF_225_5_173nolnd.txt	0.88147	0.88894	0.87973	0.8835	0.8786	0.89145	0.88621	0.87653	0.87823	0.89111	0.89057	0.88421
RF_225_5_364.txt	0.87018	0.8778	0.88484	0.88595	0.87308	0.87976	0.88486	0.87365	0.87931	0.88039	0.87933	0.86933
RF_225_5_383.txt	0.87828	0.88684	0.88518	0.88472	0.88225	0.88731	0.88732	0.88182	0.88548	0.88657	0.89201	0.88471
RF_225_5_408.txt	0.87622	0.88914	0.88751	0.88872	0.8805	0.88447	0.88632	0.88348	0.88567	0.88422	0.88768	0.88309
RF_225_5_419.txt	0.87695	0.88058	0.88822	0.88716	0.88218	0.88253	0.8882	0.88187	0.88744	0.88239	0.88722	0.88476
ZeroR173.txt	0.36112	0.35326	0.35148	0.36651	0.35133	0.35448	0.37348	0.35636	0.35897	0.35614	0.35729	0.35637

Table D.3: (Page 3 of 6) These are the pairwise Yule’s Q statistics for all combinations of two classifiers selected from the 66 models in the classifier set for the first round of majority-vote ensembles, as explained in section 4.2.1.2.

	Logistic_348.txt	Logistic_352.txt	Logistic_356.txt	Logistic_364.txt	Logistic_383.txt	Logistic_408.txt	Logistic_458.txt	Logistic_459.txt	Logistic_86.txt	Logistic_98.txt	LogitBoost_285_DecisionStump_18.txt	LogitBoost_285_DecisionStump_344.txt
AttributeSelected_Bagging_15_RBF_7_56.txt												
BayesNet_109.txt												
BayesNet_351.txt												
BayesNet_356.txt												
BayesNet_53.txt												
BayesNet_56.txt												
BayesNet_57.txt												
CostSensitive(E_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt												
CostSensitive(E_2.0)_Logistic_344_noInd.txt												
CostSensitive(E_2.0)_Logistic_86.txt												
CostSensitive(E_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt												
CostSensitive(E_2.0)_RF_225_5_383.txt												
CostSensitive(H_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt												
CostSensitive(H_2.0)_Logistic_344_noInd.txt												
CostSensitive(H_2.0)_Logistic_86.txt												
CostSensitive(H_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt												
CostSensitive(N_1.2)_Logistic_344_noInd.txt												
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt												
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt												
CostSensitive(N_2.0)_Logistic_86.txt												
DTNB_109.txt												
IBK_60_w_173.txt												
IBK_60_w_344.txt												
Logistic_0.txt												
Logistic_1.txt												
Logistic_109.txt												
Logistic_137.txt												
Logistic_150.txt												
Logistic_18.txt												
Logistic_194.txt												
Logistic_195.txt												
Logistic_196.txt												
Logistic_197.txt												
Logistic_222.txt												
Logistic_344.txt												
Logistic_347.txt												
Logistic_348.txt	1											
Logistic_352.txt	0.98918	1										
Logistic_356.txt	0.98698	0.98523	1									
Logistic_364.txt	0.98789	0.98772	0.99261	1								
Logistic_383.txt	0.99446	0.99255	0.99074	0.99451	1							
Logistic_408.txt	0.99008	0.98732	0.98626	0.99554	0.99428	1						
Logistic_458.txt	0.99839	0.99098	0.98695	0.98814	0.99553	0.98976	1					
Logistic_459.txt	0.99816	0.9915	0.98666	0.98842	0.99529	0.99024	0.99976	1				
Logistic_86.txt	0.98945	0.99091	0.98828	0.99293	0.99524	0.99428	0.99062	0.99112	1			
Logistic_98.txt	0.98523	0.98152	0.98003	0.98142	0.98171	0.98259	0.98534	0.98477	0.98131	1		
LogitBoost_285_DecisionStump_18.txt	0.98903	0.98457	0.98083	0.98279	0.98914	0.98444	0.99314	0.99323	0.98577	0.9784	1	
LogitBoost_285_DecisionStump_344.txt	0.988	0.98665	0.9877	0.98405	0.99192	0.9837	0.98889	0.98805	0.98849	0.97204	0.98434	1
MAX_RF_225_5(63.5148).txt	0.9265	0.91477	0.92449	0.93003	0.93249	0.92351	0.92628	0.92573	0.92524	0.90131	0.9239	0.925
MLP_H62_53.txt	0.87579	0.88661	0.87598	0.88218	0.88413	0.87501	0.87584	0.87785	0.87703	0.86848	0.88117	0.88588
MultiBoost_10_BayesNet_351.txt	0.94913	0.95125	0.94758	0.95995	0.95657	0.95919	0.95058	0.95087	0.95408	0.93481	0.94433	0.94333
MultiBoost_10_MLP_H62_56.txt	0.85158	0.86913	0.8586	0.85833	0.86808	0.85098	0.85529	0.85515	0.85658	0.83254	0.86664	0.87851
MultiBoost_15_BayesNet_356.txt	0.94102	0.93929	0.95549	0.94985	0.9468	0.94054	0.94119	0.94007	0.94485	0.92852	0.93245	0.94315
MultiClassClassifier_BayesNet_351.txt	0.9546	0.95656	0.95294	0.96342	0.95997	0.96246	0.95628	0.95688	0.95792	0.94289	0.95237	0.94976
MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0.96441	0.96681	0.96175	0.96402	0.97131	0.96571	0.96744	0.96683	0.96594	0.9531	0.96329	0.96463
MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0.96502	0.96627	0.96345	0.97314	0.97181	0.97322	0.96574	0.96724	0.96928	0.95492	0.96132	0.95943
MultiClassClassifier_MultiBoost_BayesNet_351.txt	0.9648	0.96581	0.96243	0.97236	0.97103	0.97233	0.96647	0.96692	0.96896	0.95411	0.96131	0.95861
RF_225_10_137.txt	0.89841	0.89097	0.89929	0.90056	0.89914	0.898	0.89646	0.89623	0.8963	0.89022	0.89903	0.89983
RF_225_5_137.txt	0.88118	0.86647	0.88275	0.88548	0.88365	0.87788	0.87932	0.87745	0.87544	0.85208	0.88414	0.88765
RF_225_5_173.txt	0.89925	0.89794	0.89843	0.89262	0.90135	0.89132	0.89729	0.89617	0.89246	0.88722	0.89961	0.90745
RF_225_5_173noInd.txt	0.88384	0.88641	0.88003	0.87445	0.88839	0.87444	0.88232	0.88073	0.88005	0.86239	0.88485	0.8996
RF_225_5_364.txt	0.87225	0.87012	0.88455	0.89254	0.88207	0.8775	0.87173	0.87108	0.87718	0.86068	0.88194	0.88549
RF_225_5_383.txt	0.88504	0.87744	0.88133	0.88388	0.89165	0.87855	0.88465	0.88295	0.88171	0.8574	0.89175	0.8981
RF_225_5_408.txt	0.8851	0.87815	0.88166	0.89013	0.88887	0.89019	0.8818	0.88082	0.88494	0.86769	0.88821	0.89202
RF_225_5_419.txt	0.88748	0.87662	0.88705	0.88574	0.88698	0.87786	0.88397	0.88361	0.88127	0.8647	0.88889	0.89108
ZeroR173.txt	0.3577	0.36047	0.37144	0.3544	0.34765	0.35911	0.34642	0.34542	0.35616	0.38729	0.38364	0.37809

Table D.4: (Page 4 of 6) These are the pairwise Yule's Q statistics for all combinations of two classifiers selected from the 66 models in the classifier set for the first round of majority-vote ensembles, as explained in section 4.2.1.2.

	MAX_RF_225_5(63.5148).txt	MLP_H62_53.txt	MultiBoost_10_BayesNet_351.txt	MultiBoost_10_MLP_H62_56.txt	MultiBoost_15_BayesNet_356.txt	MultiClassClassifier_BayesNet_351.txt	MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	MultiClassClassifier_MultiBoost_BayesNet_351.txt	RF_225_10_137.txt	RF_225_5_137.txt	RF_225_5_173.txt
AttributeSelected_Bagging_15_RBF_7_56.txt												
BayesNet_109.txt												
BayesNet_351.txt												
BayesNet_356.txt												
BayesNet_53.txt												
BayesNet_56.txt												
BayesNet_57.txt												
CostSensitive(E_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt												
CostSensitive(E_2.0)_Logistic_344_noInd.txt												
CostSensitive(E_2.0)_Logistic_86.txt												
CostSensitive(E_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt												
CostSensitive(E_2.0)_RF_225_5_383.txt												
CostSensitive(H_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt												
CostSensitive(H_2.0)_Logistic_344_noInd.txt												
CostSensitive(H_2.0)_Logistic_86.txt												
CostSensitive(H_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt												
CostSensitive(N_1.2)_Logistic_344_noInd.txt												
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt												
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt												
CostSensitive(N_2.0)_Logistic_86.txt												
DTNB_109.txt												
IBK_60_w_173.txt												
IBK_60_w_344.txt												
Logistic_0.txt												
Logistic_1.txt												
Logistic_109.txt												
Logistic_137.txt												
Logistic_150.txt												
Logistic_18.txt												
Logistic_194.txt												
Logistic_195.txt												
Logistic_196.txt												
Logistic_197.txt												
Logistic_222.txt												
Logistic_344.txt												
Logistic_347.txt												
Logistic_348.txt												
Logistic_352.txt												
Logistic_356.txt												
Logistic_364.txt												
Logistic_383.txt												
Logistic_408.txt												
Logistic_458.txt												
Logistic_459.txt												
Logistic_86.txt												
Logistic_98.txt												
LogitBoost_285_DecisionStump_18.txt												
LogitBoost_285_DecisionStump_344.txt												
MAX_RF_225_5(63.5148).txt	1											
MLP_H62_53.txt	0.83077	1										
MultiBoost_10_BayesNet_351.txt	0.91945	0.82061	1									
MultiBoost_10_MLP_H62_56.txt	0.81538	0.91062	0.79776	1								
MultiBoost_15_BayesNet_356.txt	0.91062	0.81718	0.97496	0.80341	1							
MultiClassClassifier_BayesNet_351.txt	0.92975	0.84028	0.99463	0.81664	0.96951	1						
MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0.92782	0.85305	0.96119	0.83638	0.94968	0.96154	1					
MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0.92601	0.85299	0.98673	0.83848	0.95926	0.98771	0.96818	1				
MultiClassClassifier_MultiBoost_BayesNet_351.txt	0.92702	0.85434	0.98682	0.83175	0.95915	0.98914	0.96831	0.99959	1			
RF_225_10_137.txt	0.89918	0.851	0.83051	0.84508	0.8197	0.86362	0.85505	0.87054	0.87183	1		
RF_225_5_137.txt	0.87355	0.84623	0.80129	0.8532	0.78542	0.83703	0.83816	0.85276	0.85669	0.97209	1	
RF_225_5_173.txt	0.88957	0.86578	0.8237	0.86306	0.8166	0.85685	0.85995	0.86613	0.86821	0.9813	0.96257	1
RF_225_5_173noInd.txt	0.86021	0.86249	0.79278	0.87005	0.78424	0.82838	0.84318	0.84615	0.84941	0.97213	0.97379	0.98109
RF_225_5_364.txt	0.85887	0.85421	0.79802	0.86198	0.79142	0.83592	0.83499	0.85395	0.85745	0.96806	0.98389	0.96209
RF_225_5_383.txt	0.86657	0.85473	0.79822	0.86698	0.77934	0.8325	0.84446	0.84955	0.85379	0.96308	0.98225	0.96607
RF_225_5_408.txt	0.86031	0.85246	0.80648	0.85582	0.78388	0.84058	0.84378	0.85736	0.86253	0.96763	0.9803	0.96613
RF_225_5_419.txt	0.86056	0.84857	0.79199	0.85456	0.79579	0.82802	0.83109	0.84478	0.84911	0.96821	0.97787	0.96503
ZeroR173.txt	0.10546	0.49235	0.12868	0.5571	0.12419	0.19661	0.18165	0.25047	0.25952	0.68029	0.76781	0.6784

Table D.5: (Page 5 of 6) These are the pairwise Yule's Q statistics for all combinations of two classifiers selected from the 66 models in the classifier set for the first round of majority-vote ensembles, as explained in section 4.2.1.2.

	RF_225_5_173noInd.txt	RF_225_5_364.txt	RF_225_5_383.txt	RF_225_5_408.txt	RF_225_5_419.txt	ZeroR173.txt
AttributeSelected_Bagging_15_RBF_7_56.txt						
BayesNet_109.txt						
BayesNet_351.txt						
BayesNet_356.txt						
BayesNet_53.txt						
BayesNet_56.txt						
BayesNet_57.txt						
CostSensitive(E_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt						
CostSensitive(E_2.0)_Logistic_344_noInd.txt						
CostSensitive(E_2.0)_Logistic_86.txt						
CostSensitive(E_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt						
CostSensitive(E_2.0)_RF_225_5_383.txt						
CostSensitive(H_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt						
CostSensitive(H_2.0)_Logistic_344_noInd.txt						
CostSensitive(H_2.0)_Logistic_86.txt						
CostSensitive(H_2.0)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt						
CostSensitive(N_1.2)_Logistic_344_noInd.txt						
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt						
CostSensitive(N_1.8)_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt						
CostSensitive(N_2.0)_Logistic_86.txt						
DTNB_109.txt						
IBK_60_w_173.txt						
IBK_60_w_344.txt						
Logistic_0.txt						
Logistic_1.txt						
Logistic_109.txt						
Logistic_137.txt						
Logistic_150.txt						
Logistic_18.txt						
Logistic_194.txt						
Logistic_195.txt						
Logistic_196.txt						
Logistic_197.txt						
Logistic_222.txt						
Logistic_344.txt						
Logistic_347.txt						
Logistic_348.txt						
Logistic_352.txt						
Logistic_356.txt						
Logistic_364.txt						
Logistic_383.txt						
Logistic_408.txt						
Logistic_458.txt						
Logistic_459.txt						
Logistic_86.txt						
Logistic_98.txt						
LogitBoost_285_DecisionStump_18.txt						
LogitBoost_285_DecisionStump_344.txt						
MAX_RF_225_5(63.5148).txt						
MLP_H62_53.txt						
MultiBoost_10_BayesNet_351.txt						
MultiBoost_10_MLP_H62_56.txt						
MultiBoost_15_BayesNet_356.txt						
MultiClassClassifier_BayesNet_351.txt						
MultiClassClassifier_MultiBoost_13_BayesNet_109.txt						
MultiClassClassifier_MultiBoost_13_BayesNet_351.txt						
MultiClassClassifier_MultiBoost_BayesNet_351.txt						
RF_225_10_137.txt						
RF_225_5_137.txt						
RF_225_5_173.txt						
RF_225_5_173noInd.txt	1					
RF_225_5_364.txt	0.97378	1				
RF_225_5_383.txt	0.97892	0.98081	1			
RF_225_5_408.txt	0.9766	0.982	0.98097	1		
RF_225_5_419.txt	0.97324	0.97792	0.97794	0.97496	1	
ZeroR173.txt	0.74913	0.7775	0.76346	0.77572	0.76769	1

Table D.6: (Page 6 of 6) These are the pairwise Yule's Q statistics for all combinations of two classifiers selected from the 66 models in the classifier set for the first round of majority-vote ensembles, as explained in section 4.2.1.2.

Appendix E

(Table E)

Classifier	Times used in ensembles of 3	Times used in ensembles of 5	Times used in ensembles of 7
AttributeSelected_Bagging_15_RBF_7_56.txt	0	91	459
BayesNet_109.txt	0	163	681
BayesNet_351.txt	0	0	1896
BayesNet_356.txt	0	0	1259
BayesNet_53.txt	0	0	494
BayesNet_56.txt	0	0	3454
BayesNet_57.txt	0	0	371
CostSensitive[E_1.8]_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0	44	688
CostSensitive[E_2.0]_Logistic_344_noInd.txt	0	14	635
CostSensitive[E_2.0]_Logistic_86.txt	0	0	386
CostSensitive[E_2.0]_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0	137	2903
CostSensitive[E_2.0]_RF_225_5_383.txt	0	0	1099
CostSensitive[H_1.8]_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0	0	1512
CostSensitive[H_2.0]_Logistic_344_noInd.txt	0	0	2712
CostSensitive[H_2.0]_Logistic_86.txt	0	0	456
CostSensitive[H_2.0]_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	4	83	4691
CostSensitive[N_1.2]_Logistic_344_noInd.txt	0	0	2457
CostSensitive[N_1.8]_MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0	0	2164
CostSensitive[N_1.8]_MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0	8	1820
CostSensitive[N_2.0]_Logistic_86.txt	1	58	1789
DTNB_109.txt	3	190	3725
IBK_60_w_173.txt	0	6	681
IBK_60_w_344.txt	0	5	3379
Logistic_0.txt	0	13	0
Logistic_1.txt	1	25	0
Logistic_109.txt	0	2	0
Logistic_137.txt	0	4	0
Logistic_150.txt	0	9	0
Logistic_18.txt	0	3	0
Logistic_194.txt	0	16	0
Logistic_195.txt	0	9	0
Logistic_196.txt	0	14	0
Logistic_197.txt	0	43	0
Logistic_222.txt	0	8	0
Logistic_344.txt	0	8	0
Logistic_347.txt	1	2	0
Logistic_348.txt	0	1	0
Logistic_352.txt	2	12	0
Logistic_356.txt	0	17	0
Logistic_364.txt	0	7	0
Logistic_383.txt	0	6	0
Logistic_408.txt	0	7	0
Logistic_458.txt	0	14	0
Logistic_459.txt	0	0	0
Logistic_86.txt	0	2	0
Logistic_98.txt	0	14	0
LogitBoost_285_DecisionStump_18.txt	0	0	0
LogitBoost_285_DecisionStump_344.txt	0	0	0
MAX_RF_225_5(63.5148).txt	0	0	0
MLP_H62_53.txt	0	0	0
MultiBoost_10_BayesNet_351.txt	0	0	0
MultiBoost_10_MLP_H62_56.txt	0	0	0
MultiBoost_15_BayesNet_356.txt	0	4	0
MultiClassClassifier_BayesNet_351.txt	0	58	0
MultiClassClassifier_MultiBoost_13_BayesNet_109.txt	0	3	0
MultiClassClassifier_MultiBoost_13_BayesNet_351.txt	0	5	0
MultiClassClassifier_MultiBoost_BayesNet_351.txt	0	44	0
RF_225_10_137.txt	0	9	0
RF_225_5_137.txt	0	23	0
RF_225_5_173.txt	0	78	0
RF_225_5_173noInd.txt	0	11	0
RF_225_5_364.txt	0	0	0
RF_225_5_383.txt	0	0	0
RF_225_5_408.txt	0	0	0
RF_225_5_419.txt	0	0	0
ZeroR173.txt	0	0	0

Table E.1: These are the number of times each classifier was used in ensembles that achieved at least 66% Q_3 accuracy in the first round of majority-vote ensembles, as explained in section 4.2.1.3.

Appendix F

(Table F)

	AttributeSelected_Bagging_15_RRF_7_56.txt	DTNB_109.txt	EVEN_LogitBoost_285_DecisionStump_173_(63.3877).txt	EVEN_MultiBoost_13_BayesNet_(61.5).txt	EVEN_MultiBoost_13_Logistic_(63.3).txt	EVEN_RF_225_5(63.0023).txt	IBK_60_w_344.txt	LogitBoost_285_DecisionStump_173.txt
AttributeSelected_Bagging_15_RRF_7_56.txt	1							
DTNB_109.txt	0.91072	1						
EVEN_LogitBoost_285_DecisionStump_173_(63.3877).txt	0.84788	0.93009	1					
EVEN_MultiBoost_13_BayesNet_(61.5).txt	0.80013	0.90961	0.95008	1				
EVEN_MultiBoost_13_Logistic_(63.3).txt	0.83138	0.92795	0.98985	0.95367	1			
EVEN_RF_225_5(63.0023).txt	0.83409	0.88769	0.92539	0.92027	0.93105	1		
IBK_60_w_344.txt	0.88468	0.87504	0.76627	0.70324	0.75272	0.77359	1	
LogitBoost_285_DecisionStump_173.txt	0.92599	0.97005	0.96477	0.90413	0.94973	0.89099	0.88976	1
LogitBoost_285_DecisionStump_344.txt	0.92095	0.97038	0.95734	0.90219	0.95646	0.89855	0.89625	0.99121
MAX_Logistic_(65.8767).txt	0.92598	0.96812	0.92632	0.88052	0.93002	0.87617	0.90917	0.988
MAX_MultiBoost_13_BayesNet_(64.3021).txt	0.89438	0.96792	0.93228	0.96531	0.92785	0.90005	0.84493	0.96136
MAX_RF_225_5(65.0689).txt	0.91569	0.92167	0.86466	0.85518	0.86674	0.92602	0.899	0.92499
MIP_H62_53.txt	0.86771	0.86709	0.82096	0.7636	0.79745	0.78982	0.84035	0.89177
RAISE_EN_LogitBoost_285_DecisionStump_173_(70.9).txt	0.73627	0.79959	0.92631	0.87797	0.92864	0.85985	0.63592	0.83605
RAISE_EN_MultiBoost_13_BayesNet_(71.3).txt	0.5533	0.70567	0.74073	0.80871	0.75692	0.715	0.42388	0.71198
RAISE_EN_MultiBoost_13_Logistic_(70.2).txt	0.7758	0.8556	0.95023	0.91756	0.9708	0.89689	0.68067	0.88558
RAISE_EN_RF_225_5(70.2).txt	0.7407	0.77863	0.85232	0.85119	0.86451	0.9135	0.65681	0.8005
RAISE_E_LogitBoost_285_DecisionStump_173_(78.3).txt	0.67164	0.79454	0.96035	0.92031	0.9597	0.89197	0.53675	0.84522
RAISE_E_MultiBoost_13_BayesNet_(84.4).txt	0.49187	0.65826	0.8594	0.92082	0.87373	0.84228	0.33675	0.66884
RAISE_E_MultiBoost_13_Logistic_(70.9).txt	0.77077	0.88894	0.98188	0.94886	0.99617	0.9239	0.67413	0.91213
RAISE_E_RF_225_5(86.9).txt	0.45283	0.57248	0.81443	0.82149	0.8337	0.86636	0.32816	0.58895
RAISE_HE_LogitBoost_285_DecisionStump_173_(73).txt	0.51367	0.68962	0.90988	0.85524	0.89178	0.81634	0.39119	0.70659
RAISE_HE_MultiBoost_13_BayesNet_(71.3).txt	0.49276	0.69374	0.85676	0.91728	0.86391	0.81661	0.3681	0.67318
RAISE_HE_MultiBoost_13_Logistic_(73.6).txt	0.49444	0.68698	0.88236	0.85456	0.90797	0.8159	0.3811	0.68467
RAISE_HE_RF_225_5(74.4).txt	0.4082	0.56902	0.78655	0.78448	0.79464	0.82197	0.29915	0.55571
RAISE_HN_LogitBoost_285_DecisionStump_173_(72.1).txt	0.93087	0.95765	0.92598	0.84088	0.89445	0.83228	0.91065	0.98024
RAISE_HN_MultiBoost_13_BayesNet_(71.3).txt	0.90153	0.93643	0.83036	0.87345	0.81571	0.79908	0.87771	0.92767
RAISE_HN_MultiBoost_13_Logistic_(70.4).txt	0.91333	0.96682	0.94664	0.88922	0.94798	0.87804	0.88525	0.98361
RAISE_HN_RF_225_5(74.4).txt	0.89006	0.87361	0.73029	0.6983	0.71432	0.78028	0.91005	0.86728
RAISE_H_LogitBoost_285_DecisionStump_173_(78.1).txt	0.7982	0.89743	0.96397	0.90772	0.94972	0.8796	0.72202	0.91977
RAISE_H_MultiBoost_13_BayesNet_(84.4).txt	0.6967	0.81835	0.84558	0.87085	0.83605	0.79527	0.61621	0.78723
RAISE_H_MultiBoost_13_Logistic_(86.1).txt	0.72401	0.83236	0.88011	0.827	0.86891	0.79658	0.66669	0.82367
RAISE_H_RF_225_5(91.6).txt	0.58116	0.66832	0.71299	0.68691	0.69168	0.70391	0.52848	0.63689
RAISE_N_LogitBoost_285_DecisionStump_173_(90.7).txt	0.85958	0.84845	0.71163	0.67334	0.69519	0.6851	0.86147	0.87922
RAISE_N_MultiBoost_13_BayesNet_(85).txt	0.86574	0.9014	0.79667	0.8337	0.79861	0.80081	0.83107	0.89879
RAISE_N_MultiBoost_13_Logistic_(88.9).txt	0.8897	0.90147	0.77434	0.75102	0.78575	0.75928	0.8861	0.922
RAISE_N_RF_225_5(86).txt	0.87912	0.86652	0.77264	0.76532	0.77307	0.83268	0.86812	0.87793

Table F.1: (Page 1 of 5) These are the pairwise Yule’s Q statistics for all combinations of two classifiers selected from the 37 models in the classifier set for the second round of majority-vote ensembles, as explained in section 4.2.2.1.

	LogitBoost_285_DecisionStump_344.txt	MAX_Logistic_(65.8767).txt	MAX_MultiBoost_13_BayesNet_(64.3021).txt	MAX_RF_225_5(65.0689).txt	MLP_H62_53.txt	RAISE_EN_LogitBoost_285_DecisionStump_173_(70.9).txt	RAISE_EN_MultiBoost_13_BayesNet_(71.3).txt	RAISE_EN_MultiBoost_13_Logistic_(70.2).txt
AttributeSelected_Bagging_15_RBF_7_56.txt								
DTNB_109.txt								
EVEN_LogitBoost_285_DecisionStump_173_(63.3877).txt								
EVEN_MultiBoost_13_BayesNet_(61.5).txt								
EVEN_MultiBoost_13_Logistic_(63.3).txt								
EVEN_RF_225_5(63.0023).txt								
IBK_60_w_344.txt								
LogitBoost_285_DecisionStump_173.txt								
LogitBoost_285_DecisionStump_344.txt	1							
MAX_Logistic_(65.8767).txt	0.99169	1						
MAX_MultiBoost_13_BayesNet_(64.3021).txt	0.95817	0.96026	1					
MAX_RF_225_5(65.0689).txt	0.93133	0.93701	0.92641	1				
MLP_H62_53.txt	0.88588	0.89261	0.85552	0.86704	1			
RAISE_EN_LogitBoost_285_DecisionStump_173_(70.9).txt	0.83194	0.83844	0.84216	0.8256	0.75783	1		
RAISE_EN_MultiBoost_13_BayesNet_(71.3).txt	0.70489	0.71191	0.82874	0.71468	0.58048	0.89653	1	
RAISE_EN_MultiBoost_13_Logistic_(70.2).txt	0.88417	0.88974	0.88834	0.85456	0.77929	0.98755	0.88031	1
RAISE_EN_RF_225_5(70.2).txt	0.80445	0.80427	0.82209	0.87556	0.72924	0.93941	0.87126	0.93629
RAISE_E_LogitBoost_285_DecisionStump_173_(78.3).txt	0.83397	0.78137	0.81428	0.72667	0.66816	0.91809	0.68673	0.94043
RAISE_E_MultiBoost_13_BayesNet_(84.4).txt	0.66202	0.60477	0.75602	0.61023	0.50444	0.8409	0.62731	0.8704
RAISE_E_MultiBoost_13_Logistic_(70.9).txt	0.91855	0.88222	0.89323	0.81994	0.74327	0.92288	0.72318	0.95929
RAISE_E_RF_225_5(86.9).txt	0.59537	0.52504	0.62658	0.6178	0.47132	0.82183	0.59175	0.8418
RAISE_HE_LogitBoost_285_DecisionStump_173_(73).txt	0.70481	0.58793	0.68019	0.53466	0.49962	0.62321	0.21032	0.71525
RAISE_HE_MultiBoost_13_BayesNet_(71.3).txt	0.67363	0.57528	0.75017	0.54028	0.46807	0.59353	0.20179	0.69238
RAISE_HE_MultiBoost_13_Logistic_(73.6).txt	0.70272	0.59285	0.67567	0.52881	0.47101	0.60001	0.18978	0.70261
RAISE_HE_RF_225_5(74.4).txt	0.56979	0.4437	0.57393	0.49473	0.38515	0.47261	0.04004	0.58532
RAISE_HN_LogitBoost_285_DecisionStump_173_(72.1).txt	0.97409	0.97082	0.9343	0.90004	0.8826	0.65346	0.50012	0.73958
RAISE_HN_MultiBoost_13_BayesNet_(71.3).txt	0.9273	0.92704	0.94016	0.87773	0.83485	0.55186	0.44215	0.64819
RAISE_HN_MultiBoost_13_Logistic_(70.4).txt	0.98723	0.98041	0.94973	0.90409	0.86772	0.73582	0.59166	0.80418
RAISE_HN_RF_225_5(74.4).txt	0.87551	0.87858	0.84373	0.87657	0.81741	0.42494	0.21314	0.52584
RAISE_H_LogitBoost_285_DecisionStump_173_(78.1).txt	0.91193	0.85458	0.88171	0.78026	0.76346	0.7458	0.46061	0.80762
RAISE_H_MultiBoost_13_BayesNet_(84.4).txt	0.79173	0.71736	0.80949	0.65511	0.64582	0.47108	0.20506	0.58612
RAISE_H_MultiBoost_13_Logistic_(86.1).txt	0.82956	0.75537	0.79886	0.67281	0.67718	0.52466	0.18033	0.63789
RAISE_H_RF_225_5(91.6).txt	0.64662	0.54205	0.6166	0.51868	0.51957	0.2163	-0.1717	0.34925
RAISE_N_LogitBoost_285_DecisionStump_173_(90.7).txt	0.8719	0.92607	0.85892	0.89337	0.84215	0.86909	0.74656	0.83615
RAISE_N_MultiBoost_13_BayesNet_(85).txt	0.89704	0.92743	0.93944	0.91794	0.83386	0.87954	0.83456	0.88052
RAISE_N_MultiBoost_13_Logistic_(88.9).txt	0.92053	0.96197	0.90545	0.92177	0.86308	0.86641	0.73296	0.87918
RAISE_N_RF_225_5(86).txt	0.87961	0.91065	0.88319	0.95629	0.83638	0.85846	0.78501	0.85654

Table F.2: (Page 2 of 5) These are the pairwise Yule’s Q statistics for all combinations of two classifiers selected from the 37 models in the classifier set for the second round of majority-vote ensembles, as explained in section 4.2.2.1.

	RAISE_EN_RF_225_5(70.2).txt	RAISE_E_LogitBoost_285_DecisionStump_173_(78.3).txt	RAISE_E_MultiBoost_13_BayesNet_(84.4).txt	RAISE_E_MultiBoost_13_Logistic_(70.9).txt	RAISE_E_RF_225_5(86.9).txt	RAISE_HE_LogitBoost_285_DecisionStump_173_(73).txt	RAISE_HE_MultiBoost_13_BayesNet_(71.3).txt	RAISE_HE_MultiBoost_13_Logistic_(73.6).txt
AttributeSelected_Bagging_15_RBF_7_56.txt								
DTNB_109.txt								
EVEN_LogitBoost_285_DecisionStump_173_(63.3877).txt								
EVEN_MultiBoost_13_BayesNet_(61.5).txt								
EVEN_MultiBoost_13_Logistic_(63.3).txt								
EVEN_RF_225_5(63.0023).txt								
IBK_60_w_344.txt								
LogitBoost_285_DecisionStump_173.txt								
LogitBoost_285_DecisionStump_344.txt								
MAX_Logistic_(65.8767).txt								
MAX_MultiBoost_13_BayesNet_(64.3021).txt								
MAX_RF_225_5(65.0689).txt								
MLP_H62_53.txt								
RAISE_EN_LogitBoost_285_DecisionStump_173_(70.9).txt								
RAISE_EN_MultiBoost_13_BayesNet_(71.3).txt								
RAISE_EN_MultiBoost_13_Logistic_(70.2).txt								
RAISE_EN_RF_225_5(70.2).txt	1							
RAISE_E_LogitBoost_285_DecisionStump_173_(78.3).txt	0.86408	1						
RAISE_E_MultiBoost_13_BayesNet_(84.4).txt	0.82154	0.94769	1					
RAISE_E_MultiBoost_13_Logistic_(70.9).txt	0.86407	0.98	0.9175	1				
RAISE_E_RF_225_5(86.9).txt	0.86424	0.92351	0.9346	0.88499	1			
RAISE_HE_LogitBoost_285_DecisionStump_173_(73).txt	0.59176	0.92805	0.89122	0.91873	0.86001	1		
RAISE_HE_MultiBoost_13_BayesNet_(71.3).txt	0.58064	0.88852	0.9393	0.89272	0.85394	0.9647	1	
RAISE_HE_MultiBoost_13_Logistic_(73.6).txt	0.57366	0.91608	0.88711	0.92381	0.85506	0.99168	0.96476	1
RAISE_HE_RF_225_5(74.4).txt	0.51227	0.8369	0.84597	0.83356	0.86787	0.94761	0.94116	0.94973
RAISE_HN_LogitBoost_285_DecisionStump_173_(72.1).txt	0.64441	0.72731	0.49641	0.83599	0.40326	0.61425	0.59571	0.59909
RAISE_HN_MultiBoost_13_BayesNet_(71.3).txt	0.56047	0.5976	0.49623	0.74185	0.32921	0.56418	0.62457	0.56217
RAISE_HN_MultiBoost_13_Logistic_(70.4).txt	0.71146	0.79616	0.61858	0.90941	0.53768	0.73486	0.72216	0.73474
RAISE_HN_RF_225_5(74.4).txt	0.46582	0.44728	0.25346	0.6208	0.23317	0.40888	0.42044	0.4044
RAISE_H_LogitBoost_285_DecisionStump_173_(78.1).txt	0.66651	0.86708	0.76179	0.92892	0.70518	0.93712	0.90147	0.9292
RAISE_H_MultiBoost_13_BayesNet_(84.4).txt	0.45868	0.72958	0.65809	0.81351	0.5779	0.88686	0.9188	0.88728
RAISE_H_MultiBoost_13_Logistic_(86.1).txt	0.46387	0.74529	0.63326	0.83075	0.57175	0.91363	0.87642	0.92617
RAISE_H_RF_225_5(91.6).txt	0.26939	0.59121	0.52263	0.67114	0.48894	0.8419	0.83949	0.84467
RAISE_N_LogitBoost_285_DecisionStump_173_(90.7).txt	0.81429	0.53372	0.3686	0.6161	0.31015	0.1134	0.09208	0.07247
RAISE_N_MultiBoost_13_BayesNet_(85).txt	0.86955	0.65997	0.60013	0.74023	0.49537	0.29878	0.36689	0.28454
RAISE_N_MultiBoost_13_Logistic_(88.9).txt	0.82421	0.59524	0.43739	0.70956	0.37347	0.22833	0.23351	0.24265
RAISE_N_RF_225_5(86).txt	0.90528	0.63003	0.51194	0.71407	0.52911	0.26902	0.28088	0.25453

Table F.3: (Page 3 of 5) These are the pairwise Yule's Q statistics for all combinations of two classifiers selected from the 37 models in the classifier set for the second round of majority-vote ensembles, as explained in section 4.2.2.1.

	RAISE_HE_RF_225_5(74.4).txt	RAISE_HN_LogitBoost_285_DecisionStump_173_(72.1).txt	RAISE_HN_MultiBoost_13_BayesNet_(71.3).txt	RAISE_HN_MultiBoost_13_Logistic_(70.4).txt	RAISE_HN_RF_225_5(74.4).txt	RAISE_H_LogitBoost_285_DecisionStump_173_(78.1).txt	RAISE_H_MultiBoost_13_BayesNet_(84.4).txt	RAISE_H_MultiBoost_13_Logistic_(86.1).txt
AttributeSelected_Bagging_15_RBF_7_56.txt								
DTNB_109.txt								
EVEN_LogitBoost_285_DecisionStump_173_(63.3877).txt								
EVEN_MultiBoost_13_BayesNet_(61.5).txt								
EVEN_MultiBoost_13_Logistic_(63.3).txt								
EVEN_RF_225_5(63.0023).txt								
IBK_60_w_344.txt								
LogitBoost_285_DecisionStump_173.txt								
LogitBoost_285_DecisionStump_344.txt								
MAX_Logistic_(65.8767).txt								
MAX_MultiBoost_13_BayesNet_(64.3021).txt								
MAX_RF_225_5(65.0689).txt								
MLP_H62_53.txt								
RAISE_EN_LogitBoost_285_DecisionStump_173_(70.9).txt								
RAISE_EN_MultiBoost_13_BayesNet_(71.3).txt								
RAISE_EN_MultiBoost_13_Logistic_(70.2).txt								
RAISE_EN_RF_225_5(70.2).txt								
RAISE_E_LogitBoost_285_DecisionStump_173_(78.3).txt								
RAISE_E_MultiBoost_13_BayesNet_(84.4).txt								
RAISE_E_MultiBoost_13_Logistic_(70.9).txt								
RAISE_E_RF_225_5(86.9).txt								
RAISE_HE_LogitBoost_285_DecisionStump_173_(73).txt								
RAISE_HE_MultiBoost_13_BayesNet_(71.3).txt								
RAISE_HE_MultiBoost_13_Logistic_(73.6).txt								
RAISE_HE_RF_225_5(74.4).txt	1							
RAISE_HN_LogitBoost_285_DecisionStump_173_(72.1).txt	0.47929	1						
RAISE_HN_MultiBoost_13_BayesNet_(71.3).txt	0.46875	0.96526	1					
RAISE_HN_MultiBoost_13_Logistic_(70.4).txt	0.6234	0.98429	0.94924	1				
RAISE_HN_RF_225_5(74.4).txt	0.35029	0.93389	0.94327	0.90198	1			
RAISE_H_LogitBoost_285_DecisionStump_173_(78.1).txt	0.85143	0.90228	0.85748	0.94338	0.7673	1		
RAISE_H_MultiBoost_13_BayesNet_(84.4).txt	0.85224	0.82039	0.86657	0.86891	0.75035	0.94863	1	
RAISE_H_MultiBoost_13_Logistic_(86.1).txt	0.84985	0.8554	0.84257	0.90916	0.76013	0.9777	0.96621	1
RAISE_H_RF_225_5(91.6).txt	0.86986	0.71802	0.73966	0.76479	0.69448	0.89164	0.94347	0.94848
RAISE_N_LogitBoost_285_DecisionStump_173_(90.7).txt	-0.07427	0.85136	0.77466	0.79928	0.72732	0.47608	0.2384	0.27309
RAISE_N_MultiBoost_13_BayesNet_(85).txt	0.16129	0.83637	0.8106	0.83832	0.71742	0.61691	0.43898	0.44097
RAISE_N_MultiBoost_13_Logistic_(88.9).txt	0.06478	0.89364	0.83311	0.86856	0.78616	0.59004	0.37871	0.41875
RAISE_N_RF_225_5(86).txt	0.18319	0.82616	0.78239	0.8173	0.74539	0.58346	0.3796	0.40616

Table F.4: (Page 4 of 5) These are the pairwise Yule’s Q statistics for all combinations of two classifiers selected from the 37 models in the classifier set for the second round of majority-vote ensembles, as explained in section 4.2.2.1.

	RAISE_H_RF_225_5(91.6).txt	RAISE_N_LogitBoost_285_DecisionStump_173_(90.7).txt	RAISE_N_MultiBoost_13_BayesNet_(85).txt	RAISE_N_MultiBoost_13_Logistic_(88.9).txt	RAISE_N_RF_225_5(86).txt
AttributeSelected_Bagging_15_RBF_7_56.txt					
DTNB_109.txt					
EVEN_LogitBoost_285_DecisionStump_173_(63.3877).txt					
EVEN_MultiBoost_13_BayesNet_(61.5).txt					
EVEN_MultiBoost_13_Logistic_(63.3).txt					
EVEN_RF_225_5(63.0023).txt					
IBK_60_w_344.txt					
LogitBoost_285_DecisionStump_173.txt					
LogitBoost_285_DecisionStump_344.txt					
MAX_Logistic_(65.8767).txt					
MAX_MultiBoost_13_BayesNet_(64.3021).txt					
MAX_RF_225_5(65.0689).txt					
MLP_H62_53.txt					
RAISE_EN_LogitBoost_285_DecisionStump_173_(70.9).txt					
RAISE_EN_MultiBoost_13_BayesNet_(71.3).txt					
RAISE_EN_MultiBoost_13_Logistic_(70.2).txt					
RAISE_EN_RF_225_5(70.2).txt					
RAISE_E_LogitBoost_285_DecisionStump_173_(78.3).txt					
RAISE_E_MultiBoost_13_BayesNet_(84.4).txt					
RAISE_E_MultiBoost_13_Logistic_(70.9).txt					
RAISE_E_RF_225_5(86.9).txt					
RAISE_HE_LogitBoost_285_DecisionStump_173_(73).txt					
RAISE_HE_MultiBoost_13_BayesNet_(71.3).txt					
RAISE_HE_MultiBoost_13_Logistic_(73.6).txt					
RAISE_HE_RF_225_5(74.4).txt					
RAISE_HN_LogitBoost_285_DecisionStump_173_(72.1).txt					
RAISE_HN_MultiBoost_13_BayesNet_(71.3).txt					
RAISE_HN_MultiBoost_13_Logistic_(70.4).txt					
RAISE_HN_RF_225_5(74.4).txt					
RAISE_H_LogitBoost_285_DecisionStump_173_(78.1).txt					
RAISE_H_MultiBoost_13_BayesNet_(84.4).txt					
RAISE_H_MultiBoost_13_Logistic_(86.1).txt					
RAISE_H_RF_225_5(91.6).txt	1				
RAISE_N_LogitBoost_285_DecisionStump_173_(90.7).txt	-0.00366	1			
RAISE_N_MultiBoost_13_BayesNet_(85).txt	0.17271	0.95965	1		
RAISE_N_MultiBoost_13_Logistic_(88.9).txt	0.14781	0.99155	0.96355	1	
RAISE_N_RF_225_5(86).txt	0.18272	0.95314	0.95212	0.95415	1

Table F.5: (Page 5 of 5) These are the pairwise Yule’s Q statistics for all combinations of two classifiers selected from the 37 models in the classifier set for the second round of majority-vote ensembles, as explained in section 4.2.2.1.

Appendix G

(Table G)

Classifier	Times used in ensembles of 3	Times used in ensembles of 5	Times used in ensembles of 7
AttributeSelected_Bagging_15_RBF_7_56.txt	0	39	2866
DTNB_109.txt	1	56	1172
EVEN_LogitBoost_285_DecisionStump_173_(63.3877).txt	0	21	385
EVEN_MultiBoost_13_BayesNet_(61.5).txt	0	99	1209
EVEN_MultiBoost_13_Logistic_(63.3).txt	0	220	2848
EVEN_RF_225_5(63.0023).txt	1	137	2606
IBK_60_w_344.txt	5	241	4691
LogitBoost_285_DecisionStump_173.txt	0	64	1176
LogitBoost_285_DecisionStump_344.txt	1	128	2027
MAX_Logistic_(65.8767).txt	6	149	4068
MAX_MultiBoost_13_BayesNet_(64.3021).txt	1	20	591
MAX_RF_225_5(65.0689).txt	2	49	1989
MLP_H62_53.txt	1	7	1914
RAISE_EN_LogitBoost_285_DecisionStump_173_(70.9).txt	0	12	883
RAISE_EN_MultiBoost_13_BayesNet_(71.3).txt	0	11	295
RAISE_EN_MultiBoost_13_Logistic_(70.2).txt	0	26	646
RAISE_EN_RF_225_5(70.2).txt	0	10	579
RAISE_E_LogitBoost_285_DecisionStump_173_(78.3).txt	0	18	438
RAISE_E_MultiBoost_13_BayesNet_(84.4).txt	0	12	214
RAISE_E_MultiBoost_13_Logistic_(70.9).txt	0	14	667
RAISE_E_RF_225_5(86.9).txt	0	46	279
RAISE_HE_LogitBoost_285_DecisionStump_173_(73).txt	0	121	346
RAISE_HE_MultiBoost_13_BayesNet_(71.3).txt	0	0	407
RAISE_HE_MultiBoost_13_Logistic_(73.6).txt	0	0	484
RAISE_HE_RF_225_5(74.4).txt	0	0	1311
RAISE_HN_LogitBoost_285_DecisionStump_173_(72.1).txt	0	0	1432
RAISE_HN_MultiBoost_13_BayesNet_(71.3).txt	0	0	23
RAISE_HN_MultiBoost_13_Logistic_(70.4).txt	0	0	429
RAISE_HN_RF_225_5(74.4).txt	0	0	153
RAISE_H_LogitBoost_285_DecisionStump_173_(78.1).txt	0	0	233
RAISE_H_MultiBoost_13_BayesNet_(84.4).txt	0	0	41
RAISE_H_MultiBoost_13_Logistic_(86.1).txt	0	0	83
RAISE_H_RF_225_5(91.6).txt	0	0	128
RAISE_N_LogitBoost_285_DecisionStump_173_(90.7).txt	0	0	558
RAISE_N_MultiBoost_13_BayesNet_(85).txt	0	0	228
RAISE_N_MultiBoost_13_Logistic_(88.9).txt	0	0	588
RAISE_N_RF_225_5(86).txt	0	0	1045

Table G.1: These are the number of times each classifier was used in ensembles that achieved at least 66% Q_3 accuracy in the second round of majority-vote ensembles.