OPTIMIZATION AND ANALYSIS OF WEIGHTED-WINDOW PREDICTORS OF STRUCTURAL DISORDER IN PROTEINS

APPROVED BY SUPERVISORY COMMITTEE

Zbyszek Otwinowski

Jose Rizo-Rey

Alexander Pertsemlidis

Nick Grishin

DEDICATION

To my wife, who is amazing, to my sweet children, to my parents who have encouraged me, and to God, the source of all that is good.

ACKNOWLEDGEMENTS

I thank those who have aided me in my research and in my writing, including Steve Sprang and Mischa Machius. I thank my committee, particularly Dr. Otwinowski, for requiring me to go further. I acknowledge funding provided from the HHMI and through the MSTP. Finally, I thank Nick Grishin and Lisa Kinch for their contributions and suggestions and for helping me develop as a scientist.

OPTIMIZATION AND ANALYSIS OF WEIGHTED-WINDOW PREDICTORS OF STRUCTURAL DISORDER IN PROTEINS

by

NATHAN BRENT HOLLADAY

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

May, 2007

OPTIMIZATION AND ANALYSIS OF WEIGHTED-WINDOW PREDICTORS OF STRUCTURAL DISORDER IN PROTEINS

Publication No. _____

Nathan Brent Holladay

The University of Texas Southwestern Medical Center at Dallas, Graduation Year

Supervising Professor: Nick V. Grishin

X-ray crystallographic protein structures often contain disordered regions that are observed as missing electron density. We have developed single sequence and profile-based weighted-window predictors of structural disorder in proteins, as well as a simple method for addressing disorder-prone chain termini in disorder prediction. Optimizing the parameters for these relatively simple predictors with crystallographic data using a simulated annealing type algorithm, we achieve performance similar to that of DISOPRED2. Optimized parameters from these disorder predictors provide information relating to physical processes underlying crystallographic disorder. Optimized score adjustment values suggest a simple, monotonic relationship between disorder and residue distance from termini that is nearly the same for amino- and carboxy-terminal positions. Residue disorder parameters are strongly associated with scales from certain experimental model systems that primarily reflect hydrophobic interactions. Our data do not suggest a strong association between crystallographic disorder and secondary structure beyond that explained by hydrophobicity. Our results lend support to

the idea that while hydrophobic side chain interactions are primarily involved in determining stability of the folded conformation, hydrogen bonding and similar polar interactions are primarily involved in conformational and interaction specificity.

TABLE OF CONTENTS

Please note that a supplementary workbook is also provided that includes parameter data and parameter to energy calculations

Predictors may be made available at http://prodata.swmed.edu

PRIOR PUBLICATIONS

Crozier, P.S., Rowley, R.L., Holladay, N.B., Henderson, D., and Busath, D.D. 2001. Molecular dynamics simulation of continuous current flow through a model biological membrane channel. *Phys Rev Lett* **86:** 2467-2470.

Holladay, N.B. 2001. Non-equilibrium molecular dynamics of an ion channel with bulk aqueous NaCl in a slab geometry. Brigham Young University.

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

CPU – central processing unit

DNA – deoxyribonucleic acid

HPLC – high-performance liquid chromatography

NMR – nuclear magnetic resonance

PDB – Protein Data Bank (Berman et al. 2000)

PONDR – Predictor of Naturally Disordered Regions (see www.pondr.com)

ROC – receiver operating characteristic

RONN – regional order neural network (Yang et al. 2005)

SCOP – Structural Classification of Proteins (Murzin et al. 1995)

SVM – support vector machine

VFA – very fast annealing [Ingber (1989) is cited by Cai and Shao (2002)]

XML – Extensible Markup Language

CHAPTER ONE
Introduction and Review of Literature

1.1 STRUCTURAL DISORDER IN PROTEINS

Proteins are a class of molecules that perform essential functions in living organisms. Chains of residues of various types of amino acids form the structural basis for proteins. Proteins are well known to often adopt specific folds, giving them form and stability for proper function. But it is also recognized that proteins are dynamic and some are structurally disordered, in part or in whole.

Disorder participates in a variety of physiological functions. It is common in regions in proteins that participate in protein-protein and protein-nucleic acid interactions (Wright and Dyson 1999; Dunker et al. 2001). These regions may undergo a disorder-to-order transition upon binding. Such disorder-to-order transitions may serve both to modify function (Dunker et al. 2002) and to control thermodynamic properties of binding (Huber and Bennett 1983). Linkers between protein domains may be disordered (Zdanov et al. 1994; Jacobs et al. 1999; Dunker et al. 2002). Disordered regions may also serve other functions, such as serving as a barrier or gate (Denning et al. 2003), or enabling protein degradation (Hubbard et al. 1994; Dominguez-Vidal et al. 2006). Cancer-associated and signaling proteins appear to be more likely than eukaryotic proteins in general to contain long disordered regions (Iakoucheva et al. 2002). Enzymatic proteins, on the other hand, do not appear to have such a propensity for disordered regions.

Disorder may also be involved in pathological situations. In hereditary amyloidoses, for example, destabilizing mutations may increase the likelihood of aggregation (Hurle et al.

1994; Raffen et al. 1999). On the other hand, certain mutations in p53 that cause a conditionally disordered DNA-binding region to become persistently ordered also disrupt its cancer-protective cell cycle regulatory function (Wei et al. 2003). Understanding relationships between sequence and disorder may help in formulating hypotheses regarding mechanisms by which point mutations disrupt normal function.

Types of experiments that may detect disorder include, but are not limited to, circular dichroism spectroscopy, limited proteolysis, NMR spectroscopy, and X-ray crystallography (Dunker et al. 2002). Structural disorder may also be predicted through computational methods (see below). Our work primarily considers disorder in protein structures derived through X-ray crystallographic techniques. In X-ray crystallography, electron density maps are derived from diffraction patterns produced by coherent scattering of X-rays by electrons within ordered crystals. These electron density maps and other data (particularly protein sequences) are used to 'solve' protein structures. Even though a structure may be solved for a protein, part of the protein may not yield good electron density, due to lack of consistency in its conformation (disorder) and the resulting incoherence in its scattering of X-rays. These disordered segments are often excluded from published structures, or assigned 'occupancies' of zero. The similarity between crystallographic disorder and disorder determined through other experiments has been called into question. The sample of proteins with deposited crystal structures into the Protein Data Bank (PDB) (Berman et al. 2000) obviously excludes fully unstructured proteins and disordered portions of proteins that have been removed to improve crystallization (Tompa 2002; Linding et al. 2003a).

Ordered protein structure is often depicted with static ribbon diagrams, consisting of helices, strands, and ‘loops’. But disordered protein may, by nature, not necessarily present a neat picture. Different types of disorder may occur. Disorder may be localized to only a part of a protein, or be global, involving the state of the entire protein. Different hypotheses might be put forward as to why a particular region is disordered. It may be due to the region’s being a loop that is well-solubilized; it may be due to the presence of multiple prolines that have significant variation in their isomerization states; it may be that a small, relatively internally consistent domain is oriented in different ways with respect to a crystal’s overall packing.

Although X-ray diffraction data may suggest the presence of disorder, data from a single crystal does not necessarily indicate how a particular disordered protein segment physically behaves (Huber and Bennett 1983). Distinction has been made between ‘dynamic’ disorder and ‘static’ disorder (Huber 1979; Huber and Bennett 1983) in crystal structures, with thermal motion being ‘dynamic’, and positioning in distinct conformations being ‘static’. Conformational differences or flexibility may occur at different levels—individual side chains may be positioned in different ways with respect to the backbone; local polypeptide segments may adopt different backbone conformations; or two domains may adopt different orientations with respect to each other (Huber 1979; Huber and Bennett 1983; Dunker et al. 2001).

In trying to understand disorder, researchers have linked it to various physical characteristics. Disorder has been associated with low sequence complexity, but crystallographically disordered regions often do not have notably low sequence complexity (Romero et al. 2001). Different disordered regions have been associated with various patterns

(Vucetic et al. 2003; Lise and Jones 2005), but these patterns generally appear to relate best to a minor part of disordered regions or do not have a clearly described physical interpretation. Disorder has been associated with various side chain characteristics, including hydropathy and net charge (Uversky et al. 2000), coil propensity (Linding et al. 2003b), coordination number, β-sheet propensity, hydrophobicity (Williams et al. 2001; Dosztányi et al. 2005), and others. Such an association can be demonstrated through straightforward statistics or through showing that the characteristic is useful in predicting disorder.

Simple predictors of disorder, optimized using X-ray crystallographic data, are described in this dissertation. Their data-optimized parameters shed light on the physical nature of disorder.

1.2 HISTORY OF DISORDER-RELATED PREDICTORS

Folds, or shapes, that proteins tend to adopt are primarily determined by their residue sequences (Anfinsen 1973). Efforts to predict various ordered structural elements (globular α-helices, strands, and transmembrane helices, for example) from their sequences have been complemented by efforts to predict where disordered regions occur

Several researchers have approached the problem of predicting disordered regions. Especially since 2003, disorder predictors from various groups have proliferated. But the history of related predictors goes back further. Here are briefly described and discussed several prediction methods that are directly targeted toward or related to prediction of disorder. Secondary structure prediction methods that predict residues as being within either

'helix', 'strand', or 'coil' regions are related, but are not discussed here. The number of disorder predictors has grown, and the list below is not necessarily comprehensive.

1.2.1 Rose hydrophobicity plots

In 1978, George Rose described how turns in protein structures could be predicted as minima on a hydrophobicity plot, where the hydrophobicities are derived from the hydrophobicity scale published by Nozaki and Tanford (Nozaki and Tanford 1971), and plots are smoothed using Savitzky-Golay polynomial smoothing (Savitzky 1964). For similar hydrophobicity plots, peaks were shown to correspond with regions that formed the 'hydrophobic core' of the protein, while 'dominant valleys' were shown to correspond with 'large solvent-exposed loops' of lysozyme.

1.2.2 Hopp-Woods antigen prediction program

Over twenty years ago, Hopp and Woods (1981) described a window-based method, which they indicated was for "predicting the locations of protein antigenic determinants" (Hopp and Woods 1983). The method used a 6-position (evenly weighted) sliding window to average 'hydrophilicity values' assigned to residues by residue type. The hydrophilicity values were largely based on the work of others (Nozaki and Tanford 1971; Levitt 1976). A basic idea of Hopp and Woods' work was that the top peaks in the results of their method would suggest likely antigenic regions of proteins.

1.2.3 Kyte-Doolittle hydropathy plots

Kyte and Doolittle (1982) published a window-based method for generating 'hydropathy' plots. They developed a well-known 'hydropathy' scale with the intent to reflect hydrophobic and hydrophilic properties of residues. For example, in their hydropathy scale, the values of tryptophan and tyrosine are shifted significantly with respect to values of other hydrophobic residues in comparison with hydrophobicity scales such as that of Nozaki and Tanford due to their containing hydrophilic moieties in addition to being hydrophobic. Kyte and Doolittle not only pointed out that their plots reflected interior vs. exterior regions of proteins, but also highlighted the ability of their method to discern transmembrane-spanning segments of proteins.

1.2.4 PONDR (or related predictors)

Dunker and colleagues performed pioneering work in the field of disorder prediction. They developed several disorder predictors (Romero et al. 1997a; Romero et al. 1997b; Li et al. 1999; Obradovic et al. 2003; Vucetic et al. 2003), some or all known as Predictors of Natural Disordered Regions (PONDR®); see www.pondr.com. Most or all are based on various residue 'attributes' (see also http://www.pondr.com/background.html). Some of these predictors may be relatively limited in availability (see http://www.pondr.com). The majority of these predictors are neural network-based. Some of these predictors appear to be lacking in performance (Ward et al. 2004), possibly in large part due to inadequate training data and/or transformation of residue types using attribute scales not ideally associated with

disorder. The recent VSL2 predictors (Peng et al. 2006) are more impressive in their apparent performance.

1.2.5 GlobPlot

Linding et al. (2003b) describe GlobPlot as a "graphical tool" that may be used to attempt to "measure and display the propensity of protein sequences to be ordered or disordered." As of writing, GlobPlot2 is available at http://globplot.embl.de. It plots a smoothed running sum of disorder propensities, as defined by the scale one selects. The default scale, the Russell/Linding scale is a set of 'coil propensities', where coil regions are non-α-helical, non-β-strand regions, "based on the hypothesis that the tendency for disorder can be expressed as P = RC – SS where RC and SS are the propensity for a given amino acid to be in 'random coil' and regular 'secondary structure', respectively." Although there is evidence that GlobPlot can effectively help to locate domains, its efficacy in predicting disordered regions using coil propensities is called into question by results presented here.

1.2.6 DisEMBL

DisEMBL (Linding et al. 2003a) includes three different simple sequence-based neural network predictors: one for 'coils', one for 'hot loops', and one for X-ray crystallographic missing coordinates, as defined by REMARK 465 entries (informative lines, labeled as REMARK's are found at the beginning of PDB files, and 465 is the code for a missing coordinate remark; however, files do not always provide REMARK 465 entries when there are missing coordinates.)

1.2.7 DISOPRED

DISOPRED was produced by Ward and Jones in an initial version (Jones and Ward 2003) and in a second version (DISOPRED2) (Ward et al. 2004), which is available at http://bioinf.cs.ucl.ac.uk/disopred/. DISOPRED is neural-network based. DISOPRED bases its prediction on PSI-BLAST generated alignments. DISOPRED2 was trained using a support vector machine method to try to balance the influence of different cases on predictor training. (Perhaps this could have some detrimental effect on overall performance, depending at least partly on how it is measured.)

1.2.8 IUPred

IUPred, so named as a predictor of 'intrinsically unstructured proteins', is based on the hypothesis that disorder is inversely related to the strength of local interactions (Dosztányi et al. 2005). IUPred bases its predictions on a matrix of statistically-derived 'interaction energies' between different pairs of residue types. This paper describes the decomposition of this matrix into separate eigenvectors, the eigenvector with the largest eigenvalue being qualitatively associated with hydrophobicity. They show that the predictor is effective at predicting residues within 'intrinsically unstructured proteins' as being disordered. However, they do not show how the predictor performs in predicting disorder in locally disordered segments of otherwise structured domains. They also do not show whether the interaction matrix used is optimal for the purpose of predicting disorder.

1.2.9 RONN

Yet another option for disorder prediction is RONN (regional order neural network; Yang et al. 2005). RONN, version 3, is a predictor that uses 'bio-basis function neural networks'. The essential idea of this predictor is that it compares sequence against disorder-related 'prototype sequences', utilizing the BLOSUM62 matrix (Henikoff and Henikoff 1992). Yang et al. (2005) show that RONN gives the best 'probability excess measure' among nine different disorder predictors. However, this probability excess measure is dependent on the prediction threshold, and does not necessarily reflect the overall performance of the predictor very well. In their Fig. 3, comparing performance of different predictors at binary classification of residues as ordered or disordered, DISOPRED2 and DisEMBL are conservative in prediction of disorder in that they show high specificity with moderate sensitivity. It is apparent from the plot that RONN's better 'probability excess' measure may just be due to selection of a less conservative prediction threshold, which increases sensitivity, but also substantially decreases specificity.

1.2.10 PreLink

PreLink (Coeytaux and Poupon 2005) predicts disordered/linker regions using a unique algorithm. It uses composition of a window of 21 residues surrounding a position. In cases that are less clearly determined by composition, it also may utilize rules involving 'cluster distance'—essentially the distance in residues from the nearest 'hydrophobic cluster'. In identifying clusters, residues are reduced to three categories: hydrophobic

('VILFMYW'), proline, or other. PreLink was developed using a relatively large dataset based on PDB structures aligned to corresponding SwissProt sequences.

1.2.11 DRIPPRED

DRIPPRED (MacCallum 2004) uses a 'self organizing map' that is trained so that fifteen-position PSI-BLAST-generated profile windows can be mapped to it. The prediction of disorder for a particular residue in the middle of such a window is related to information on the position to which it maps on the self organizing map—basically the degree to which profile windows from SCOP-based data vs. profile windows from UniProt-based data map to that position. In other words, the prediction relates to whether there is a relative absence of examples in proteins with experimentally determined structures of regions that are like the part of a protein surrounding a particular position whose disorder is being predicted.

1.2.12 FoldIndex

FoldIndex (Prilusky et al. 2005) is a web-based tool that makes predictions based on the simple prediction scheme that Uversky et al. (Uversky et al. 2000) describe for predicting whether proteins are 'natively unfolded', in which prediction is based upon a protein's 'hydrophobicity' (actually, a measure based on the Kyte-Doolittle hydropathy scale) and net charge. Rather than making whole-protein predictions as in the original Uversky et al. paper, FoldIndex uses a non-weighted sliding window for making position-by-position predictions. Prilusky et al. claim in their abstract that FoldIndex has an "error rate comparable to that of more sophisticated fold prediction methods." However, this should be considered in the light

of the statistics provided in their Table 1. On a test set of 39 ‘intrinsically unfolded’ and 151 ‘folded’ proteins, FoldIndex is shown to have a sensitivity of 77% and a specificity of 88%. DISOPRED2 is shown to have a sensitivity of 56%, but a specificity of 99%. Sensitivity and specificity are interdependent values, however. Had DISOPRED2’s prediction cutoff been adjusted to yield either a sensitivity or specificity like that of FoldIndex, it is likely that the other performance measure (specificity or sensitivity, respectively) would have been higher than the corresponding one for FoldIndex. (This provides an example of why it is good to show a ROC curve, rather than simply giving measures of performance at one cutoff.)

1.2.13 Weathers et al. SVM predictors

Weathers et al. (Weathers et al. 2004) describe a set of support vector machine-based predictors of disorder, which make predictions for entire protein segments. These predictors use a variety of linear, compositional approaches. Different parameter sets used by the predictors include the twenty standard residue types, reduced residue alphabets of various sizes, and dimer or trimer permutations. Their paper demonstrates that using reduced alphabets leads to some loss of prediction accuracy, although the loss from using reduced alphabets is not generally remarkable in their view. They do not show a significant improvement using dimers (400 in all), but do show improvement using trimers (8000 in all). Their training data included approximately 718 disordered and 1190 ordered segments. Weathers et al. were able to graphically display meaningful parameters resulting from the SVM training, including disorder parameters for the twenty amino acids.

1.2.14 Predictor history conclusion

Predictors vary in their approaches to predicting disorder, and may also vary in the type of disorder they may most effectively predict. Different predictors may produce widely varying results. Important factors in the behavior of a predictor, besides the choice of algorithm, may include the choice of data upon which the predictors are trained/based and/or the choice of parameters for predictors which are not purely data-optimized.

One primary problem with several predictors is that the amount/variety of data used for training may not be sufficient to yield very near-optimal performance. Some of the neural network predictors (which include PONDR predictors, DisEMBL, DISOPRED, RONN, and DRIPPRED), in particular, have several parameters and risk overfitting to certain cases when adequate variety and balance in data are not present. Another potential source of problems for some predictors is that residue types may be translated into values based on certain 'attribute' scales that might not be sufficiently associated with disorder to yield near-optimal efficiency, e.g., GlobPlot and FoldIndex. Some predictors incorporate the use of simple sliding windows in smoothing, which may be significantly less effective than appropriate weighted window averaging or some other smoothing technique.

Some disorder predictors show efficacy using rationally selected sets of parameters (e.g., Kyte-Doolittle hydropathies) (Li et al. 1999; Uversky et al. 2000; Linding et al. 2003b; Dosztányi et al. 2005). Efficacy, however, does not equal optimality, and thus the possibility may remain that parameters may be found that are substantially better related to disorder. On the other hand, some neural network-based predictors have been developed using 'machine learning' to optimize predictors without using predetermined parameters (Jones and Ward

2003; Linding et al. 2003a; Ward et al. 2004), but understandably, resulting networks of parameters have not been well explained in physical terms (Lise and Jones 2005). (With such complex predictors, there may also be a possibility that substantially better parameters might be obtained, depending upon the process and the data used to 'train' them.)

1.3 WHY FURTHER PREDICTORS

With so many disorder predictors available, it might be asked what developing further predictors could contribute. Many available predictors rely on parameters that are not directly obtained from disorder data. Some predictors have data-optimized parameters, but these predictors are typically neural network predictors and the optimized parameters are not readily interpretable in terms of physical mechanisms behind disorder (Lise and Jones 2005). It may not be known how special cases influence their predictions. Only one other set of disorder predictors (of which the author is aware) utilizes data-optimized parameters that can be readily interpreted—a set of SVM predictors of disorder reported by Weathers et al. (2004) However, their predictions are apparently for whole protein segments rather than individual residue positions. In the work described herein, simple, data-optimized predictors of disorder are developed to make predictions for individual residue positions. The resulting, optimized parameters provide additional scientific information not provided by parameters from other disorder predictors.

The approach of training relatively simple predictors on a large amount of data might be expected to be beneficial in a number of ways. It is more likely that such a predictor would achieve near-optimal performance than a predictor that used the same algorithm with a

rationally selected set of parameters. By approaching the maximum potential performance of the simple predictor through data-based optimization, it might then be determined whether the complexity of an algorithm such as a neural network is really contributing to the prediction or not. If a simple predictor's performance is nearly as good as a complex one, the simple one might be better for use if coding or computational time is an issue. Furthermore, because the more complex predictor might be more prone to overfitting to certain types of cases, the simple predictor's performance might be less prone to be based on excellent performance on special cases combined with sub-par performance on other, general cases. The parameters resulting from direct data-based optimization of a simple predictor might actually yield useful information on the subject of the prediction—disorder—rather than just creating another prediction tool. In a draft of my qualifying exam proposal related to this research, it was stated, "The proposed research involves using advanced parameter optimization procedures with different prediction algorithms to find improved predictors of disorder and to build a better understanding of the sequence patterns underlying disordered regions." Predictors were developed by taking simple steps, with relatively simple prediction algorithms, with an idea that taking such a careful, rational approach might eventually yield superior predictors. This approach has yielded the production of predictors with performance similar to that of DISOPRED2, a support vector machine-based predictor with neural network architecture that utilizes profiles. The discovery of multiple patterns in disordered regions was deemphasized, but the data-optimized parameters are indeed informative, and have, in particular, reinforced understanding of a major pattern related to crystallographic

disorder (and other forms of disorder), with a strong relationship with experimental hydrophobicity.

CHAPTER TWO
Methodology

2.1 METHODS INTRODUCTION

Development of a predictor includes establishing an algorithm (prediction method), obtaining an effective set of parameters, and testing the predictor. Disorder predictors described here use a weighted window summation method to assign positional disorder scores, a higher score for a residue suggesting a greater likelihood that it is disordered. Predictions are based upon either a simple sequence or a profile derived from a PSI-BLAST generated alignment of sequences related to the query sequence (the sequence for which the prediction is being obtained). Due to the tendency for chain termini to be disordered, calculation of disorder scores for residues near the ends of sequences optionally includes the addition of 'tail adjustments' that increase the disorder scores for these residues. Predictors' parameters (such as window weights and disorder values for different residue types) are optimized using X-ray crystallographic data. Data includes domains from five major SCOP classes of globular domains, representing nearly 2000 SCOP families. To aid in assessing the quality of results, a five-way cross validation scheme is used in developing and testing the predictors. Different computational tools have been developed that have facilitated the development and analysisof different predictors.

2.2 PREDICTOR DETAILS

Predictors use a simple, sliding window-based algorithm. An initial value is assigned to each position in a sequence. A weighted window sum of initial disorder values is used to calculate a disorder score for a residue position:

$$S_i = \sum_{j=-t}^{t} w_j s_{i+j} \quad [2.2\text{-}1]$$

where S_i is the score at position i; w_j is the window weight at window position j, which ranges from $-t$ to t, where t is the tail length of the window; and s_{i+j} is the initial disorder value at position $i+j$. For the simple sequence predictors, this value is assigned according to the residue type:

$$s_i = \sigma_r \quad [2.2\text{-}2]$$

where s_i is the starting value at position i and σ_r is the disorder propensity parameter for residue type, r. For profile-based predictors, this starting value is determined using a weighted sum of the disorder values for the different residue types, weighted according to the profile at that position (see below for more on profiles):

$$s_i = \sum_{r} v_r \sigma_r \quad [2.2\text{-}3]$$

where v_r is the position profile's weight for the residue type r.

Because windows 'hang off' the end of the sequence for some positions near sequence ends, 'ghost' residues may be placed on either end of the sequence, having initial disorder values of 0. Due to how disorder parameters are normalized, this essentially represents the typical, ordered residue. If the sequence represents a full polypeptide chain rather than some internal fragment thereof, the termini, which are less constrained due to lack of backbone continuation, are generally more likely to be disordered—in this case, the addition of tail adjustments improves predictions. If tail adjustments are included, a positive value is added to the scores of residues near the termini, based on the residues' positions in relation to the amino- or carboxy-terminus. In the amino-terminal case,

$$S_i = \sum_{j=-t}^{t} w_j s_{i+j} + \tau_{N,i} \quad [2.2\text{-}4]$$

and in the carboxy-terminal case,

$$S_i = \sum_{j=-t}^{t} w_j s_{i+j} + \tau_{C,L-i-1} \quad [2.2\text{-}5]$$

where $\tau_{N,k}$ or $\tau_{C,k}$ is the tail adjustment parameter, for the amino- or carboxy-termini, respectively, at distance, *k*, from terminal residue.

In summary, the predictors described in this work use weighted sliding windows and base their scoring on either the query sequence alone or a PSI-BLAST generated profile. The standard sliding window length is 35. Tail adjustments, if included, are added to each of the 30 positions at either end of the sequence. Adjustable parameters include disorder values based on residue type, window position weights, and N- and C-terminal tail adjustment values.

2.3 CROSS VALIDATION

Cross validation allows predictors to be trained and tested on independent data, while still (in this work) taking all data into account in generating final predictor parameters. Cross validation provides a sense of data-dependent variance of optimized parameters and test performance measures, as well as a means of judging whether substantial overfitting is occurring. The data are separated into five different groups. For each predictor, five optimizations are performed, each time excluding one of five data subsets from the 'training' data, resulting in five different parameter sets (see Fig. 2.3-1). Parameter sets may be normalized and adjusted. The performance of each parameter set is tested on the data subset not used in its optimization. Five parameter sets from individual optimization runs are averaged to give a final parameter set for the predictor. Results from five individual tests may also be averaged to give summary test results for the predictor. Tests of 'outside' predictors and statistics may also be obtained on the five individual test sets and then averaged.

2.4 DATASET

Both quality and variety of data used for training and testing predictors are important factors that have been taken into account in development and testing of predictors described here. Predictors were optimized on X-ray crystallographic data from the Protein Data Bank (PDB) (Berman et al. 2000) to discriminate between 'missing' and 'non-missing' residues. A 'missing' residue's C-α carbon is either missing coordinates or is assigned an occupancy of 0 (see below on 'counted' residues). Structures dated before 2000 or with a resolution 'worse'

than 3.0 angstroms are not in the dataset. Data include 1912 families from the first five classes (alpha and/or beta domain) of SCOP (Murzin et al. 1995), version 1.67, thus excluding potentially problematic domains, including membrane-spanning domains, 'small proteins', and coiled-coil domains. If a protein chain contained any SCOP domains, regions of the chain not assigned to any domains were assigned to a nearest domain. Multi-chain domains were not included in the data set. To enhance variety, rather than selecting a single representative from each family, any proteins that were not excluded for some reason were kept in the dataset.

Some structures were excluded due to PDB file-related issues. For example, gaps were found by aligning the sequence given in SEQRES lines with the sequence of residues found in a chain's coordinates; if a mismatch occurred between the sequence found in the SEQRES entries and the sequence found in the coordinates, then the structure was not used. Non-standard residues included selenomethionine and all other non-standard residue types. These were uncommon, in general, and selenomethionine was more frequent than all other nonstandard types combined. Selenomethionine and other nonstandard residue types were considered two distinct residue types in terms of disorder predictors. Any positions with nonstandard residue types except for selenomethionine were not counted in performance analyses but were counted in calculating log odds ratios (equation 2.6-1).

Profiles were generated on alignments extracted from results of PSI-BLAST searches on chain sequences, with the PSI-BLAST option of showing alignments for up to 1000 database sequences in the output file. This large number of sequences in the output apparently triggered memory problems for some queries, for which profiles were not

therefore obtained. Thus, the size of the dataset for profiles was reduced; see Table 2.4-1. Up to three PSI-BLAST passes were performed, and the e-value cutoffs for generating profiles for subsequent iterations and for displaying resulting aligned sequences were both 0.001. DISOPRED2 and PSIPRED results were obtained. A sequence identity limit of 97% was used in generating alignments from PSI-BLAST output (i.e., if multiple sequences were more similar than the threshold allowed, only one was used). Code for a version of COMPASS (Sadreyev and Grishin 2003) was modified to generate profiles. Use of position-specific independent counts reduced overrepresentation of closely related sequences (Sunyaev et al. 1999). Pseudocounts (Tatusov et al. 1994) were generated using the BLOSUM62 (Henikoff and Henikoff 1992) matrix. Profiles were normalized at each alignment position to yield fractional weights with a sum of 1.

The full dataset ('simple') was used to train simple-sequence based predictors. The subset of the data for which profiles were successfully generated ('profile') was used to train profile predictors and make comparisons of performance shown here. Table 2.4-1 provides summary statistics for residues included in analyses of performance, and Table 2.4-2 provides such statistics for individual cross validation data subsets. To my knowledge, this dataset represents more data than that used in developing and testing any previously published predictor of disorder.

To prevent spurious parameter results or overlap between training and testing sets, certain residues were included in sequences when making predictions, but were excluded in various training and testing situations from analyses of performance (including ROC score calculations). Residues belonging to stretches of missing residues less than four residues

long were excluded. SCOP domains were expanded to include unassigned regions of chains, but nine residues on either side of a domain division were excluded from performance analysis, to prevent training/testing data overlap. Exclusion of terminal residues from ROC score calculations during training of certain predictors removed bias for methionine and for histidine (due to polyhistidine tags), and based window and residue disorder parameters primarily upon internal disordered stretches. When training or testing included all terminal residues, thirty residues were excluded on sequence ends containing polyhistidine tags, due to observed influence on tail adjustment values of residues internal in sequence to polyhistidine tags (see section 4.4).

Table 2.4-1. Dataset statistics. The simple dataset was used to train the simple sequence-based predictors. The profile dataset was used to train the profile-based predictors. See text for descriptions of how residues are counted as missing or non-missing. (*) When zero residues are excluded at the termini, sequence ends with detected polyhistidine stretches are excluded, at least the first thirty residues from the end, consistent with training and testing.

Dataset	No. of SCOP families	No. of domains	No. of terminal residues excluded	No. of missing residues	No. of non-missing residues
Simple	1912	28128	0*	183902	5563922
			18	85490	4957018
			30	66460	4473062
Profile	1773	23386	0*	157195	4496369
			18	71792	3958638
			30	55243	3531262

Table 2.4-2. Statistics on cross validation subsets. The sequence dataset was used for training the simple sequence-based predictors. The profile dataset was used in training the profile-based predictors and in testing/comparison of various predictors. (Refer to main Table 2.4-1 caption regarding residue statistics).

Dataset	No. of terminal residues excluded	Test set no.	No. SCOP families	No. of domains	No. of missing residues	No. of non-missing residues
Sequence	0*	1	382	5629	35306	1067876
		2	382	6034	38590	1069879
		3	382	5191	37320	1075863
		4	383	5709	39818	1152927
		5	383	5565	32868	1197377
	18	1			15625	955137
		2			17089	936184
		3			18136	960262
		4			19492	1025865
		5			15148	1079570
	30	1			12105	863519
		2			13368	829356
		3			13768	869406
		4			15472	924087
		5			11747	986694
Profile	0*	1	359	4645	31248	856247
		2	354	5129	34961	908560
		3	357	4275	30491	851557
		4	347	4615	32485	881081
		5	356	4722	28010	998924
	18	1			13648	753994
		2			14991	786695
		3			14946	749576
		4			15652	774341
		5			12555	894032
	30	1			10471	670977
		2			11498	689772
		3			11435	670049
		4			12312	689077
		5			9527	811387

When making comparisons of one's own predictor with others, one is generally 'playing on home field'. One of the 'home' advantages is that the training and testing data were obtained in the same fashion. Also, the ROC score (which appears to be a good evaluation of performance for optimization, given success in obtaining parameters) is directly related to sensitivity/specificity curves. On the other hand, a favorable aspect of testing in this work is that the testing data as a whole comprise a large part of the PDB. DISOPRED2 also has an advantage in that no efforts were made to exclude examples used in its training or similar to those used in its training, from testing. Thus, actual differences in potential performance between simple predictors such as ours and neural network predictors might better be quantified if developed with consistent training. Although the window size selected for standard predictors (35) is longer than that of DISOPRED, it is not expected that this difference has a significant effect on overall performance (note in Fig. 3.2-4b that the window weights at the three positions on either end are relatively small).

2.4.1 Test set 3

In several instances it has been observed that test set 3 seems to be an outlier in terms of its behavior. This appears to be related to the presence in test set 3 of a group derived from one particular SCOP family, 'RNA-polymerase beta-prime' (ID = 64490), which contains four representatives in the 'simple' dataset : PDB entry 1I50, chain A; PDB entry 1IW7, chains D and N; and PDB entry 1K83, chain A. Each of these contains missing regions of substantial size. Two of the chains, PDB 1I50, chain A, and PDB 1K83, chain A, contain large stretches of sequence that imperfectly repeat the sequence, 'PSTPSYS' (only these two representatives of the family are present in the 'profile' dataset).

2.5 PREDICTOR OPTIMIZATION

2.5.1 Overview

A simulated annealing variant optimization scheme, depicted in Fig. 2.5-1, was used to obtain predictor parameters. Predictor performance was measured using a receiver operating characteristic score (Gribskov and Robinson 1996), calculated by:

$$ROC_n = \frac{\sum_{i=1}^{n} a_i}{nA} \quad [2.5\text{-}1]$$

where a_i is the number of true positives that sort by score above the ith false positive; A is the total number of true positives; and n is the limiting number of false positives. Instead of using a fixed value for n, n was calculated as the greatest integer less than or equal to a certain fraction of the number of false positives in the data subset, the fraction typically being 0.5 (this is denoted as a '$ROC_{0.5}$' score). Throughout training runs, different random subsets of the entire training dataset were used. In a five-way cross validation, the data were divided evenly by SCOP family into five sets for the cross-validation. For the sake of balance, each SCOP family was represented with approximately equal frequency, except for families with less than five sequences.

2.5.2 Details

Parameter optimizations were performed using a simulated annealing variant algorithm, diagrammed in Fig. 2.5-1. In optimizations, 700 annealing temperature steps were typically performed. A typical annealing step consisted of 250 cycles, each cycle involving a newly selected subset of the data. A subset of data was obtained by selecting approximately one half of SCOP families from a training data set, and from these families generally

selecting a sequence at random, except for families containing less than five examples, for which selections were less frequent. A cycle would run until two perturbations had been successfully retained or twenty had been tried, whichever came first. Analysis of performance was based on the ROC score for the performance of the predictor with a given set of parameters on the cycle's given data subset, calculated over all 'counted' residues in the data subset, based on each residue's score and status as 'gap' or 'nongap'. Perturbations for each mutable parameter were generated according to the 'very fast annealing' (VFA) distribution, generated by the function:

$$f = \mathrm{sgn}(r - 0.5)T\left[\left(1 + \frac{1}{T}\right)^{|2r-1|} - 1\right] \quad [2.5\text{-}2]$$

[presumably this may be attributed to Ingber (1989), cited by Cai and Shao (2002)]

where f is the fractional value ($-1 \leq r < 1$; ideally, the range would include the value 1, but the exclusion of this single value in the set of random numbers is presumed to have had minimal impact due to the large number of discrete random numbers possible in the random number function used—function MyMath::real_rand() found in my header my_math.hpp) by which the maximum perturbation increment is multiplied to obtain a perturbation; r is a random number between 0 and 1; and T is the perturbation 'temperature' (which is on a different scale than the temperature used for the Metropolis decision—see Fig. 2.5-1). A set of C++ classes was developed for flexible optimization of parameters, with XML-style input/output formatting. This allowed individual parameters to be optionally fixed, bounded, perturbed along a logarithmic scale, and/or normalized within a certain group of parameters

(normalization enabling comparison across various optimizations). Details of optimizations and code also may be publicly available at http://prodata.swmed.edu.

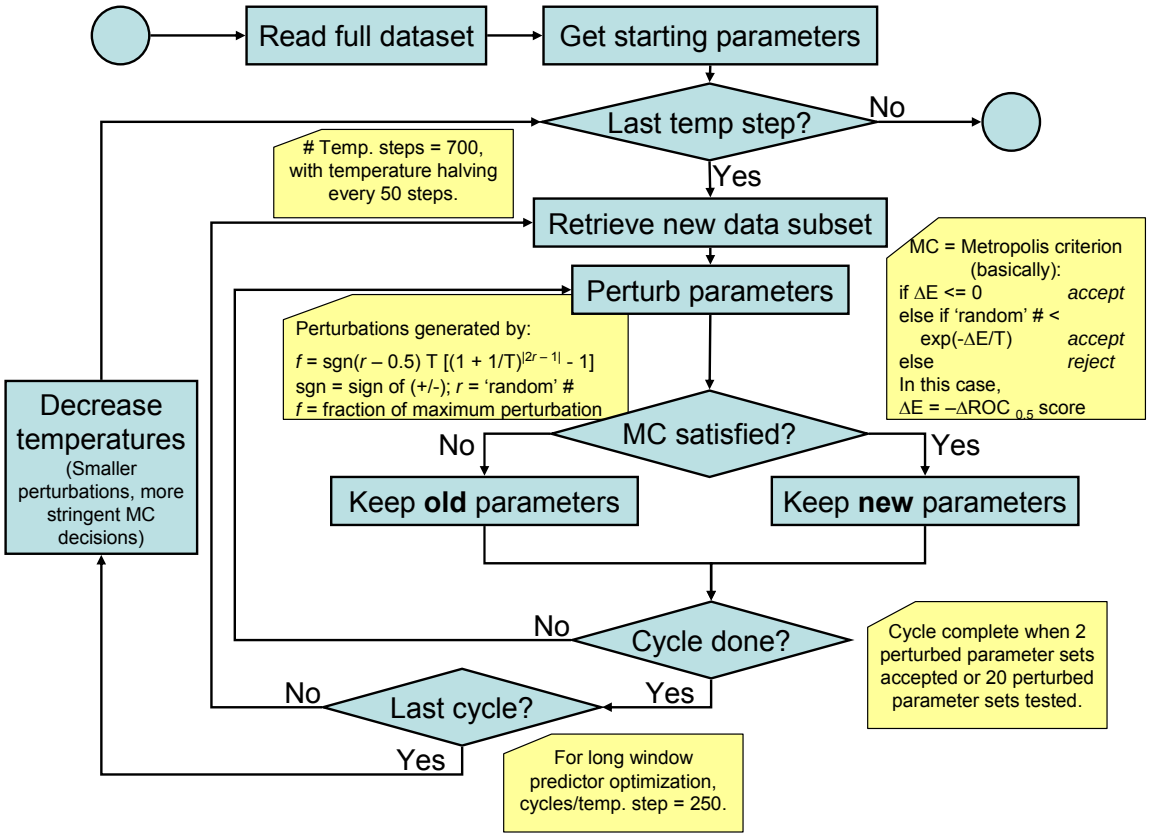

Figure 2.5-1. Optimization scheme. The Metropolis criterion (Metropolis et al. 1953) basically indicates how decisions were made to accept or reject perturbed values.

Although the Metropolis criterion (Metropolis et al. 1953) (see Fig. 2.5-1) was basically used in deciding whether to reject or retain parameters (with a coarse random function, rand()), it is likely that simply retaining parameter sets that performed better than or equal to previous parameter sets would have worked just as well, with the simple, linear nature of the predictors; the use of VFA perturbations; and the rotation of data sets.

For the simple and profile basic window predictors, the residue type and window position weight parameters were optimized in a five-way cross validation (see Table 2.4 2 for statistics on subsets), using each combination of four out of five test sets for training. In optimizing these parameters, the performance on 18 terminal residues on each end (in the case of a polyhistidine tag—detected as at least four histidines in a row near a sequence end—the terminus up through the polyhistidine stretch and then 18 residues further in) was excluded from analysis of performance. These values were then normalized to yield (on respective training sets) score distributions with mean value of 0 and standard deviation of 1. Values for the general nonstandard residue type were optimized with other residue types, but when normalization was performed, this parameter was assigned a value of 0. The 'ghost extension' and N terminal methionine (for the predictor without tail adjustments) residue types were also given the value of 0. In optimizing tail adjustments for the simple and profile predictors, the final, normalized parameters from the basic predictor optimizations were held fixed while tail adjustment values were optimized. The special disorder value for N terminal methionine was optimized for the simple sequence predictor along with its tail adjustments. Terminal (sequence end) residues were not excluded by default, but thirty residues at a sequence end were automatically excluded if a polyhistidine stretch was detected at the

sequence end. For the final predictors, parameters from different cross validation training runs have been averaged together.

2.6 TESTING/STATISTICAL ANALYSIS

In testing predictors (e.g., generating ROC curves) and in other statistical analyses (e.g., calculating disorder vs. order log odds ratios for each residue type), standard datasets were used to provide consistency. These datasets were balanced in a manner similar to the way data were balanced during optimization. Generally, for each family in a particular training or testing set, 100 representatives were chosen from among available members of the family unless the family contained less than five available members, in which case, the number of representatives chosen is roughly equal to n / 5 * 100, where n is the number of available members in the family. Of course, with this scheme a particular member of a family could be represented more than once. Average ROC curves are generated by averaging corresponding points from five separate ROC curves generated on five different testing data subsets.

2.6.1 Log odds ratios

Disorder vs. order log odds ratios may be calculated as follows:

$$LOR = \ln \frac{\left(\frac{P(i \mid miss)}{1 - P(i \mid miss)} \right)}{\left(\frac{P(i \mid pres)}{1 - P(i \mid pres)} \right)} \quad [2.6\text{-}1]$$

where the log odds ratio is calculated for residue type, *i*. *P*(*i*|*miss*) denotes the probability of the occurrence of residue, *i*, given that residue is in the *miss*ing state. *P*(*i*|*pres*) can be read the same way, where *pres* refers to the *pres*ent, or non-missing, residue state.

In calculating log odds ratios for residues based on sequences, probabilities were calculated from discrete values obtained from counts of occurrences of different residue types among 'missing' vs. 'non-missing' residues. In the cases of profiles, 'log odds ratios' were calculated in essentially the same way as for simple sequences—instead of using discrete counts, however, the frequency of a residue type in the set of 'missing' or 'non-missing' positions was obtained from summing fractional weights for the particular residue type over all counted residues within that category ('missing' or 'non-missing' residues).

Penultimate residue statistics were obtained to obtain evidence that methionine is sometimes entirely missing from the chains, with methionine aminopeptidases being selective according to the type of the residue adjacent to the methionine (the penultimate residue). Statistics were calculated using only sequences for which the first residue in the sequence was methionine and for which the adjacent (penultimate) residue was ordered. Log odds ratios were calculated for N terminal methionines that were missing vs. present in these cases.

2.6.2 Paired t-test

To assign some statistical measure to differences in performance between predictors, the two-tailed, pairwise t test was used. There is evidence that there would be marked skew in distributions of differences between performance for predictors, especially considering multiple instances in which ROC score differences for test set 3 (see section 2.4.1) often are

substantially different from ROC score differences for test sets 1, 2, 4, 5, which are all closer to each other (see, for example, Fig's 3.2 2, 5.1 3). Because of this, t test-derived probabilities might be expected to underestimate the significance of differences between predictors. (Thus, if the t test probability is less than 0.05, there should not be a problem with concluding that a particular difference is significant.)

2.7 SUMMARIES OF OPTIMIZATION RUNS

Several optimization runs have been run, with a variety of prediction algorithms. Optimization runs were assigned 'run base names' that serve as identifiers, and predictors may be identified by the optimization run (set) from which the predictor's parameters were derived. The names typically indicate the general algorithm being used (e.g., sw35_? stands for simple window, length 35) and then have a number associated (e.g., sw35_1, sw35_2, ...) for different optimizations of parameters for the same algorithm. Examples of differences in optimization include differences in the general dataset being used, differences in the 'annealing' details (e.g., how many cycles per temperature step), or differences in the which residues were excluded from the measures of predictor performance (e.g., how many residues, if any, at domain boundaries were excluded). Table 2.7 1 gives information on different runs. Code for various optimizations is provided in appendix. The 'standard predictors' include sw35_8 (simple sequence), sw35_st30_8 (simple sequence with tails), p2w35_4 (profile), p2w35_st30_6 (profile with tails).

Table 2.7-1. Information on optimization runs.

Run base name	Sequence or Profile	Tail Adjustments	Window length	Optimization performance measure	Cycles per temp. step	Term. resd's excl.	Parameter starting temp's
sw35_7	Sequence		35	ROC 0.05	250	18	0.4
sw35_8	Sequence		35	ROC 0.5	250	18	0.4
p2w35_4	Profile		35	ROC 0.5	250	18	0.4
sw35_st30_8	Sequence	Tails (30)	35	ROC 0.5	250	0	0.4
p2w35_st30_6	Profile	Tails (30)	35	ROC 0.5	250	0	0.4
sw9_1	Sequence		9	ROC 0.5	150	18	0.4
sw9_2	Sequence		9	ROC 1.0	150	18	0.4

2.8 COMPUTATIONAL METHODS

The work described required significant coding time and CPU time. A variety of techniques/tools/strategies were implemented. A few significant developments are here described that might be regarded as novel and may have contributed significantly in accomplishing work efficiently. A set of classes was developed that allowed facile control of parameters in optimization. Software architecture was designed to allow interchangeability of software components.

2.8.1 Parameter organization classes

As in other work (Wehrens and Buydens 1998), our optimization scheme took advantage of a biological analogy. Names of classes used in storing parameters reflected genetic organization (see Fig. 2.8 1). A genome is an entire parameter set. An element describes an individual parameter with its associated attributes/settings. Chromosomes and genes provide two levels of subgrouping. This way of organizing parameters provides multiple convenient features. It allows logical grouping of parameters. This is convenient

both in terms of conceptualizing and working with parameters directly from a human standpoint. It also facilitates a modular approach to programming, in which each component of a predictor may, in turn, read in externally controlled parameters that it uses, if any. For example, in a weighted window-based predictor that also uses tail adjustments, the predictor could utilize a chromosome containing four genes—one that contains residue disorder value parameters; one that contains window position weights; one that contains amino-terminal tail adjustment values and one that contains carboxy-terminal tail adjustment values. If such a predictor actually used two sub-predictors, with two different parameter sets organized in the same fashion, depending upon conditions, the identically-organized parameter sets for each sub-predictor could be placed into two chromosomes. The genes also allow the potential use of parameters with different types of values, including Boolean, discrete, and 'continuous' values. All of the parameters within a gene must be of the same basic value type, but different genes in a chromosome may contain sets of parameters with different base data types, from one gene to another. Furthermore, a gene may be specialized to allow interactions between parameters. In practice, some genes were set to normalize their parameter sets—for example, genes containing window position weights were often set so that their parameters would be normalized to yield an average parameter value of 0.5.

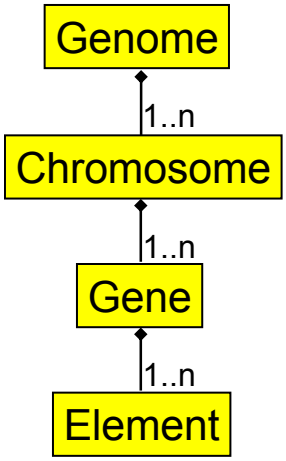

Figure 2.8-1. Organization of parameter classes.

A further feature of this system of organizing parameters is that it provides XML-style input and output. With the ability to organize parameters and set individual parameter attribute values and to describe organization and settings in an XML format, direct human manipulation of parameters was facilitated in some ways.

2.8.2 Software engineering

In the process of predictor optimization and testing, it was recognized that good organization of code could contribute to more efficient development. A modular approach in engineering software facilitated modification of code and implementation of different related tasks. A software component could be reused a number of times in combination with various other components. Tasks included optimizing, analyzing, and implementing predictors.

Some general component types are clients (which organize and use other components), datasets, predictors, and analyzers. The fundamental data type is the residue record, which varies based on the needs of the tasks at hand. Data are organized in sequences and are loaded and stored in a ‘central’ location that basically remains fixed. Various operations may be performed in succession on the data, often passing pointers to the

applicable sequences, rather than copying the data from one memory location to another. Different fields of the basic residue data structure that are required are detected automatically at compile-time using a template meta-programming technique, so that the basic residue record structure of the central data contains all the necessary components for various processing/analysis steps. By thus using templates, different sets of components may be integrated without reprogramming for different data structures that may be required for different combinations of components. This facilitates the introduction of new/modified components, such as new types of predictors and new types of analyzers of predictions.

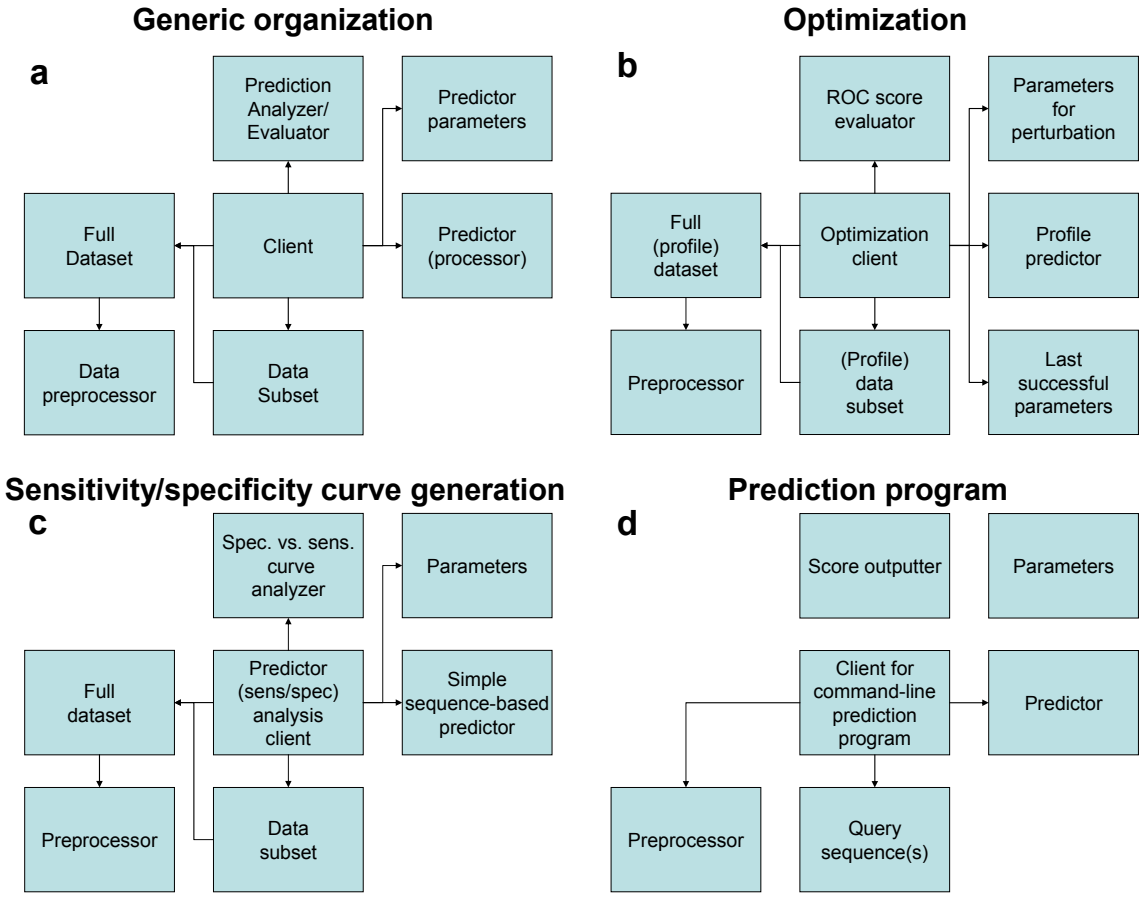

Figure 2.8-2. Software architectures. a) Generic organization of different components. b) Instance of optimization. c) Instance of analysis, in which sensitivity/specificity curves are generated (which may be transformed into ROC curves). d) In a prediction program.

CHAPTER THREE
Predictor parameter and performance results

3.1 INTRODUCTION

We developed linear predictors of crystallographic disorder that use a weighted sliding window-based, compositional prediction strategy. Predictions are based upon either the query sequences alone or profiles derived from PSI-BLAST (Altschul et al. 1997)-generated alignments. An option is available to enhance scoring of disorder-prone amino- and carboxy-terminal regions (Li et al. 1999) by adding simple values to disorder scores of terminal residues, based on their positions at the termini ('tail adjustments'). From a bioinformatics standpoint, it is shown below that these predictors are effective. From a structural science standpoint, the simple, linear algorithms allow informative interpretation of optimized parameters. The simple sequence predictor's data-optimized residue disorder values are largely reducible to a single physical property, hydrophobicity, providing evidence that crystallographic disorder is closely linked to a side chain's tendency to interact with aqueous surroundings as opposed to retaining hydrophobic interactions.

3.2 RESULTS

3.2.1 Predictor performance

Comparing the performance of predictors may assist in determining how well various algorithms and their parameters model disorder. Figure 3.2-1 compares performance of standard predictors (see section 2.7)—including simple sequence and profile-based predictors, with and without tail adjustments—among themselves and with DISOPRED2

(Ward et al. 2004). DISOPRED2 is a support vector machine/neural network-based predictor of disorder that uses PSI-BLAST-generated sequence alignment profiles in its prediction. As shown with DISOPRED (Jones and Ward 2003; Ward et al. 2004), using profiles enhances overall prediction performance for weighted window-based predictors, but this improvement should be considered in the context of performance results as described here.

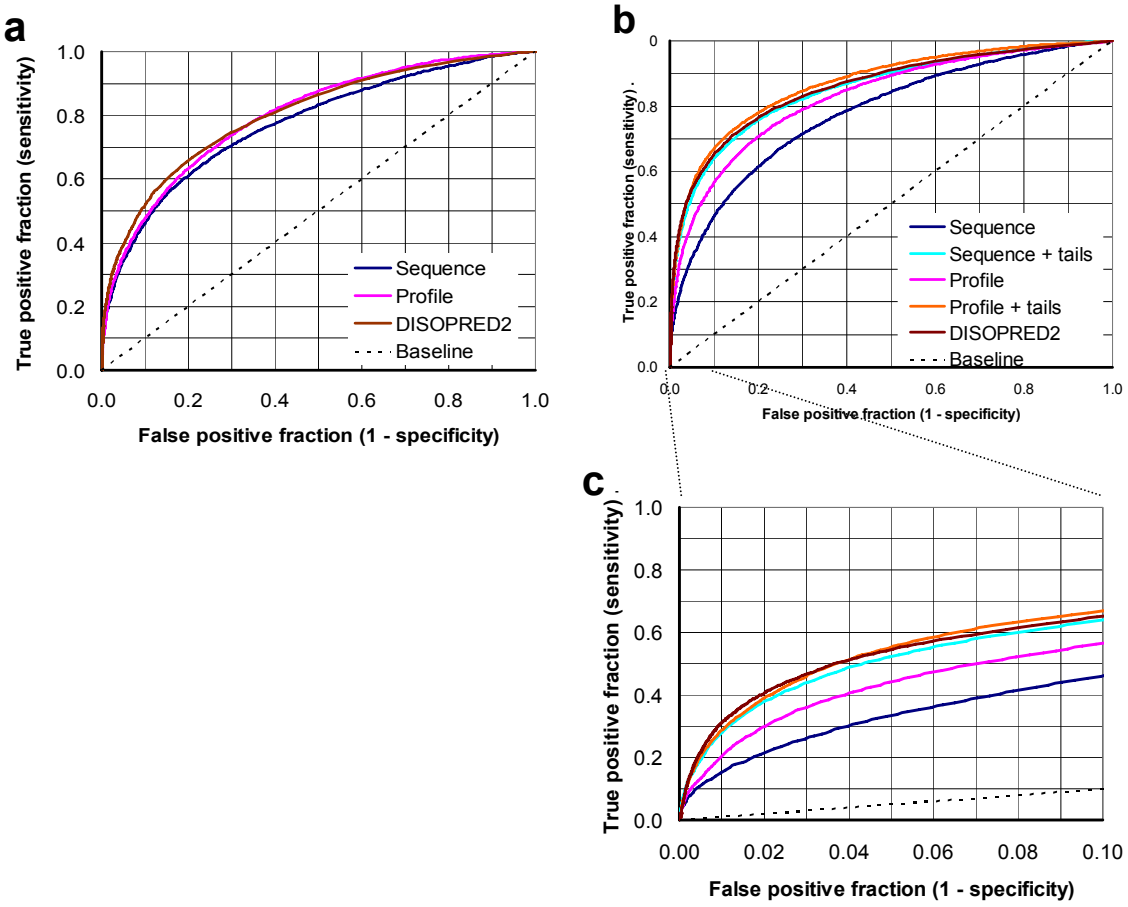

Figure 3.2-1. Test performance comparison for standard predictors and DISOPRED2 (Ward et al. 2004). a) 'Average' ROC curves excluding 30 terminal residues in performance analysis. b) Average ROC curves including terminal residues. c) Same as b, with change in scale. See Methods (Chapter 2) for details on average ROC curves and terminal residue exclusion.

Disorder in polypeptide chain terminal regions makes up more than half of the disorder in datasets (Table 2.4 1). For non-terminal protein regions, the simple sequence-based predictor performance is similar to that of the profile-based predictor and DISOPRED2 (Fig. 3.2-1a). The profile predictor shows modest improvement in performance over the simple sequence predictor primarily in the lower specificity range of the ROC curve—i.e., for lower-scoring residues. Figure 3.2-1b shows performance measures with termini included. Performance differences between DISOPRED2, the profile predictor, and the simple predictor primarily occur at the sequence termini (compare Figs. 3.2-1b and 3.2-1a). Including terminal regions in measures of performance yields substantial performance improvement in DISOPRED2, some improvement in the profile predictor (without tail adjustments), and the least improvement in the simple sequence predictor (without tail adjustments). Adding tail adjustments to the simple sequence and profile predictors substantially narrows the performance gap between the simple sequence and profile predictors and DISOPRED2. Although the predictors with tail adjustments show, on average, performance similar to that of DISOPRED2, comparing performance on individual cross validation data subsets (see Table 2.4 2), reveals substantial difference in behavior (Fig. 3.2 2; see section 2.4.1 for discussion of test set 3).

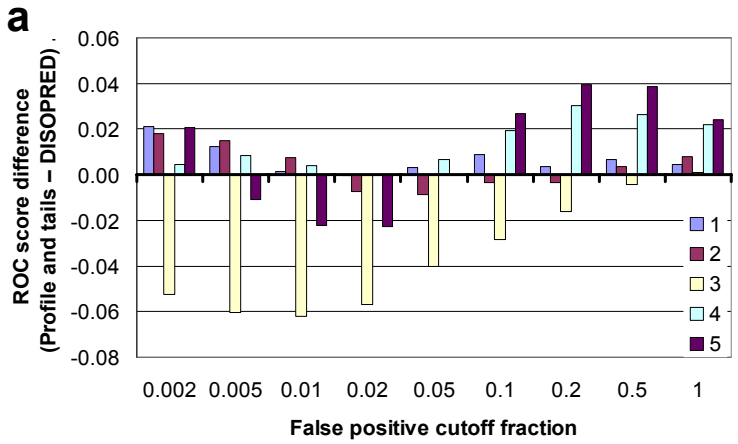

Figure 3.2-2. Differences between performance of profile with tail adjustments predictor and DISOPRED2 on individual testing data subsets. ROC scores at different cutoffs obtained by subtracting DISOPRED2's ROC score from the profile with tail adjustments predictor's ROC score.

A single prediction by a profile-based predictor may take minutes due to the time required for PSI-BLAST to generate sequence alignments, while the simple sequence-based predictor may take less than a second. The quality of profiles may vary widely from case to case, and neural networks introduce the increased possibility of unknown sources of bias in predictions. Because the profile predictor with tail adjustments and DISOPRED2 do not perform substantially better than the simple sequence predictor with tail adjustments, using the simple sequence-based predictors may be advantageous in certain instances, such as comparing proteins that have similar sequences, utilizing such a prediction as a step in another bioinformatic method, or other instances in which speed is an important factor or there is a good reason to avoid bias toward special cases. It is acknowledged that other predictors are now available that may perform significantly better than DISOPRED2 (Peng et al. 2006), and our predictor has not been compared directly with these predictors.

3.2.2 Optimized parameters

Due to the transparency of the predictor algorithms, the data-optimized parameters yield insight into what is contributing to prediction—including whether unwanted bias is present (see section 4.4 regarding polyhistidine tag treatment)—and provide scientific insight on disorder. These parameters include residue disorder values, window weights, and tail adjustment values. Optimized parameters generally agree well between cross-validated parameter sets (Fig. 3.2 3), and, because of their consistency, they may be averaged to give final parameters (Fig. 3.2 4, see Appendix A for tables of parameter values, particularly Table A-1 for final parameter values for the standard predictors).

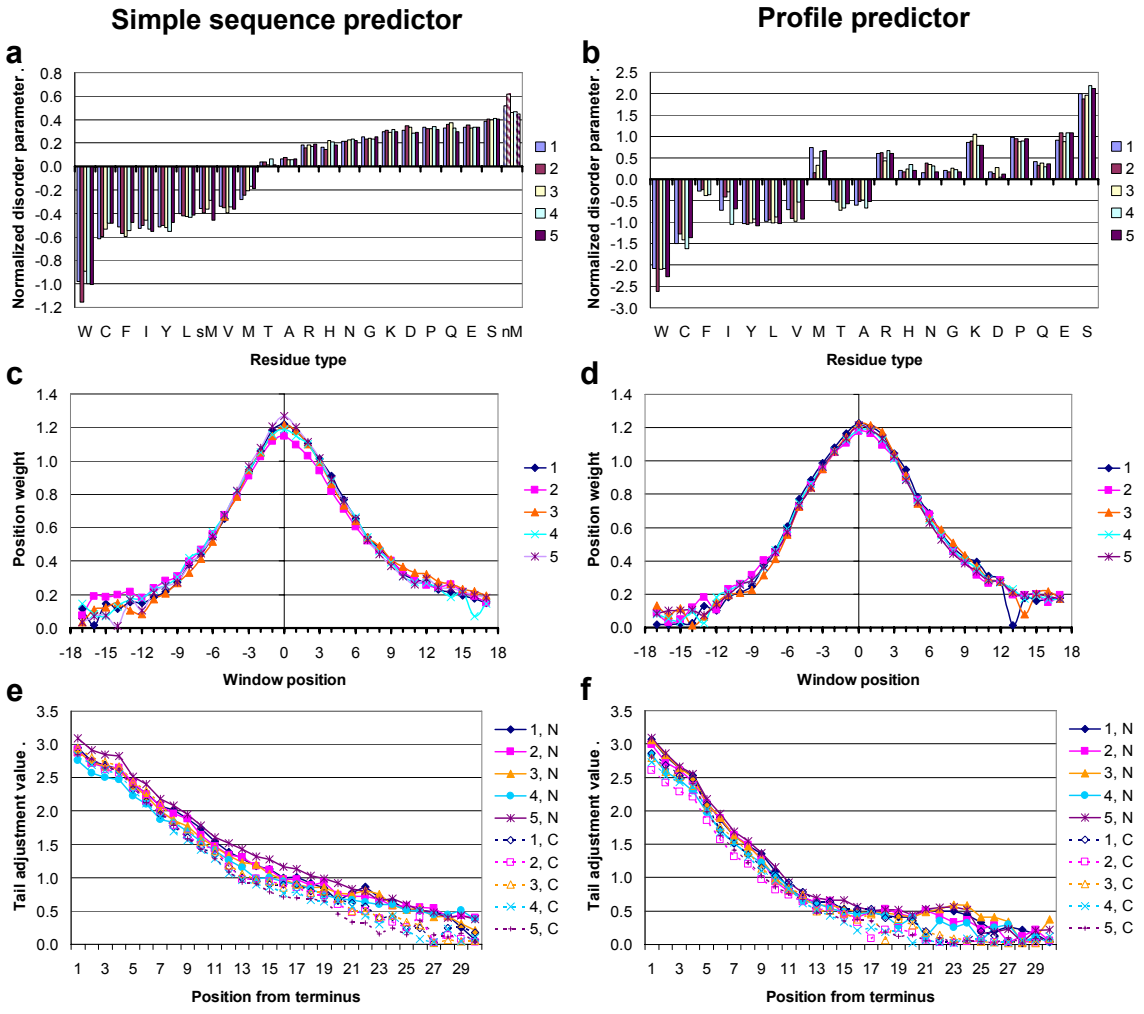

Figure 3.2-3. Optimized parameters from individual cross-validation runs. Numbers in individual panel legends indicate cross-validation run number. a) Normalized optimized residue parameters, simple sequence predictor (nonstandard/unknown residue type also optimized but then set to 0 at normalization, before testing—not included; ghost extension residue type also set to 0, not included). Striped bars for nM (N-terminal methionine) represent that these values are obtained for the simple sequence predictor with tail adjustments—this value is 0 for the simple sequence predictor without tail adjustments. b) Profile predictor residue disorder values. c) Simple sequence predictor window weights. d) Profile predictor window weights. e) Simple predictor N- and C-terminal tail adjustment values. f) Profile predictor N- and C-terminal tail adjustment values.

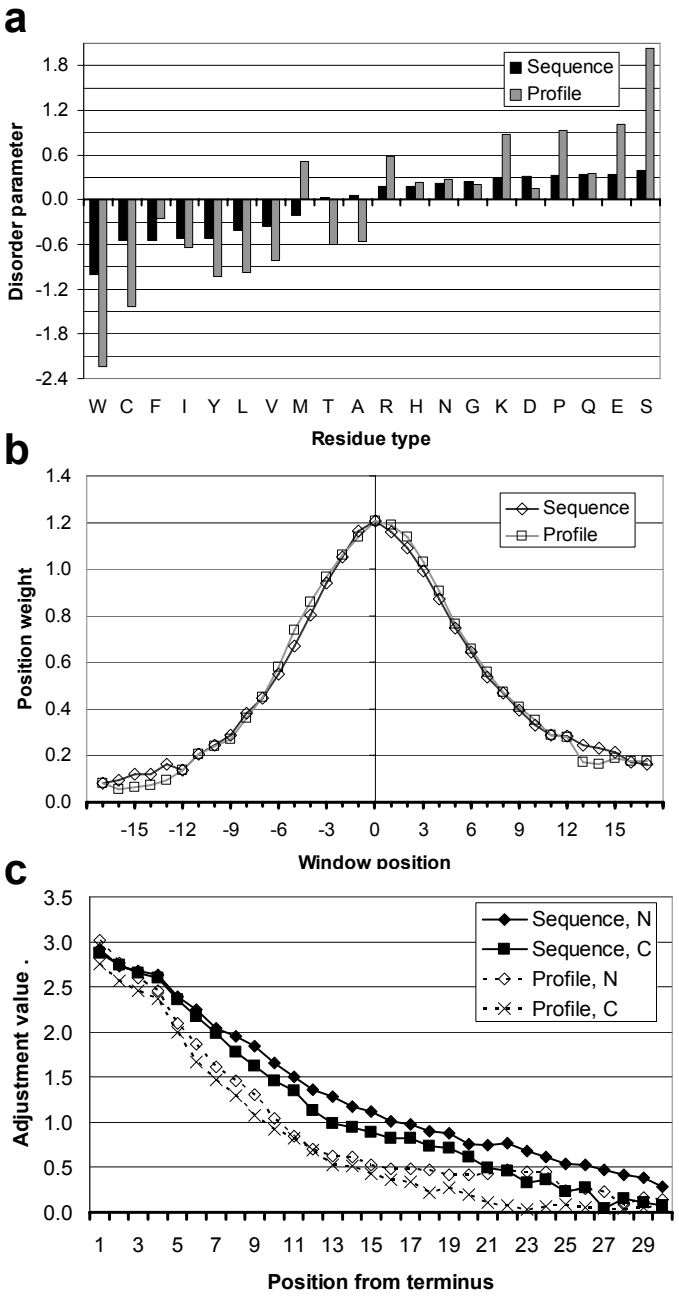

Figure 3.2-4. Final predictor parameters. a) Disorder parameters for standard residues, after optimization, normalization, and averaging. (Special residue types not shown; see also Table 2.2-1.) b) Average optimized window position weights. c) Average optimized N- and C-terminal tail adjustments.

Disorder score distributions, particularly for ordered residues, are approximately normal, with higher scores indicating greater disorder tendency. Residue disorder parameters for the profile and sequence predictors were normalized to yield final score distributions for ordered residues that approximate the standard normal distribution (see Fig. 3.2-5). Thus, scores represent an approximate Z-score, the number of standard deviations from the average ordered residue.

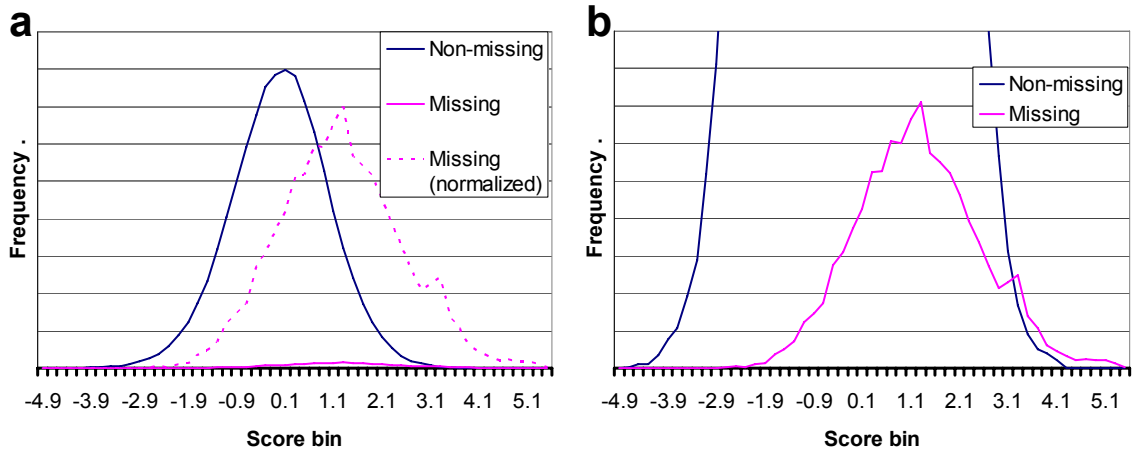

Figure 3.2-5. Simple predictor score distributions for missing ('disordered') and non-missing ('ordered') residues. Residues within 18 positions from the termini were excluded in this analysis (see Methods). Score bin values are bin centers, with score bin widths of 0.2. a) Scale allows full view of 'non-missing' (basically, ordered) residue distribution of scores. Relative frequency of missing vs. non-missing residues can be seen (solid blue and pink lines). b) Scale allows clear view of distribution of 'missing' residues.

Table 3.2-1. Standard disorder values (average normalized optimized disorder parameters for simple sequence predictor) for common residue types and selenomethionine. (See also Fig. 3.2-4).

Residue type	Optimized disorder value
W	-1.00739
C	-0.540732
F	-0.540414
I	-0.514274
Y	-0.513589
L	-0.418184
sM	-0.373603
V	-0.358167
M	-0.216377
T	0.0333232
A	0.0642762
R	0.176914
H	0.18568
N	0.221683
G	0.241088
K	0.300523
D	0.313504
P	0.32731
Q	0.336406
E	0.33729
S	0.400289

‘Optimized disorder values’, ‘disorder parameters’, or similar phrases refer to averaged normalized optimized residue disorder parameters for the standard simple sequence predictor (sw35_8) unless otherwise specified. Table 3.2-1 displays these disorder propensities. Residue disorder parameters for the sequence and profile based predictors’ follow different patterns (Fig. 3.2 4a, Fig. 3.2 6), but for both the profile and simple sequence predictors, tryptophan is clearly the most order-associated residue type and serine is the most disorder-associated residue type. The simple sequence optimized disorder values are well correlated with their statistical disorder propensities (log odds ratios; see equation 2.6-1) for the simple sequence predictor, but not the profile predictor (Fig. 3.2-7), showing that optimal prediction parameters may or may not show a close correlation with values obtained through a more conventional statistical approach, and giving evidence that a linear algorithm is more appropriate for a simple sequence-based predictor than a profile-based predictor, which takes more advantage of special clues. Threonine and alanine, with disorder values close to 0, have approximately average ordering propensity.

Other groups have found sets of disorder propensities similar to our optimized disorder propensities, including sets of disorder propensities independently derived from X-ray crystallographic, NMR, and CD data by Williams et al. (Williams et al. 2001), and especially propensities obtained for a support vector machine-based predictor of disorder (Weathers et al. 2004). The missing coordinates propensities shown by Linding et al. (Linding et al. 2003a) also demonstrate some similarity to ours. The values shown by Peng et al. (Peng et al. 2005), fig. 2, show substantial differences from our optimized disorder propensities, but this appears to be because their values are calculated as differences in the

absolute fraction of residues found in disordered vs. ordered data sets (e.g., tryptophan has a smaller negative value because the overall number of tryptophans is smaller than other residue types, making the fractions of tryptophan among disordered and ordered residues smaller).

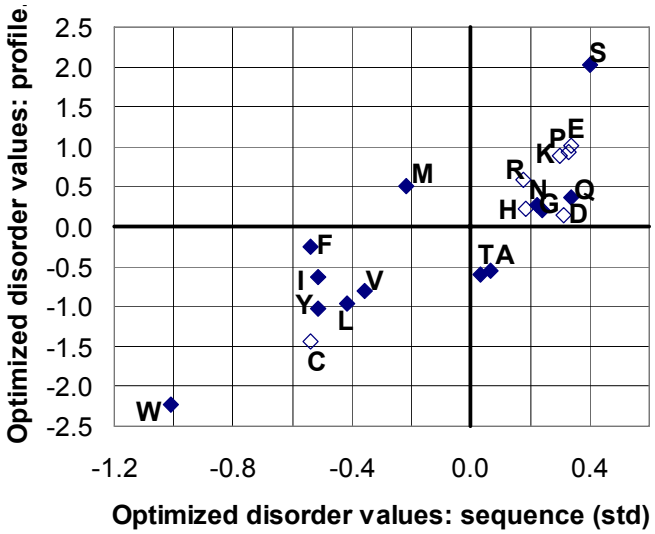

Figure 3.2-6. Optimized disorder values for profile predictor vs. simple sequence predictor. (Open diamonds represent C, P, and ionic residues).

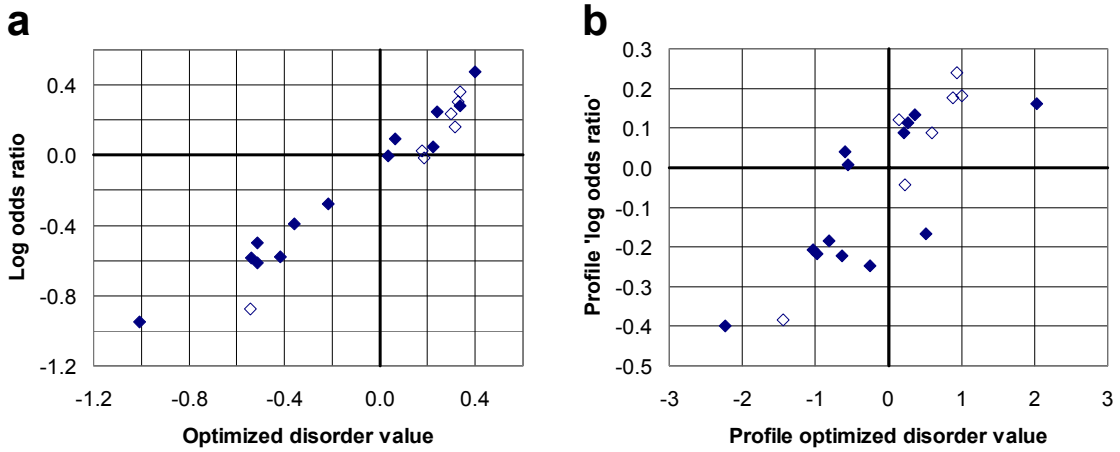

Figure 3.2-7. Correlation plots of optimized disorder propensities and log odds ratios. a) Log odds ratios vs. simple sequence predictor optimized values. b) Profile 'log odds ratios' vs. profile optimized values. (Open diamonds represent C, P, and ionic residues).

Profile and simple sequence predictors optimized window weights are similar (Fig. 3.2-4b). Both sets of window weights show a small skew, with weights in the C-terminal direction (positive positions) greater than their N-terminal counterparts (negative positions).

Amino- and carboxy-termini show similar propensities for becoming disordered (Fig. 3.2-4c), implying that N- and C-termini have similar physical propensities for disorder due to decreased constraint, exclusive of the effects of sequence on disorder in these regions. Tail adjustments for the profile and simple sequence predictors are consistent in magnitude close to the termini, suggesting that the residue composition-based scores for the profile and simple sequence predictors are equivalent (recall that the residue disorder propensities were normalized to produce Z-score-like disorder scores prior to the optimization of tail adjustment values). The profile predictor shows a consistent relative drop in both the amino and the carboxy-terminal adjustment values, with respect to those of the simple predictor.

(One possible explanation for this is that there is some variation in the composition of alignments in both amino- and carboxy-terminal positions that enhances prediction of disorder in these regions.)

The relatively high value for methionine in the profile disorder parameters is presumably due to N-terminal methionines in aligned sequences. The high special disorder value for N-terminal methionine (see Table A-1; Fig. 3.2 3) provides evidence that disordered termini are often erroneously excluded from chain sequences in PDB files (see section 5.3.5). Tail adjustments thus may underestimate actual propensities for terminal positions to be disordered.

3.2.3 Correlation of disorder and hydrophobicity

The most notable information from optimized parameters comes from optimized disorder values. Optimized disorder values (see Table 3.2-1) for the standard side chains were compared against various residue scales that reflect different physical/experimental/statistical properties of residues.

Secondary structure is sometimes divided into three categories: helix, strand, or coil. Optimized disorder values and various coil propensity scales are not well correlated (see, for example, Fig. 3.2-8a), confirming the finding of Linding et al. (Linding et al. 2003a) that crystallographically disordered regions and coil regions in general have substantially different compositions. The association between coil propensity and values optimized for the profile predictor is even weaker (not shown; $R^2 = 0.276$ for residues excluding C, P, and ionic residues).

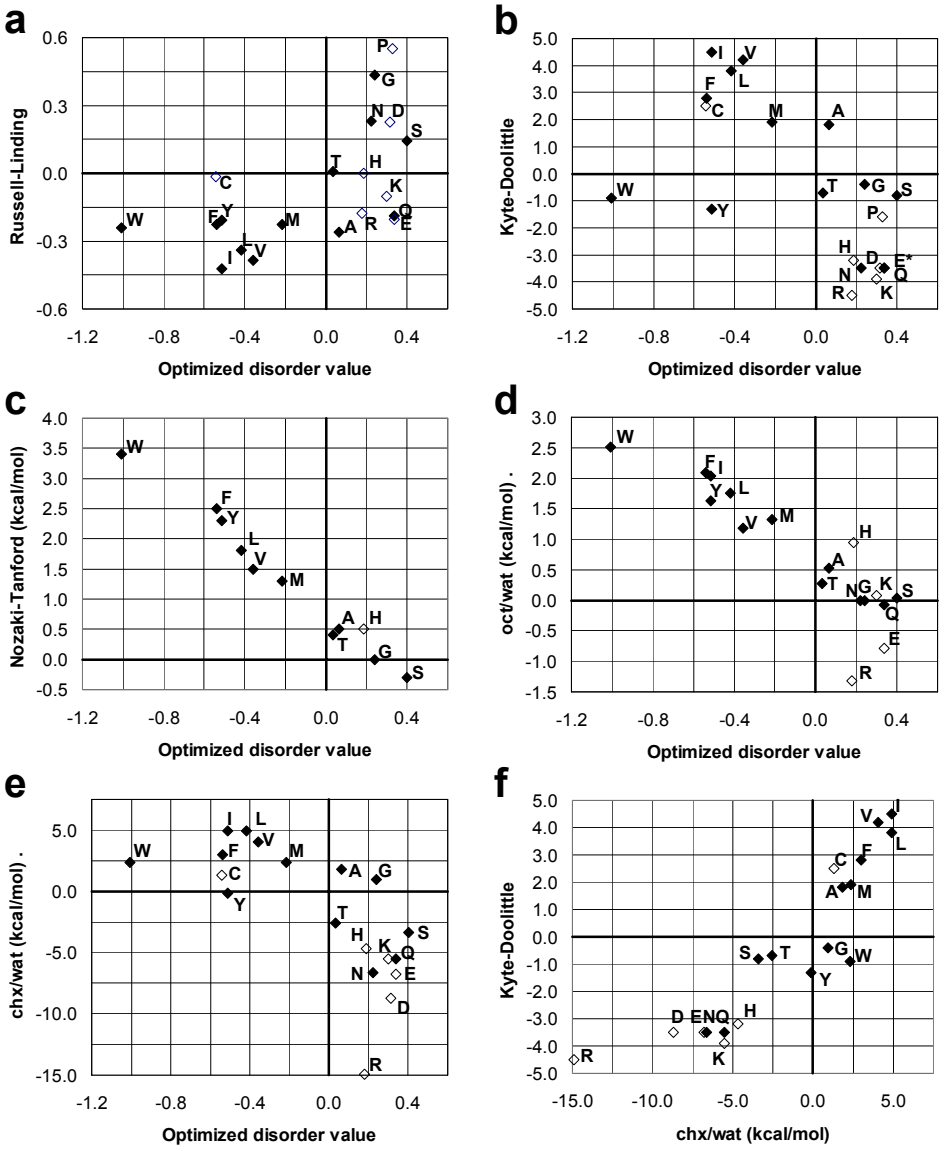

Figure 3.2-8. Correlation plots of various residue scales. R^2 values are given for residue types marked by closed diamonds (excluding D, E, R, K, H, C, and P, which are marked by open diamonds). a) Russell-Linding coil propensities (Linding et al. 2003b) vs. disorder value; $R^2 = 0.427$. b) Kyte and Doolittle (1982) hydropathy vs. disorder value; $R^2 = 0.225$. (*) E and Q overlap in this plot. c) Nozaki and Tanford (1971) side chain transfer energies vs. disorder value; glycine, as the reference amino acid, is given a value of 0; $R^2 = 0.982$. d) 'Octanol/water' transfer energies, which include influence from organic solute(s) other than octanol (Guy 1985; Radzicka and Wolfenden 1988) vs. disorder value; $R^2 = 0.941$; note that the Radzicka/Wolfenden version was plotted, which is the inverse of the Guy scale, and excludes proline. e) Cyclohexane/water transfer energies (Radzicka and Wolfenden 1988) vs. disorder value; $R^2 = 0.491$. f) Kyte-Doolittle hydropathy vs. cyclohexane to water transfer

energies; $R^2 = 0.818$. For the entire Nozaki-Tanford subset of residues, R^2 values are respectively a) 0.463, b) 0.089, c) 0.977, d) 0.881, e) 0.390, f) 0.730.

Strengths of association with different 'hydropathy' scales vary greatly. The association between optimized disorder values and Kyte and Doolittle (1982) hydropathies is weak (Fig. 3.2-8b). The Hopp and Woods scale (1981) is associated with disorder propensities with an unimpressive R^2 of 0.649 for all residues, but has a clearly visible linear relationship with disorder values for the majority of non-ionic residues (see Fig. 5.1-2b). The Hopp-Woods scale traces back (Levitt 1976) to the Nozaki and Tanford (1971) hydrophobicity scale, an incomplete scale based on various experiments measuring amino acid solubilities in organic solvent (primarily ethanol), water, and combinations thereof (see section 5.2.1). Figure 3.2-8c clearly shows a strong inverse correlation between the Nozaki-Tanford hydrophobicities and optimized disorder values. Disorder values have a similar relationship with 'octanol/water' partitioning energies (Guy 1985; Radzicka and Wolfenden 1988) (see Fig. 3.2-8d), which are derived not only from octanol/water transfer experiments, but also from experiments utilizing other organic solutes including methanol and ethanol, while cyclohexane/water partitioning energies (Radzicka and Wolfenden 1988) show more similarity to the Kyte-Doolittle scale (see Fig. 3.2-8e, f).

Substituting various scales that might be used *a priori* in disorder prediction, as in GlobPlot (Linding et al. 2003b), in place of optimized disorder values, can cause the simple sequence predictor performance to drop significantly (see Fig. 3.2-9), showing the importance of the pattern of disorder propensities. A scale, such as the Hopp-Woods scale, may show a strong association with disorder propensities for the majority of residues, but because of residues with special behavior, the strength of such relationships may not be reflected in predictor performance or in routine statistical measures of association. This may

help to explain why the strength and significance of the relationship between hydrophobicity and disorder have not been fully appreciated (see section 5.2), and illustrates how the rational selection of parameters can be problematic.

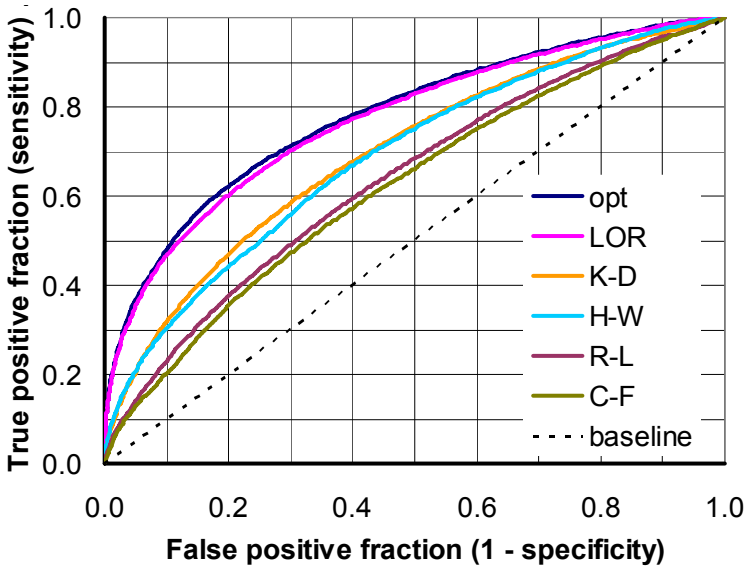

Figure 3.2-9. ROC curves for simple window predictor substituting various residue scales (same window position weights used for all). Abbreviations: opt – optimized disorder values; LOR – log odds ratios; K-D – Kyte-Doolittle scale (Kyte and Doolittle 1982); H-W – Hopp-Woods scale (Hopp and Woods 1981); R-L – Russell-Linding coil propensity scale (Linding et al. 2003b); C-F – Chou-Fasman coil propensity scale (Chou and Fasman 1974).

3.3 DISCUSSION

3.3.1 A hydropathic spectrum

General phenomena that significantly affect the physical behavior of proteins include dispersion forces, ionic and hydrogen bonding interactions, other electrostatic interactions, and the hydrophobic effect. These forces have different degrees of influence in various 'hydropathy' scales that attempt to describe how amino acids or side chains partition between different environments. Without considering this, differences between hydropathy scales may

not be well explained. Understanding such differences is important in interpreting results. The terms ‘hydrophobicity’ and ‘hydrophilicity’ are sometimes used as simple opposites, referring only the directionality of some hydropathic property. In usage here, the hydrophobic effect is basically the tendency of a ‘less polar’ part of a molecule to interact with other less polar entities instead of water, for which such an interaction is unfavorable. As counterpart to the hydrophobic effect, the ‘hydrophilic effect’ is essentially the tendency of polar and ionic groups to interact preferably with water (or aqueous solution), over other environments. The terms ‘hydrophobicity’ and ‘hydrophilicity’ are used accordingly.

Radzicka and Wolfenden (1988) compared results from different partitioning experiments that serve as model systems where hydrophobic and hydrophilic effects are dominant, respectively. Octanol/water (or equivalent polar organic/aqueous) partitioning (Guy 1985; Radzicka and Wolfenden 1988) may primarily reflect hydrophobicity for non-ionic residues, because hydrogen bonding potential is satisfied to a similar degree in both polar aqueous solvent and water. On the other hand, interaction potentials for polar groups are not well-satisfied in cyclohexane (‘wet’ cyclohexane does not contain much water). Subtracting the ‘octanol/water’ scale from the cyclohexane/water scale to produce ‘cyclohexane/octanol transfer energies’ (Radzicka and Wolfenden 1988) thus provides a hydrophilicity scale. Comparing the magnitudes of the ‘octanol/water’ and ‘cyclohexane/octanol’ scales (Radzicka and Wolfenden 1988) suggests that the strength of the ‘hydrophilic effect’, in full force, is greater than that of the hydrophobic effect, in terms of potential influence on side chain partitioning energies. Various hydropathy-related scales may be deconvolved into approximate hydrophobic and hydrophilic components by finding a

linear combination of the ‘octanol/water’ and cyclohexane/octanol scales that is optimally associated with that scale (see Fig. 3.3-1).

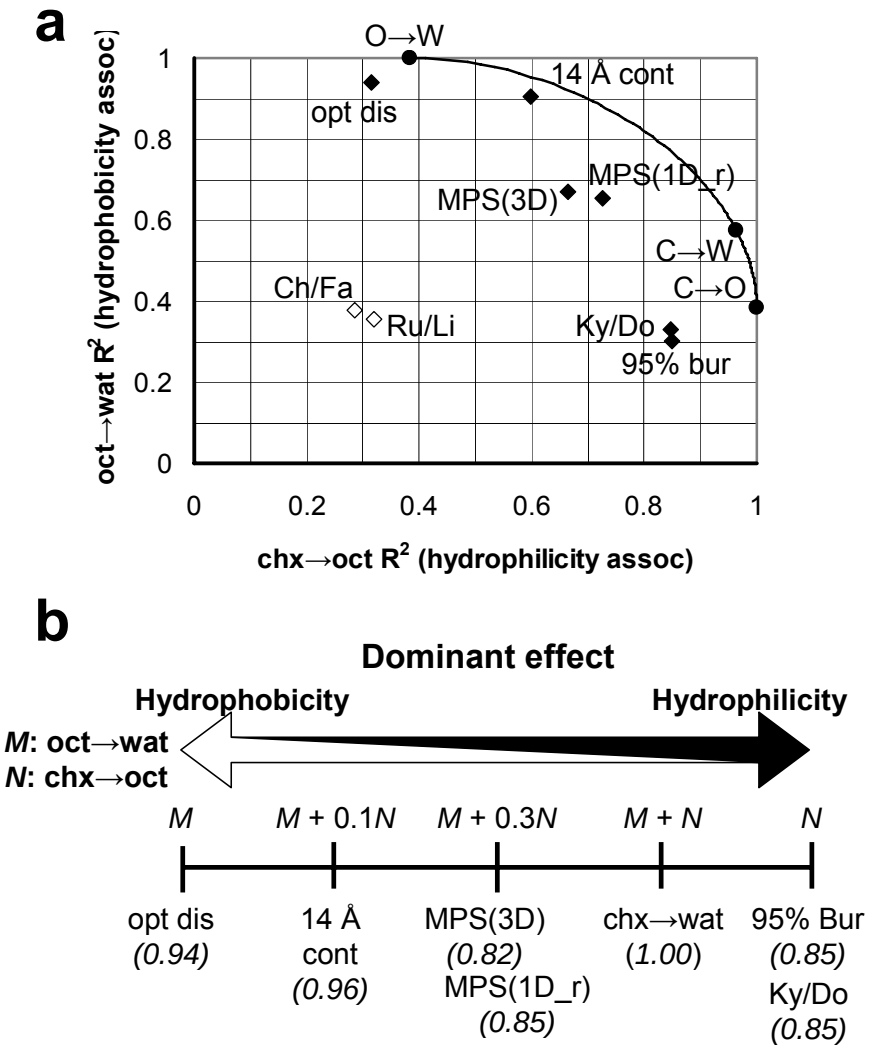

Figure 3.3-1. Approximate deconvolution of various scales into hydrophobic and hydrophilic components using 'octanol/water' (Guy 1985; Radzicka and Wolfenden 1988) and 'cyclohexane/octanol' (Radzicka and Wolfenden 1988) partitioning energies. a) R^2 values for various scales against the 'octanol/water' and 'cyclohexane/octanol' transfer free energies, always excluding C, P, H, R, K, D, E; obtained by subtracting the 'octanol/water' scale from the cyclohexane/water scale (Radzicka and Wolfenden 1988). The curved line represents correlations for exact combinations of the two scales. b) Approximate locations of different scales along a hydropathy 'spectrum'; locations on spectrum are given by linear combinations of the 'octanol/water' scale (M) and 'cyclohexane/octanol' scale (N) that approximate the relative degrees to which the hydrophilic and hydrophobic effects are present (the general magnitude of the hydrophilic effect is greater than that of the hydrophobic effect). Below names of scales, in italics, are strengths of associations (R^2) of those scales with the respective linear combination of M and N noted above it.

Abbreviations: oct→wat: ‘octanol/water’ transfer energies (Guy 1985; Radzicka and Wolfenden 1988); chx→oct: ‘cyclohexane/octanol’ transfer energies (Radzicka and Wolfenden 1988); chx→wat: cyclohexane/water transfer energies (Radzicka and Wolfenden 1988); opt dis: standard optimized disorder scale; opt dis: optimized disorder scale (standard); 14 Å cont: 14 Å contact number (Nishikawa and Ooi 1986); MPS(3D): Punta-Maritan X-ray diffraction/NMR-based transmembrane scale (Punta and Maritan 2003); MPS(1D_r): Punta-Maritan non-X-ray diffraction/NMR experiment transmembrane scale (Punta and Maritan 2003); Ky/Do: Kyte/Doolittle scale (Kyte and Doolittle 1982); Ru/Li: Russell/Linding coil propensity scale (Linding et al. 2003b); Ch/Fa: Chou/Fasman coil propensity scale (Chou and Fasman 1974).

Statistics discriminating between almost fully buried residues and residues that are more exposed on the surface have been shown to be better associated with the 'cyclohexane/octanol' (hydrophilicity) scale than the 'octanol/water' scale (Radzicka and Wolfenden 1988). Scales related to transmembrane helix prediction are also associated with hydrophilicity. The data-derived Punta-Maritan transmembrane scales (Punta and Maritan 2003) reflect roughly equal influence from hydrophobicity and hydrophilicity, and they are better correlated with linear combinations of 'octanol/water' and 'cyclohexane/octanol' partitioning scales than with either alone (see Fig. 3.3-1b). No combination of these two scales is strongly correlated with coil propensity. Disorder propensities are well explained by hydrophobicity alone. In summary, there is a hydropathic spectrum, ranging from hydrophobic to hydrophilic effect dominance, determined by the strength of the hydrophilic effect relative to the hydrophobic effect (see Fig. 3.3-1b).

3.3.2 Modeling disorder

Global disorder falls into different structural categories (Dunker et al. 2001; Uversky 2002). There are also many possible ways local disorder may occur. It is not just the good association between optimized disorder values and hydrophobicity, but their lack of separate association with hydrophilicity that support some models or explanations of disorder above others.

The good association between hydrophilicity and 95% residue burial statistics (Radzicka and Wolfenden 1988) (Fig. 3.3-1) reflects that polar and ionic groups on side

chains are not easily buried within a protein, as ionic/hydrogen bonding potential must be satisfied to prevent a large energetic penalty. With the strength (Radzicka and Wolfenden 1988) of the 'hydrophilic effect' (compare 'octanol/water' and cyclohexane/water scales, Fig. 3.2-8d, e), buried hydrophilic residues without good hydrogen-bonding partners would be expected to disrupt otherwise good potential crystal contacts. This is consistent with evidence that mutating hydrophilic surface residues to alanines can improve crystallization (Derewenda 2004). It is concluded that the optimized disorder values do not simply reflect propensities for residues to participate in crystal contacts, since a hydrophilic component should then likewise be reflected in disorder propensities (see, for example, 14 Å contact number, fig. 3.3-1).

Our scale associations suggest that in model systems that are well-associated with disorder propensities, both the aqueous and organic phases provide hydrogen bonding and van der Waals interactions for side chains; a primary difference between phases appears to be in the ability of one phase to offer hydrophobic protection to hydrophobic portions of side chains. The strong inverse relationship between hydrophobicity scales (from such systems) and disorder propensities suggests that, like the organic phase, the close environments of ordered residues tend to reduce unfavorable side chain/water interactions (Rose et al. 1985) (through interactions with other hydrophobic parts of the same protein or possibly neighboring proteins), while still allowing favorable ionic and hydrogen-bonding interactions with aqueous surroundings. The environment for disordered residues is more like the aqueous phase, with side chains being more indiscriminately exposed to aqueous surroundings than ordered surface residues.

Disordered backbone is presumably flexible. Pappu and Rose (2002) calculated energies of different alanine dipeptide conformations using soft atomic repulsion potentials, approximating a chain in 'good solvent'. Their calculations show a broad, relatively flat energy basin (within the $-\phi, +\psi$ quadrant) that includes polyproline II helix (as their global energy minimum) and β-strand backbone conformations. The flatness of the basin indicates that this region of conformational space is not highly constrained when a chain is interacting largely with solvent, as is suggested here to be often the case. Disordered backbone likely often adopts conformations within this relatively flexible region of conformational space as the backbone and residues interact with aqueous surroundings.

If a disorder-related category (e.g., linkers, proline-rich regions, molten globules, etc.) has consistent statistical propensities that cannot be reasonably well associated with hydrophobicity, then other models of disorder should be considered. Crystallographic disorder has been subdivided into static disorder and dynamic disorder (Huber 1979; Huber and Bennett 1983). It would appear that among the missing residues used in predictor development, static disorder resulting from small domain rotations is less abundant than dynamic disorder, given that disorder propensities might then be expected to reflect localization to the surface vs. deep burial (i.e., with a hydrophilic component) and/or coil propensity. For similar reasons, variable conformations of surface-bound loop/coil regions do not appear to significantly influence disorder values. Physical differences in disorder may translate into functional differences—for example, a highly soluble disordered region may be more easily degraded. Because of exclusion of residues in the terminal and domain junction regions in optimizing residue disorder parameters (see section 2.5), disorder values may be

less applicable for terminal or domain-linking disordered regions than for intra-domain regions.

The accompanying supplementary workbook (Excel spreadsheet file) contains a parameter-energy calculations worksheet that shows how optimized disorder values may be tentatively converted into estimated average energies of transfer from ordered to disordered states using log odds ratios and information on score distributions (Fig. 3.2 5). The majority of side chains are order-promoting. With the linear relationship between these energies obtained from disorder values and experimental hydrophobicities, the intercept and slope may be interpreted respectively as the average contribution of the backbone to order/disorder transfer energies and relative degree of hydrophobic protection. The backbone appears to favor disorder by a relatively small energy. Slopes of hydrophobicity vs. disorder trends suggest that the average difference of protection of hydrophobic groups in the ordered state (for surface/coil residues) vs. the disordered state is significantly less than difference between the protection afforded in organic vs. aqueous phases. Section 5.4 contains further discussion on transforming parameters into estimated energies.

The idea of normally inefficient hydrophobic protection may help to explain seemingly paradoxical associations between disorder and aggregation. For example, polyglutamine stretches are associated with various diseases including Huntington's disease and are known to aggregate. Given glutamine's relatively low hydrophobicity/high disorder propensity, polyglutamine sequences would not tend to form good strand-strand interactions with most protein sequences. Perhaps lining polyglutamine strands together (Khare 2005;

Sambashivan 2005) results in unusually efficient hydrophobic protection of side chains with their consistent length, without significantly hampering amide hydrogen bonding.

CHAPTER FOUR
Further predictor details and comparison

4.1 INTRODUCTION

In the previous chapter, a few standard predictors were highlighted. This chapter provides more information on those and other predictors.

4.2 STANDARD SIMPLE SEQUENCE-BASED PREDICTOR

The basic predictor is the simple sequence predictor, which consists of assigning each residue an initial disorder value based on its residue type and then obtaining a final value through weighted window summation. The standard simple sequence predictor (sw35_8) has a window length of thirty-five, and was optimized using $ROC_{0.5}$ performance measures. Figure 4.2-3 shows training $ROC_{0.5}$ scores throughout the course of optimization (panel a), as well as testing scores for parameter sets saved through the course of optimization (panel b). Figure 4.2-3c suggests the degree of overfitting, which, as would be expected given the similarity in parameters resulting from different runs (Fig. 3.2-3a, b), is fairly small. Cross-validation performance for the standard predictor is fairly consistent across the various cross validation runs (see Fig. 4.2-1).

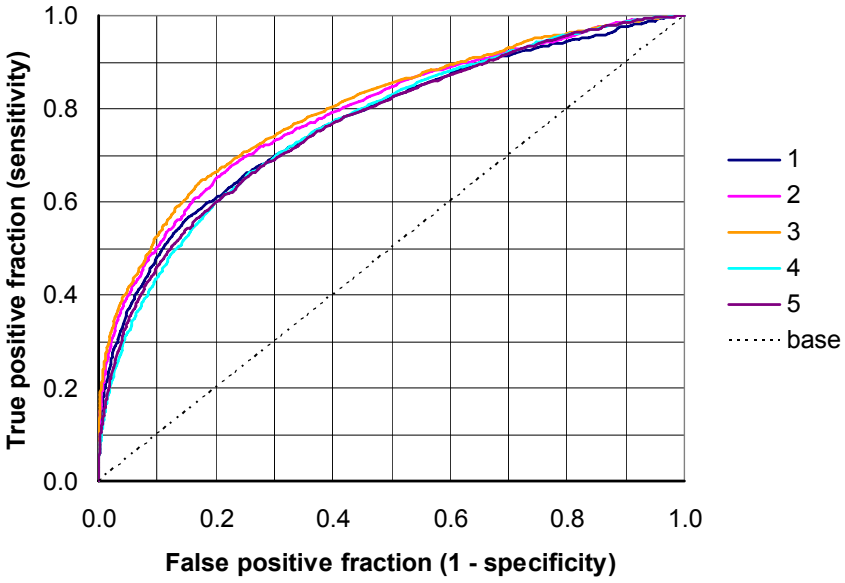

Figure 4.2-1. Different test performance curves for sw35_8 (on the simple sequence dataset, with 18 terminal residues excluded in testing).

When comparing the performance of this simple predictor with DISOPRED2 on non-terminal residues, the performance of the simple predictor approaches that of DISOPRED2 fairly well (Fig. 3.2-1). At certain false positive cutoff fractions, on some test sets, it even shows better performance, but, as might be expected, on test set 3, DISOPRED2 is markedly better (Fig. 4.2-2; see section 2.4-1).

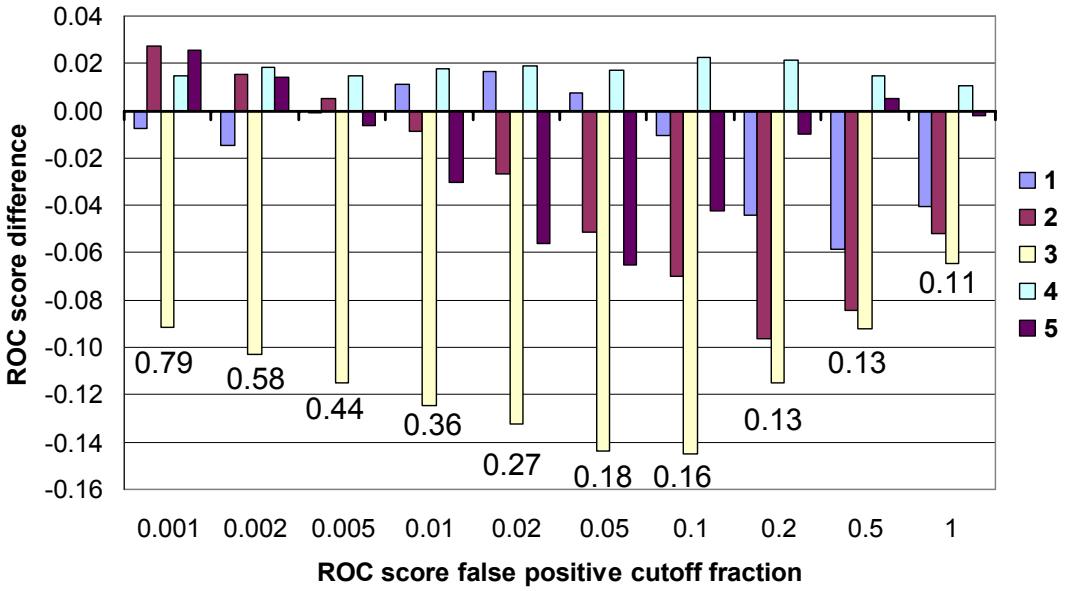

Figure 4.2-2. ROC score differences for individual test sets, for the simple predictor (sw35_8) vs. DISOPRED2 (sw35_8 – DISOPRED2).

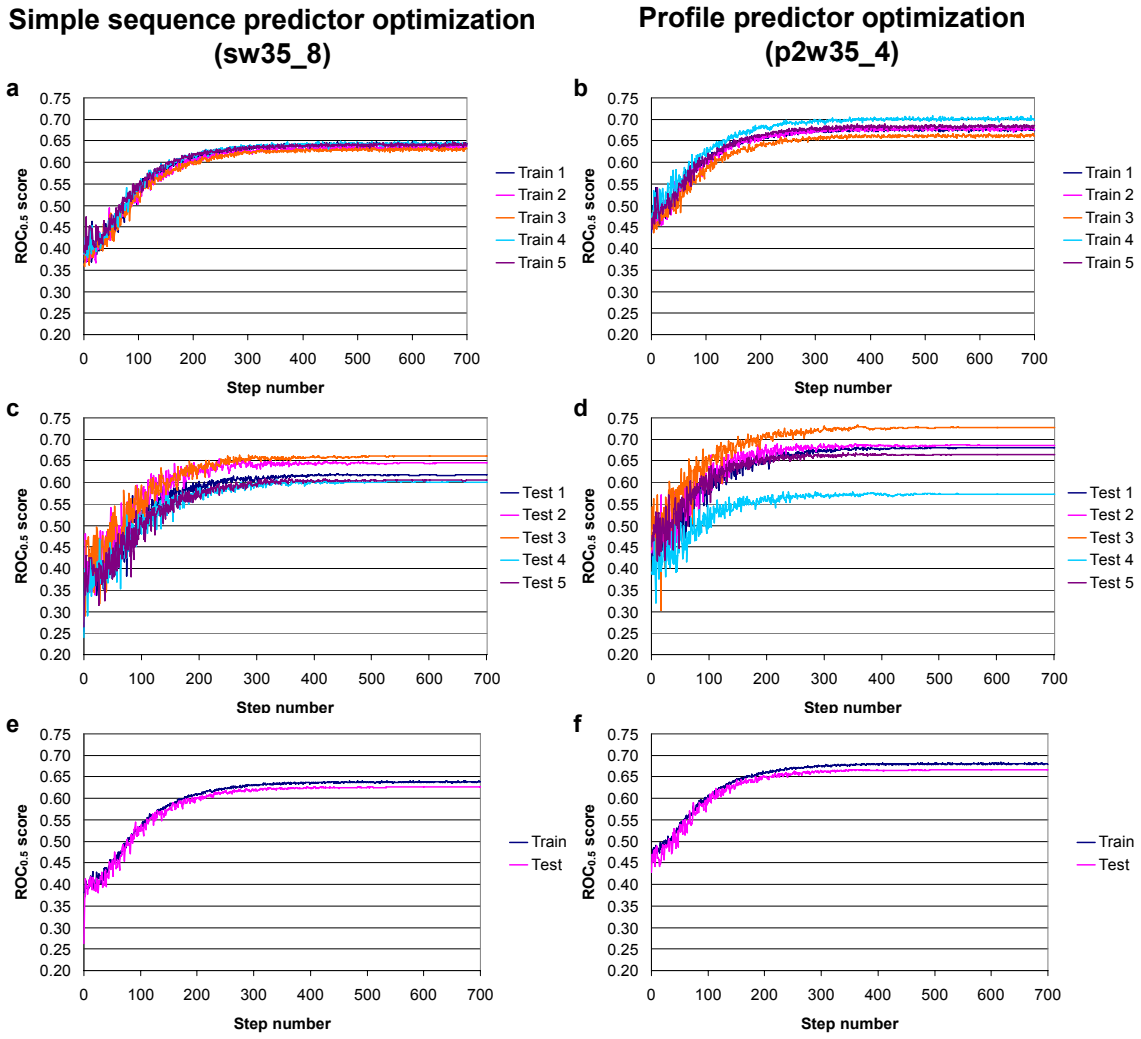

Figure 4.2-3. Score progressions over sw35_8 (simple sequence) and p2w35_4 (profile) predictor optimizations. Panels a, c, e are for simple sequence predictor, and b, d, and f are for profile predictor. a, b) Training score progressions for individual optimization runs. Scores are average begin-of-cycle scores at each temperature step. (When parameter sets, used for testing—see b—are recorded, a parameter set is recorded prior to 'annealing'—set 0—and then parameters are saved at the end of each temperature step—set 1 and the end of temperature step 1, etc. Thus, to keep consistent with testing scores—see b,c—since the score represents an average of begin-of-cycle scores through all cycles in a temperature step, the graph locates the score at a temperature step value of 0.5 less than the actual temperature step number.) c, d) Testing score progressions for individual test sets—$ROC_{0.5}$ scores on test sets using parameter sets 0 (the beginning of optimization, 'randomized' parameter set) through 700 (the end-of-optimization parameter set), which are not 'normalized' as done for the final parameter set (see section 2.5.2). c) Average training and testing score progressions.

4.3 PROFILE VS. SIMPLE WINDOW

4.3.1 Residue type

When excluding sequence ends from evaluation of predictor performance, performance of the profile predictor (p2w35_4) appears to be modestly better than that of the simple sequence-based predictor at low specificity cutoffs (Fig. 3.2-1a). However, the statistical significance of this difference is questionable (Fig. 4.3-1a). On the other hand, the profile predictor performs substantially better than the simple sequence-based predictor when sequence ends are included in performance analysis (Fig. 3.2-1b), and this difference is statistically significant (Fig. 4.3-1b; see discussion of paired t-test, section 2.6.2).

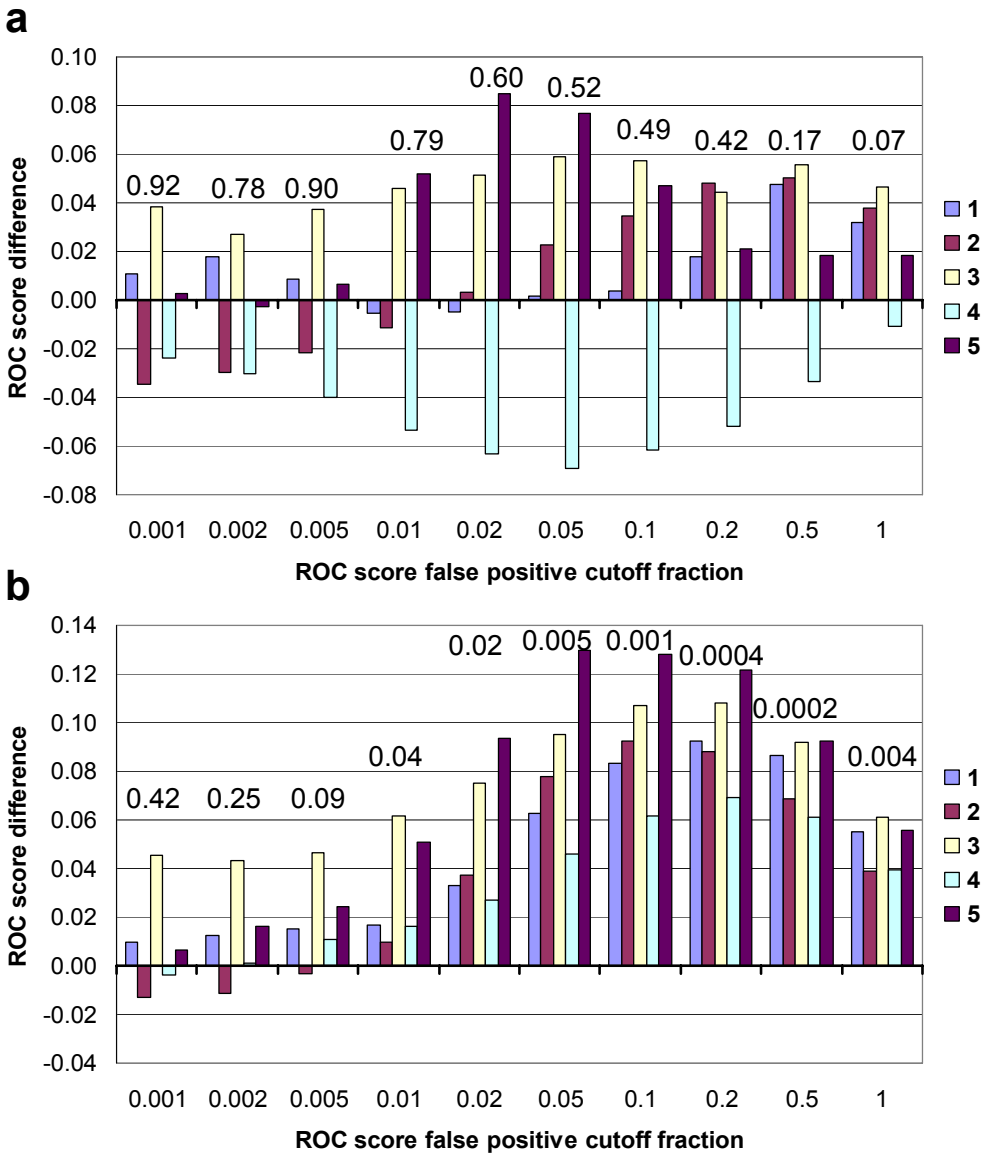

Figure 4.3-1. ROC score differences for individual test sets, for the profile predictor (p2w35_4) vs. the simple sequence-based predictor (sw35_8) (p2w35_4 – sw35_8). Numbers in plot area represent p-values calculated from a two-tailed, paired *t*-test for sets of ROC scores from both predictors at some false positive cutoff. a) With exclusion of thirty residues at the sequence ends. b) With the inclusion of sequence ends (except for ends containing polyhistidine tags).

Profile window predictor (p2w35_4) optimized residue disorder parameters differ significantly from residue disorder parameters optimized for the simple sequence predictor (see fig 4.3-2, 4.3-5). The source of these differences is not immediately obvious, but a close look at this problem yields some insight into how an alignment-based profile may contribute to understanding of a given sequence region.

The profile log odds ratio's correlate better with the simple sequence predictor optimized values (Fig. 4.3-3) than with the profile predictor optimized values (Fig. 3.2-7b), and even better yet with simple sequence log odds ratio's (fig. 4.3-4). There are likely some unique aspects of prediction when using profiles—ways that the predictor uses profiles (beyond simply looking at residue frequencies that more accurately reflect the environment of a particular region of sequences) that somehow enhances prediction enough to justify a significant deviation of optimized residue values from log odds ratios. Presumably, the profile takes special cases into more account, and this improves performance overall but also makes performance more disparate from case to case. This appears to be reflected in that there is more variation in training and testing scores from data subset to data subset than for the simple sequence predictor (see Fig. 4.2-3). Interestingly, the profile predictor does not appear to have much greater overfitting than the simple sequence predictor (see Fig. 4.2-3, panels e, f).

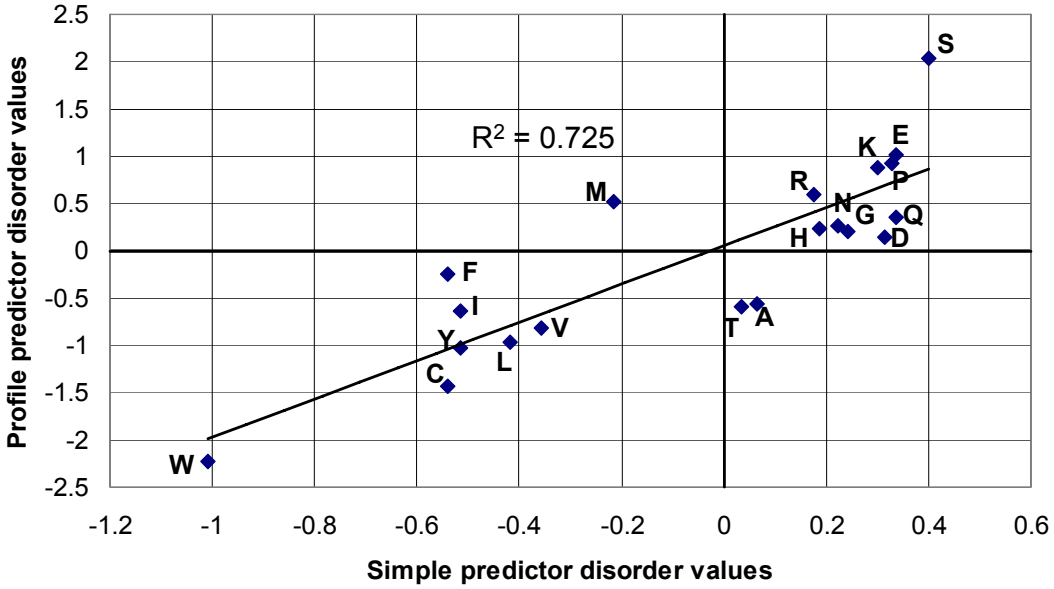

Figure 4.3-2. Average optimized residue disorder values for profile predictor vs. those for simple sequence predictor. A least-squares fitted trend line is provided to give an idea of the direction and size of significant deviations.

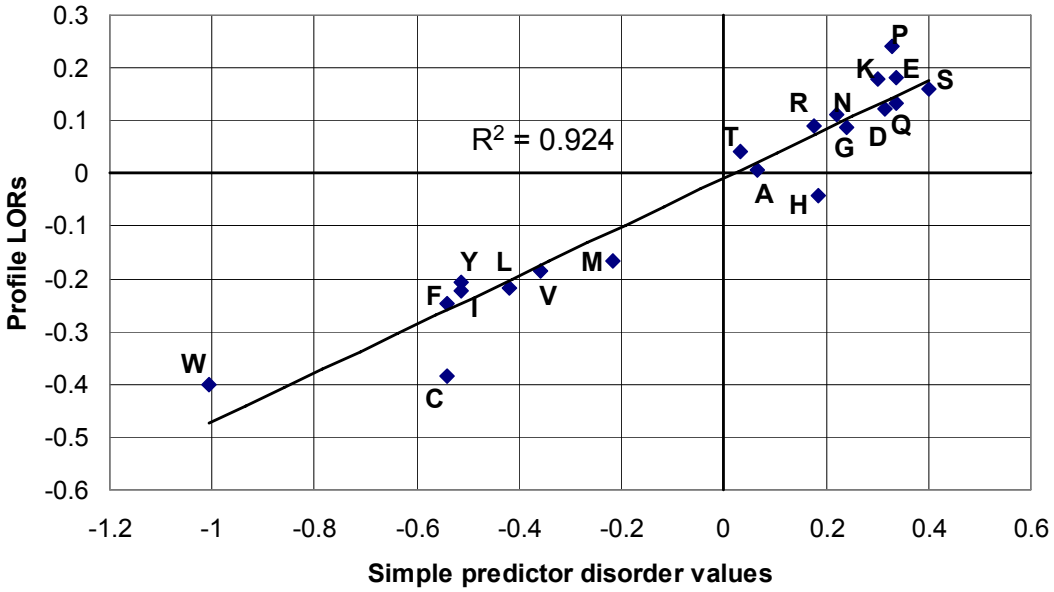

Figure 4.3-3. Correlation of log odds ratios (disordered vs. ordered) of frequencies for different residues in profiles with average optimized simple sequence predictor disorder values.

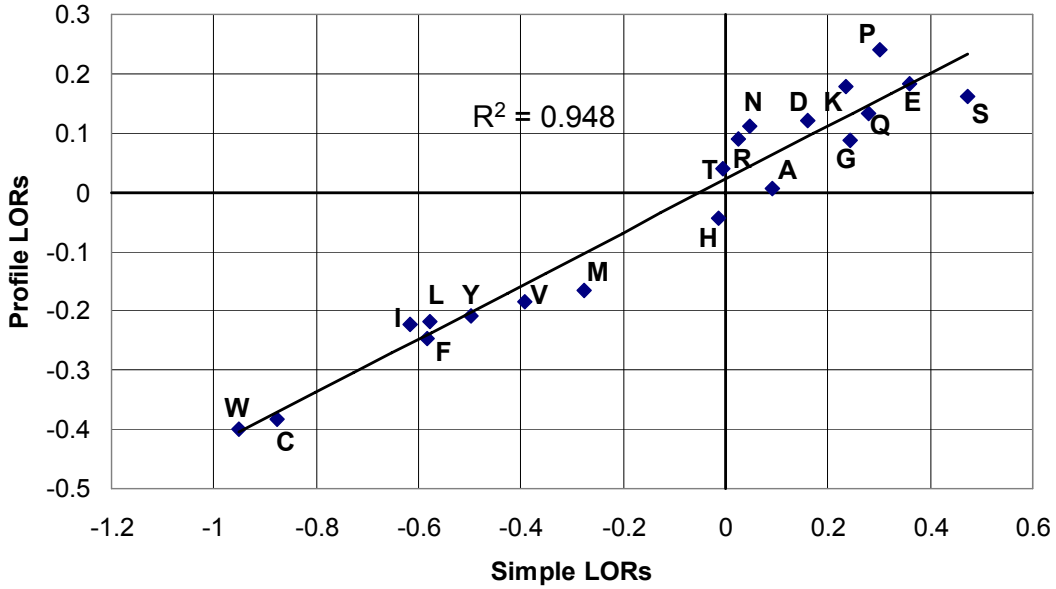

Figure 4.3-4. Correlation of log odds ratios (disordered vs. ordered) of frequencies for different residues in profiles with average with log odds ratios of residue frequencies in simple sequences (note that two different datasets were used—these reflect the frequencies in the actual training sets used for the profile and simple window predictors.) Note that cysteine is close to the trend line here as opposed to when comparing log odds ratios and optimized disorder values (Fig. 4.3-2).

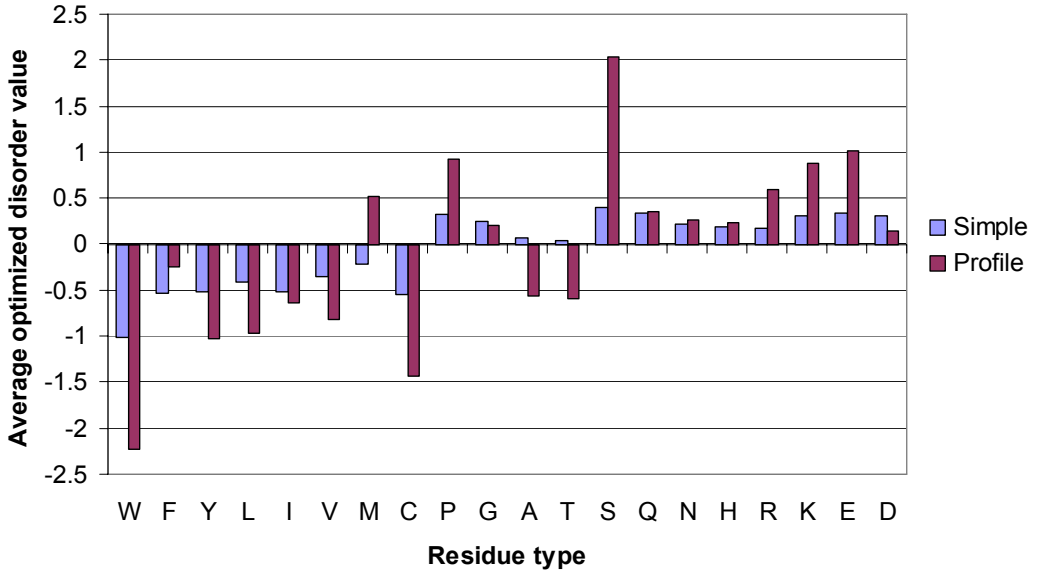

Figure 4.3-5. Comparison of optimized disorder residue type parameters (averaged over the five optimized parameter sets) for simple sequence and profile window predictors, normalized to yield residue score distributions that approximate the standard normal distribution (mean = 0, standard deviation = 1).

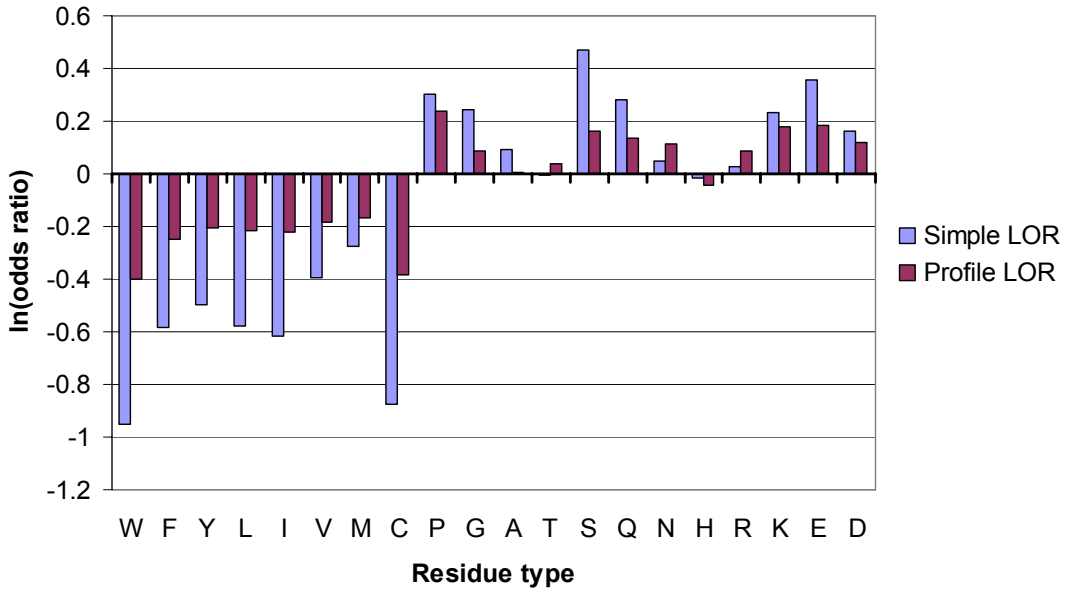

Figure 4.3-6. Comparison of average disorder vs. order 'log odds ratios' for different residue types, calculated from simple sequences and profiles.

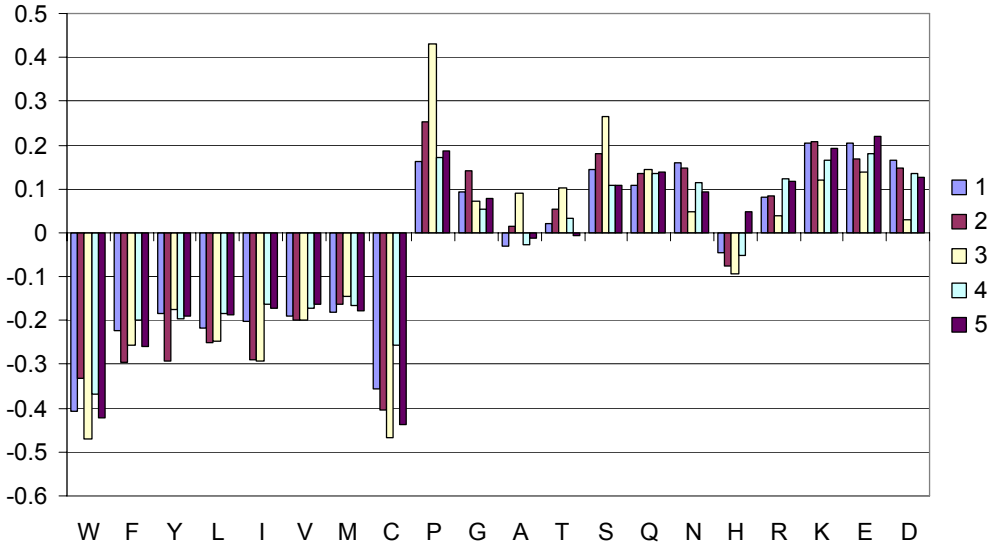

Figure 4.3-7. Log odds ratio values of different residue types' frequencies in disordered vs. ordered regions, calculated for the five standard profile test sets. Note that in test set 3, values for residue types including proline and serine stick out (see section 2.4 1).

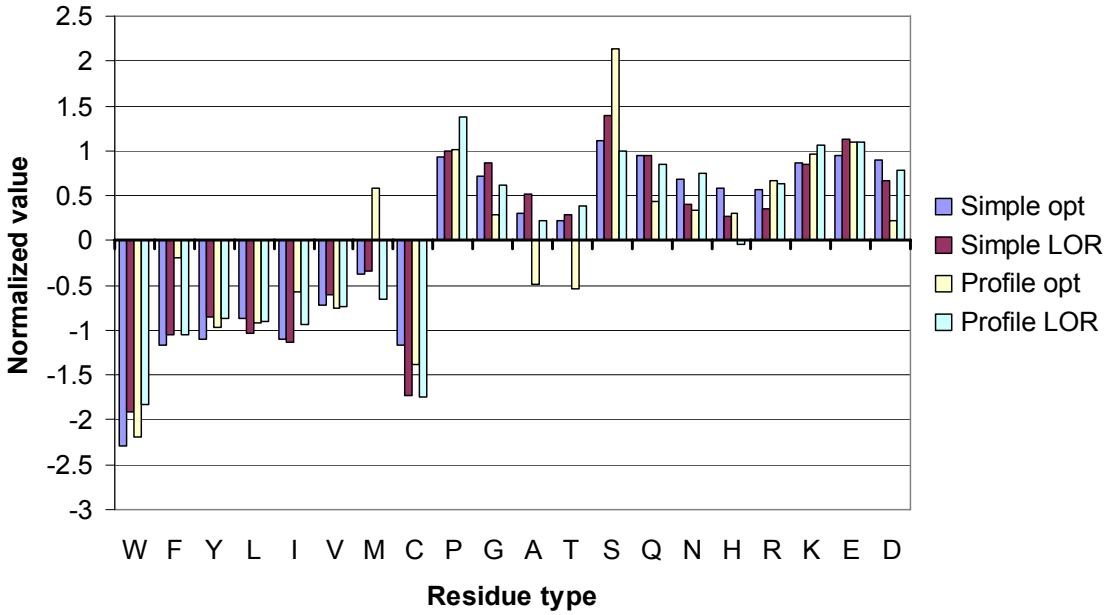

Figure 4.3-8. Average disorder vs. order log odds ratios and average optimized values for profiles and simple sequences, all normalized so that each set of values has a mean of 0 and a standard deviation of 1, to allow for comparison.

In a sequence alignment, some positions are, of course, better conserved than others. Highly variable positions in the alignment will not yield much information, but well-aligned positions, in which a certain subset of residues is more frequent than others, will yield more signal. If a residue's disorder value deviates significantly from what would be expected from frequencies, it may not have as strong an effect in positions where it is simply making some background contribution to the disorder score (particularly when the disorder values of other residues that tend to co-occur with it balance it out), as when it occurs more strongly in a position where it is more conserved. The predictor is likely taking advantage of more conserved positions. Some residues that appear to be significant outliers, with some reasonable explanation, are discussed below:

4.3.1.1 Serine

Serine shows a dramatic increase in value relative to other parameters, in the profile-optimized parameters, as opposed to those optimized using simple sequences. However, when simply comparing the disorder vs. order log odds ratio for serine in profiles with that for serine in simple sequence, it actually appears to drop (see Fig. 4.3 4). The values of threonine and alanine (the two most substituted residues for serine according to the BLOSUM matrices, threonine being the most), on the other hand, appear to drop significantly.

Consider a position where serine is conserved. There also may be a number of threonines in that position. If the threonine to serine ratio is higher, the position might be expected to be more likely to be ordered, and vice versa. Thus, if the disorder value for serine is increased and the disorder value for threonine is decreased, then threonines at a

serine/threonine position would counterbalance serine in an ordered region, but in a disordered region, with a lower ratio of threonine to serine, the effect of serine's high disorder score would be significant. Such a counterbalancing effect appears to be indeed taking place. If one starts with the original, optimized parameters for a profile predictor and perturbs the serine value, then one would expect that a counterbalancing residue's disorder value, if re-optimized, would change in the direction opposite to the serine. If there is no strong coupling between the two residue types, on the other hand, the average change in the value of the residue being re-optimized would be expected to be near zero. This indeed appears to be the case for threonine (see fig. 4.3-9). It was expected to occur for alanine also, but the effect, although statistically significant, was not strong like the one for threonine.

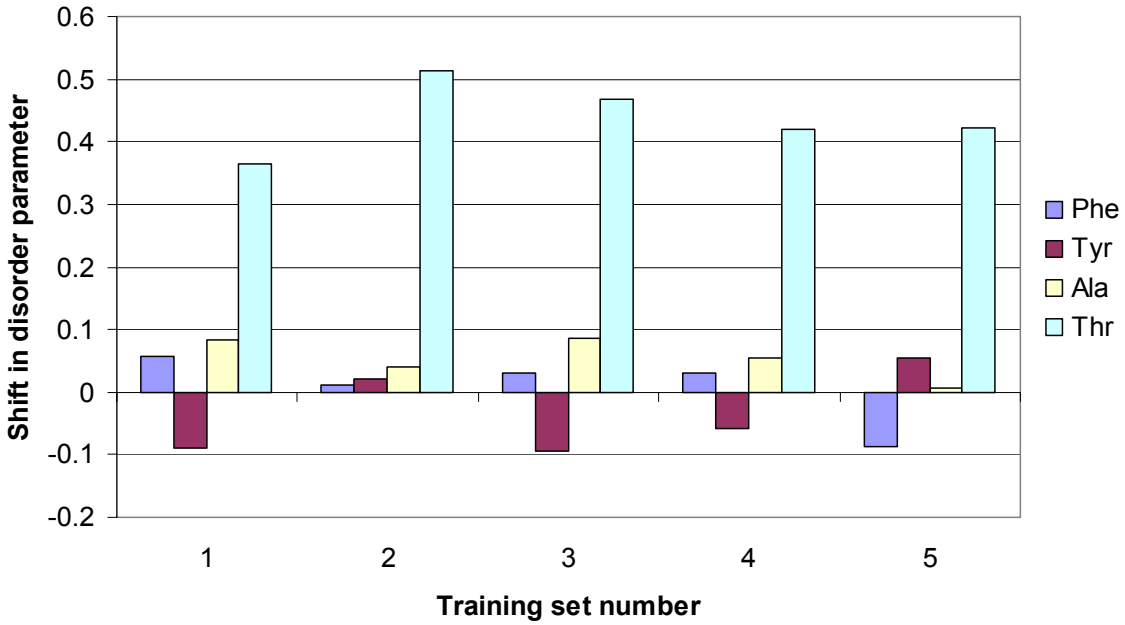

Figure 4.3-9. Shifts in disorder parameters for different residue types when serine is perturbed by a value of -1 (bringing it close to the value it might be expected to have according to optimized simple window values – see fig. 4.3-1 – or from logs odds ratios of residue type frequencies in ordered or disordered regions). Two experiments were done. In one, the disorder parameters for Ala and Thr were freed to change; and in the other, the values for Phe and Tyr were allowed to change. It was expected that Thr and Ala would show significant shifts, while Phe and Tyr were included as controls. The *p*-values for differences, calculated from two-tailed, paired *t*-tests are as follows: Phe: 0.75; Tyr: 0.33; Ala: 0.021; Thr: 0.000063.

One might propose that the dramatic increase in serine's optimized disorder value for the profile predictor vs. the simple sequence predictor is largely due to the existence of polyserine stretches in disordered regions, allowing easier alignment of serines. The number of instances of three or more consecutive serines was counted in the filtered, nonredundant sequence set that was used in obtaining the profiles. (Note that, since it was filtered for low complexity sequence, very long polyserine stretches would not be included in the dataset.) The number of such occurrences was 548,122, and the total number of serines in such stretches was 1,767,602. The results for serine could be compared with those for glycine, which experienced a shift in the opposite direction from serine in its optimized disorder value. The number of occurrences of three or more glycines in a row was 411,817, and the total number of glycines within such stretches is 1,317,297. One would expect that polyglycines, like polyserines, would occur often in disordered regions.

4.3.1.2 Methionine

When window-only predictors were trained, residues were excluded from performance evaluation if they were within 18 residues of the termini. This effectively removed any influence of the N-terminal methionines on the prediction, and therefore, the prediction parameter value for methionine reflected the effects of non-N-terminal methionines. When profiles were optimized, however, this did not exclude the effects of the N-terminal methionines in sequences related to the query sequence, whose N-terminal methionines aligned to a position on the query sequence close to, but not at, the start of the

query sequence. The fact that the profile parameter for methionine was relatively much higher than that of the simple window parameter thus makes sense.

4.3.1.3 Glycine

Glycine may be significantly concentrated above its background frequency in positions where it serves to act as, say, a helix-breaker, or otherwise facilitate some kink in a structure. In this way, glycine would be promoting a consistent structural feature. One might expect, then, that well-conserved glycines, not as easily substitutable as random glycines, would have a higher order to disorder log (odds ratio) than random glycines.

4.3.1.4 Phenylalanine/Isoleucine

The dramatic increase in the residue value of phenylalanine, but not tyrosine, might be explained by the predominant existence of phenylalanine, in comparison to other hydrophobic residue types, in some loops. That the same may also be true for isoleucine is suggested by increases in isoleucine's residue value, as well as some similarity in shape between phenylalanine and isoleucine. As with serine, the ratios of phenylalanine and isoleucine to other hydrophobic residue types in loops can be compared with those in helices and strands.

4.4 PREDICTORS WITH TAIL ADJUSTMENTS

For both the simple sequence-based and profile-based predictors, addition of tail adjustments produces dramatic change in predictor performance when considering the sequence ends in the predictor performance (see Fig's 3.2-1, 4.4-1).

The tail adjustments serve as an example of how the optimized parameters of a simple predictor might help to detect biases in the data/optimization. It was noted in a plot of tail adjustment parameters from an older run, sw35_st30_5 (see Fig. 4.4-2; code may be made available on the web at http://prodata.swmed.edu) that there was a marked leveling off of parameters among the middle positions before continuing to fall off as position from terminus increased. It was hypothesized that this was due to the how it was being determined whether or not residues were counted in performance analysis, in that when a polyhistidine tag was detected, the residues terminal to and through the polyhistidine stretch were not counted, but the first residue internal to the polyhistidine stretch could be counted in analysis of performance. Due to apparent enhancement of terminal disorder by polyhistidine tags, a residue immediately internal to a polyhistidine stretch might then be expected to have a higher likelihood of being disordered/missing than another residue equally distant from the sequence end, but without a polyhistidine stretch in proximity. Thus, two different curves might be obtained for tail adjustments, depending on whether sequence ends with polyhistidine tags were counted in optimizations or not, with the curve being shifted upward if polyhistidine tag-containing ends were included. Thus, it was hypothesized that the leveling off in the tail adjustments was due to a combination of exclusion of histidine stretches and residues external to these stretches and inclusion of residues internal to the

polyhistidine stretches. An adjustment to the treatment of the data was made so that at sequence ends where polyhistidine stretches were detected, at least the first thirty residues on that end of the sequence were excluded from analyses of performance. This appeared to reduce the leveling off effect (see Fig. 4.4-2).

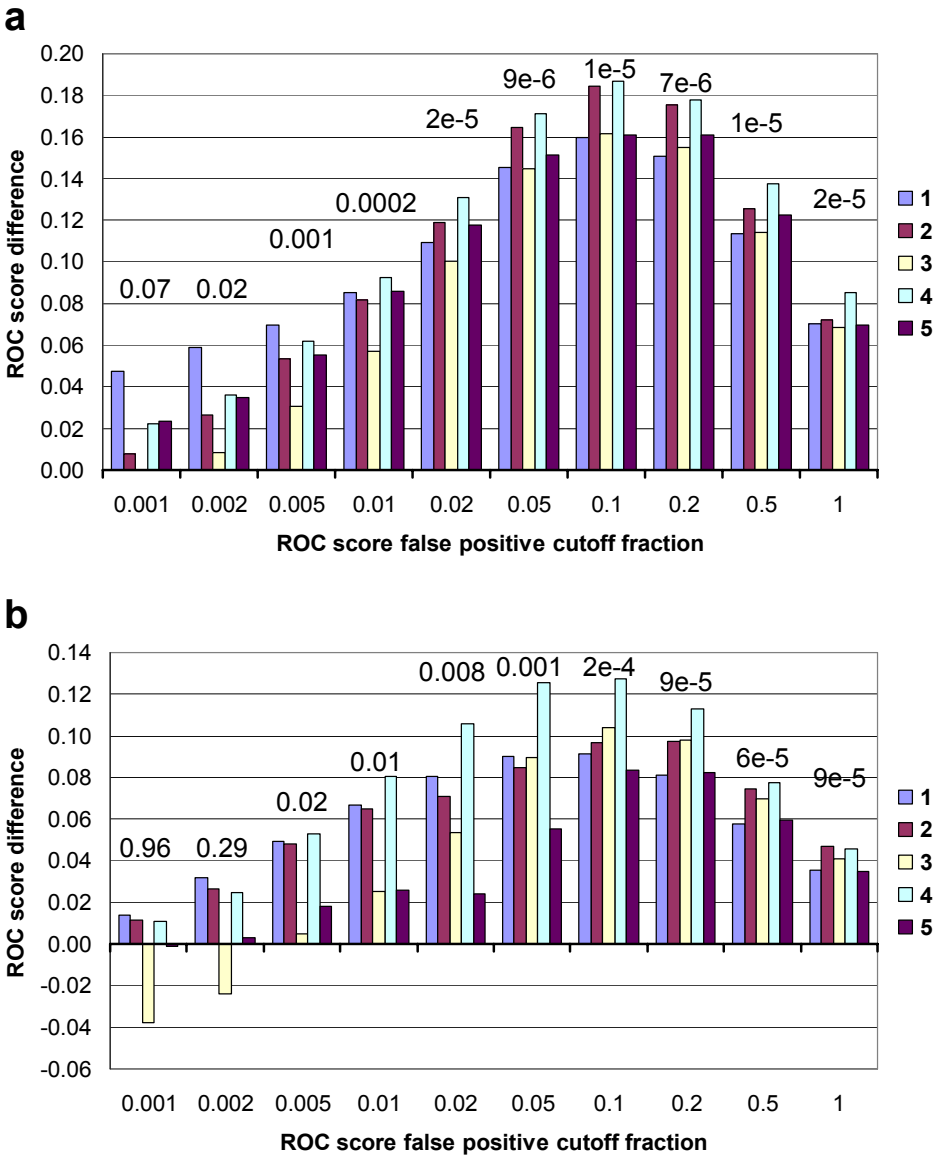

Figure 4.4-1. ROC score differences for predictors with tail adjustments vs. respective predictors without tail adjustments. Numbers in plot area represent p values calculated from a two-tailed, paired t test for sets of ROC scores from both predictors at some false positive cutoff (t-test is not a perfect statistical measure in estimating p-values for ROC score differences: see section 2.6.2). a) With exclusion of thirty residues at the sequence ends. b) With the inclusion of sequence ends (except for ends containing polyhistidine tags).

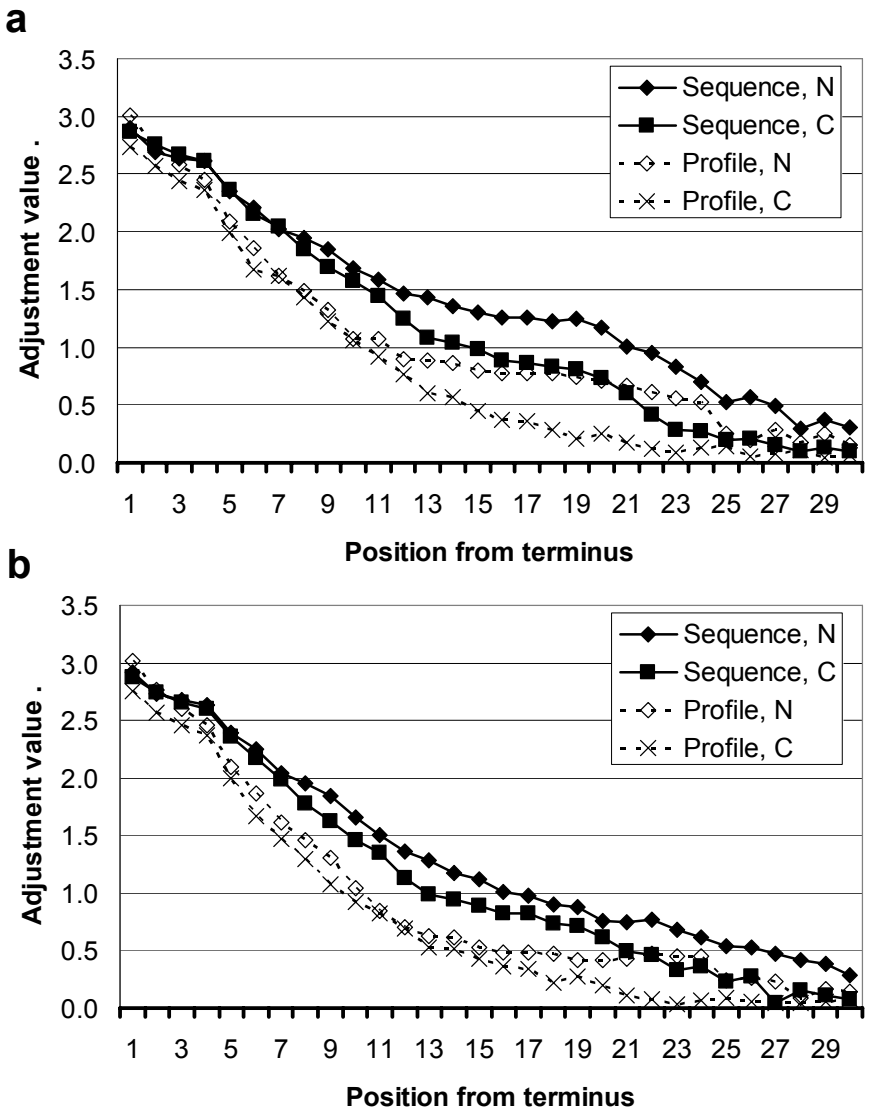

Figure 4.4-2. Tail adjustment parameters from different predictors. Allows comparison between parameters optimized with different treatment of sequence ends that contain polyhistidine stretches. a) From predictors that were optimized with 'old' method, in which residues external to and including polyhistidine stretches were excluded from analyses of performance, but residues immediately internal to such polyhistidine stretches could be included. From sequence-based predictor, sw35_st30_5 and from profile-based predictor, p2w35_st30_4. b) From predictors that were optimized with 'new' method, in which, in the case of a sequence end where a polyhistidine stretch was detected, the first thirty residues from the end were automatically excluded from analyses of performance. From sequence-based predictor, sw35_st30_8 and from profile-based predictor, p2w35_st30_6. (See Table 2.7-1 for a list of optimization runs.)

4.5 HIGH SPECIFICITY PREDICTOR

The performance measure used to optimize the predictors is the ROC_f score, where f is the fraction used to determine the false positive cutoff in calculating the score. The standard predictor (sw35_8) was optimized using a $ROC_{0.5}$ score. Essentially, this means that optimization considered the ability of the predictor to sort disordered from ordered residues for residues that had higher-than-average scores. Disordered regions in the low-specificity range (with low scores) were not considered, a rationale for this being that low-scoring disordered regions might be expected to represent 'noise' or regions that tend to be different from typical normal disordered regions in their cause of disorder. Some justification for this might be found in differences in tryptophan disorder residue parameters for the 9-position window predictors optimized with $ROC_{0.5}$ (Fig. 5.3-3b) and $ROC_{1.0}$ (Fig. 5.3-3c) scores.

On the other hand differences might also be expected between a predictor optimized using the $ROC_{0.5}$ score and a 'high specificity' predictor, like one optimized using a $ROC_{0.05}$ score (sw35_7). The optimization of the high specificity predictor considers only the sorting of disordered vs. ordered residues among those residues with the highest scores. Differences between optimized parameters for the high specificity predictor and the standard predictor are briefly discussed here.

4.5.1 Performance

The high specificity predictor (sw35_7) modestly better than the standard predictor (sw35_8) in the high specificity region (see Fig. 4.5-1), and modestly worse in the low specificity region. Even though the differences appear to be small, it may be better to use

such a predictor if looking for longer disordered regions. As in other cases, test set three (see section 2.4.1) appears to have special behavior (see Fig's 4.5-1c, 4.5-2).

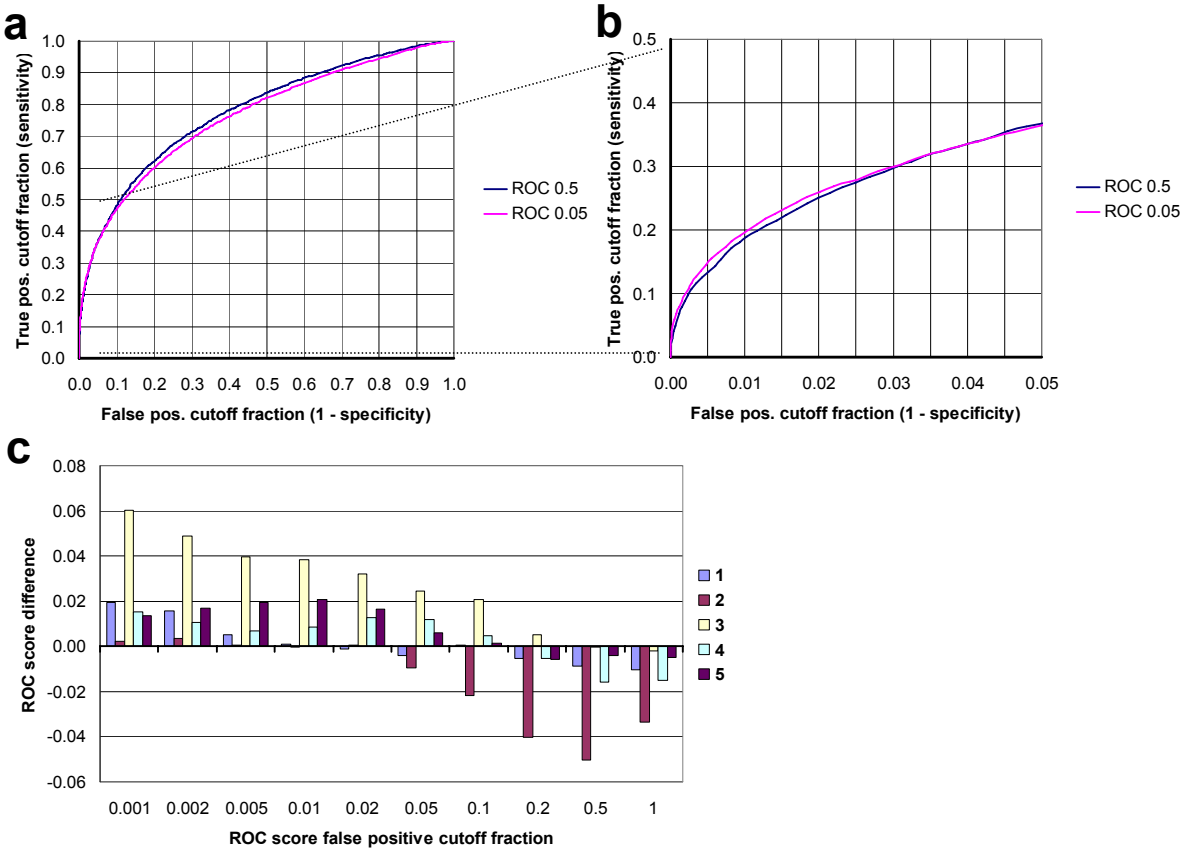

Figure 4.5-1. ROC curves and differences for high specificity (sw35_7), standard (sw35_8) predictors. a) Average ROC curves, full scale. b) Average ROC curves, different scale. c) ROC score differences at different false positive cutoff fractions (sw35_7 – sw35_8).

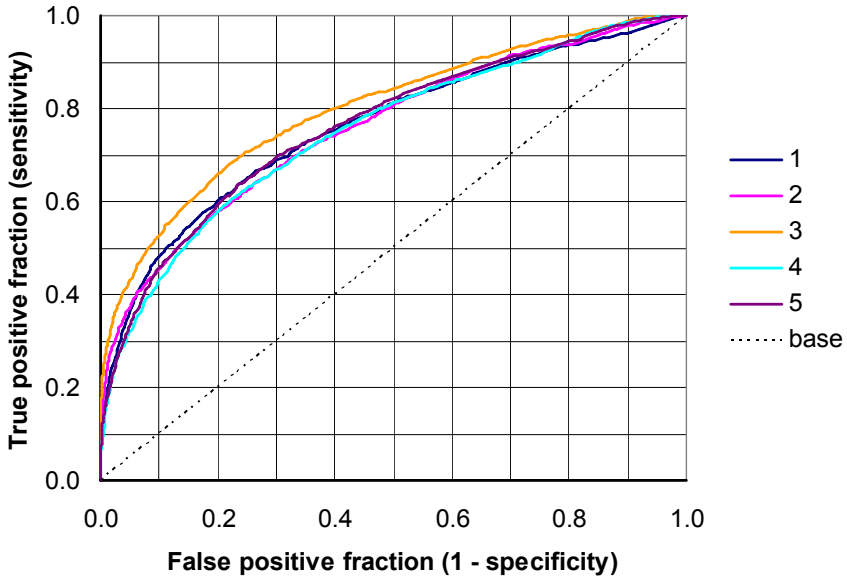

Figure 4.5-2. High specificity curve individual test ROC curves.

4.5.2 Residue disorder values

Residue disorder values for the high specificity and standard predictors are generally similar (see Fig's 4.5-3, 4.5-4). The primary exception is tryptophan, and the next strongest outlier is histidine. Even though performance measurements appear to be significantly affected by test set 3, differences in parameters from those of the standard predictor do not appear to be largely attributable to the special behavior of test set 3, as the disorder values resulting from the optimization of training set 3 (which excludes test set 3; see section 2.3 on cross validation) do not appear to be significant outliers (see Fig. 4.5-5). That tryptophan is an outlier makes sense, given that tryptophan is the most hydrophobic (and thus, the most order-promoting and 'stickiest') residue. Tryptophan might thus disrupt the normal behavior of a long disordered loop, and therefore, sequences of long disordered regions might reasonably tend to exclude it. It is not very clear why histidine is such an outlier.

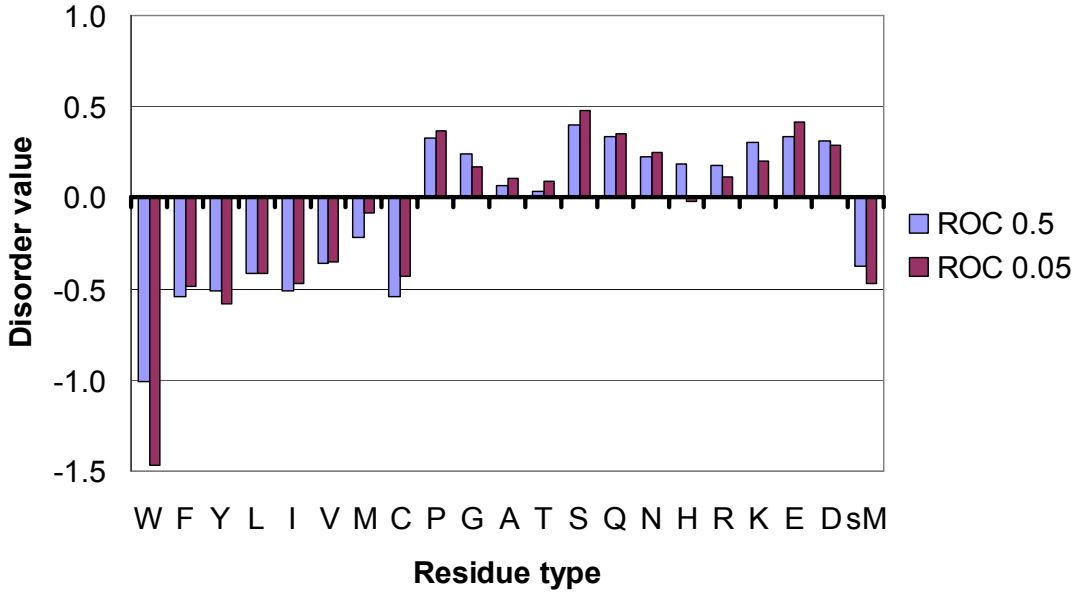

Figure 4.5-3. High specificity ($ROC_{0.05}$-optimized) and standard ($ROC_{0.5}$-optimized) residue disorder values.

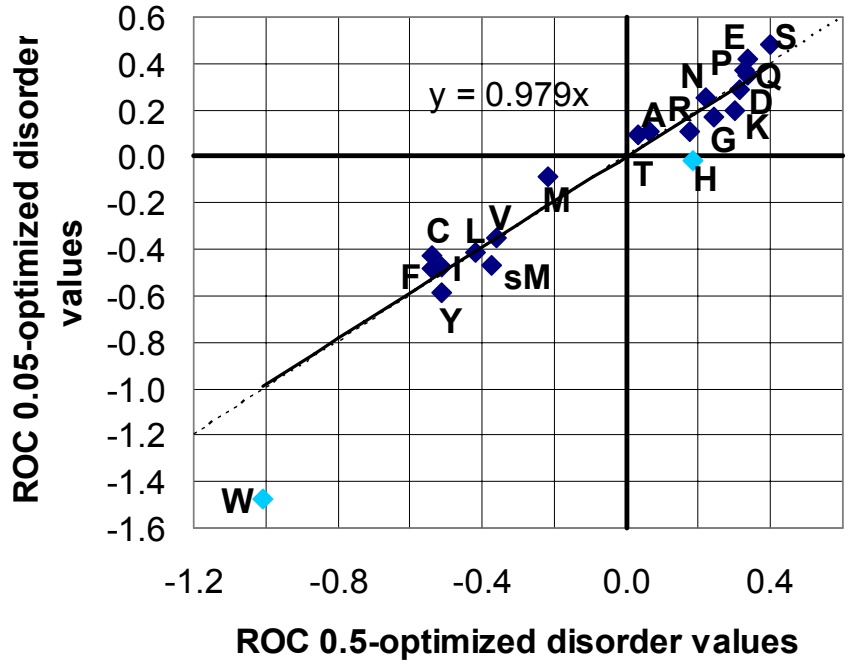

Figure 4.5-4. Correlation plot of high specificity ($ROC_{0.05}$-optimized) vs. standard ($ROC_{0.5}$-optimized) disorder values. Dotted line: y = x; solid line: fit to residues marked by dark blue diamonds (all except W and H)—equation shown on plot (intercept constrained to 0).

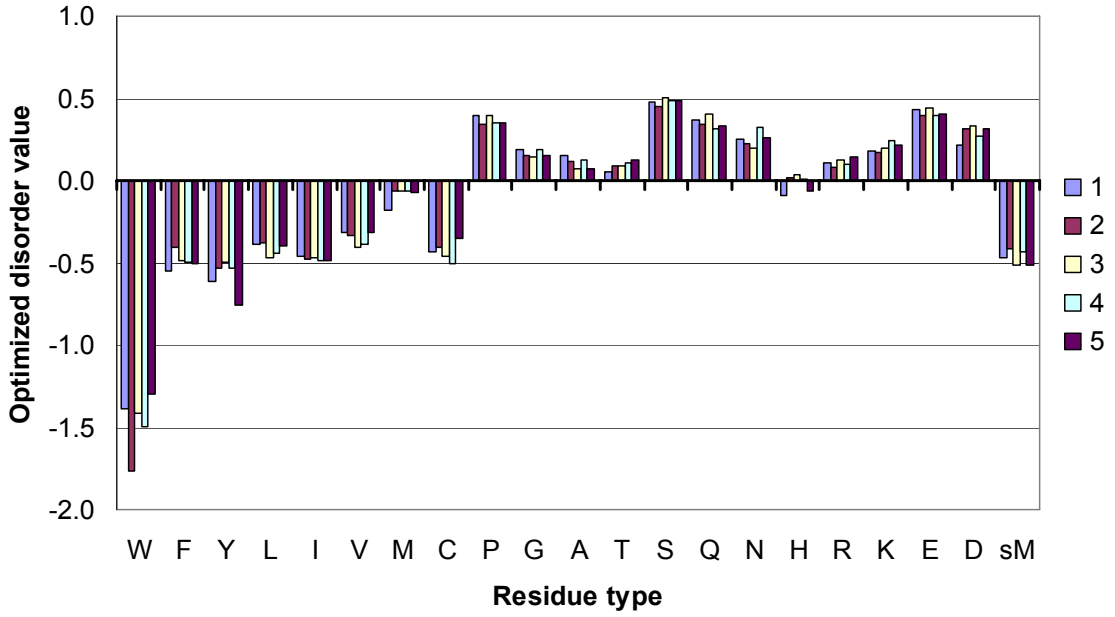

Figure 4.5-5. High specificity predictor (sw35_7) normalized optimized disorder values for standard residue types and selenomethionine (sM).

4.5.3 Window position weights

A primary difference between the high specificity predictor and the standard predictor is in the window position weights (Fig. 4.5-6). The window weight pattern for the high specificity predictor is broader, as would be expected, since long low-hydrophobicity stretches of residues would generally be more likely to be disordered than short ones. Thus, the high specificity predictor might be considered more of a predictor of long disordered regions (see Peng et al. 2006 regarding separation of prediction of long and short disordered regions).

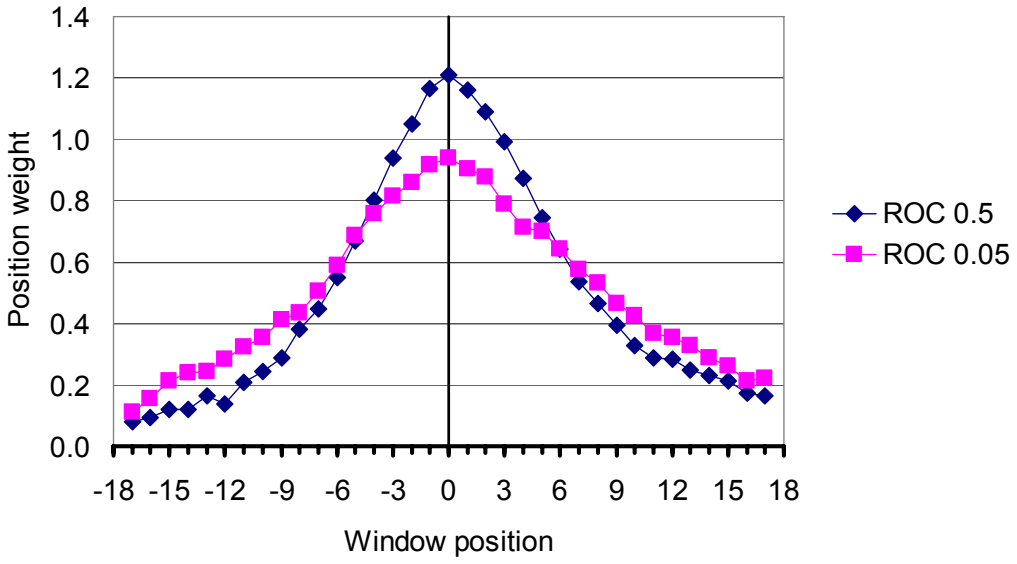

Figure 4.5-6. Comparison of standard predictor and high specificity predictor window position weights. The standard predictor was optimized using $ROC_{0.5}$ scores, and the high specificity predictor was optimized using $ROC_{0.05}$ scores.

4.6 OTHER PREDICTION METHOD ATTEMPTS

Although they will not be covered in detail, other less successful measures were tried in attempting to augment or improve prediction. Previous to the work described, attempts were made to use a double sliding window method, in which a short window was run over initial (simple sequence-based) disorder values and then a 'cooperativity adjustment' was made, essentially with the intent to enhance the effect of strongly disorder-promoting regions, and a second, larger window was then passed over the adjusted values. Any improvement gained from this 'double window' method was quite modest. Also, there were attempts with the simple sliding window and double window methods, to simultaneously optimize low- and high-specificity scorers, where the high specificity score was counted for high-scoring positions, and the low-specificity score was otherwise used. This did not yield

notable improvement—indeed, for the single window method, performance dropped, suggesting poor optimization.

Also, inclusion of residue-residue statistics was tested using full and reduced ‘alphabets’. This was tested for ‘dimers’ (R-R permutations) without any substantial improvement in performance with the variants tried, in general agreement with the results of Weathers et al (2004). Perhaps consideration of R-x-x-R or R-x-x-x-R statistics might have yielded more improvement in performance, given the potential for interaction between residues spaced thus in helices, but this was not tried.

CHAPTER FIVE
Disorder/hydrophobicity association and other residue type-related issues

The tight linear association between disorder and hydrophobicity has significant implications not strongly supported by previous associations of disorder with 'hydrophobicity' (Uversky et al. 2000; Williams et al. 2001; Dostanyi et al. 2005; Linding et al. 2003a). Relationships were discovered initially through visual inspection of correlation plots of various scales with the optimized disorder values, including the Nozaki-Tanford scale, Wimley-White scales, and Radzicka-Wolfenden scales. No single hydrophobicity scale is perfect. Careful study of these scales, however, and a subsequent search of AAIndex1 (Kawashima et al. 1999), using knowledge gained from visual inspection of other scales, helped to confirm the idea that hydrophobicity is indeed strongly associated with disorder, better than any other property.

Such a tight relationship begs explanation. This relationship suggests that a model of general crystallographic disorder can and should be linked directly to the concept of hydrophobicity. Significant deviations from the relationship may signify some special, residue type-specific effect on disorder. Disorder is qualitatively related to other properties, presumably generally through their associations with hydrophobicity, but these properties should not necessarily be used as a direct basis for inferring the nature of crystallographically disordered regions.

Before optimizing parameters for a predictor, it might be argued why certain parameters should work well, but intuition may not always provide parameters that perform near-optimally. The standard optimized residue disorder values (averaged,

normalized/adjusted results from sw35_8) are compared with various other possible sets of residue disorder values that might have been used, including values obtained from relative frequencies of residues in random coil vs. other secondary structures (coil propensity scales), values estimating hydrophobicity (hydropathy scales), and decreases in performance are discussed.

Here, I also further discuss evidence that the tight disorder/hydrophobicity association is real and is not just specific to the Nozaki-Tanford scale. I discuss why certain deviations may appear in certain scales and specific things that have been done that help to disclose the tight association. Discussion is added on issues specific to certain residue types and on interpreting the linear relationship (quantified) between disorder and hydrophobicity.

5.1 DIFFERENCES IN SCALES FROM OPTIMIZED DISORDER VALUES

In this section differences in coil propensity and other scales from optimized disorder values are presented. These scales might be expected to be well associated with disorder, but are not. In this section will also be shown in more detail the effects on predictor performance of various types of differences in scales from optimized disorder values.

5.1.1 'Coil propensity' scales

Various coil propensity scales have been constructed, calculated from a residue's propensity for being in 'random coil' segments of structure vs. segments identified as having some other specific secondary structure—typically helix or strand. None of the coil propensity scales discussed in this section show good correlation with average optimized disorder values (Fig. 5.1-1), highlighting that there is a significant difference between just

being in a coil region and being in a region that is disordered. Linding et al. (Linding et al. 2003a) also demonstrate a difference between residue composition in coils (by their definition, any structure other than α-helix, β-strand, and 3_{10}-helix as defined by DSSP (Kabsch and Sander 1983)) and remark 465 regions in X-ray crystallographic structures, but do not give a strength of correlation. Thus this provides a unique look at the issue of the difference between disordered and coil regions.

Perhaps the most well-known coil propensity values are those calculated by Chou and Fasman (1974). Chou and Fasman calculated propensities of various residues being within different types of secondary structural elements, including helices, β sheets, and 'coil' conformations. Propensities, P, were calculated by

$$P_{c,i} = f_{c,i} / \langle f_c \rangle$$

where c is the secondary structural conformation and i is the residue type; fc,i is the frequency with which a given residue occurs in one conformation vs. the other conformations; and ‹fc› is the average frequency of residues a given conformation.

Deléage and Roux (1987) developed a secondary structure prediction algorithm that used parameters similar to those of Chou and Fasman, with β-turns in their own category (see Fig. 5.1-1b).

The recommended residue coil propensity values for the globularity/disorder program, GlobPlot (Linding et al. 2003b), are the Russell Linding propensities, which were calculated from a SCOP superfamily representative set of proteins by subtracting the 'secondary structure propensity' of a residue from its 'random coil' propensity. Perhaps partly due to improved statistical power of the dataset, the (all residue) R^2 value of the

Russell-Linding coil propensity values of 0.326 is better than the Chou Fasman or Deléage Roux R^2 values (0.208 and 0.162, respectively), but it is still worse than any of the hydrophobicity-scale R^2 values, and its overall performance in disorder prediction is only slightly better than that of the Chou-Fasman values (see Fig. 3.2-9, Fig. 5.1-3).

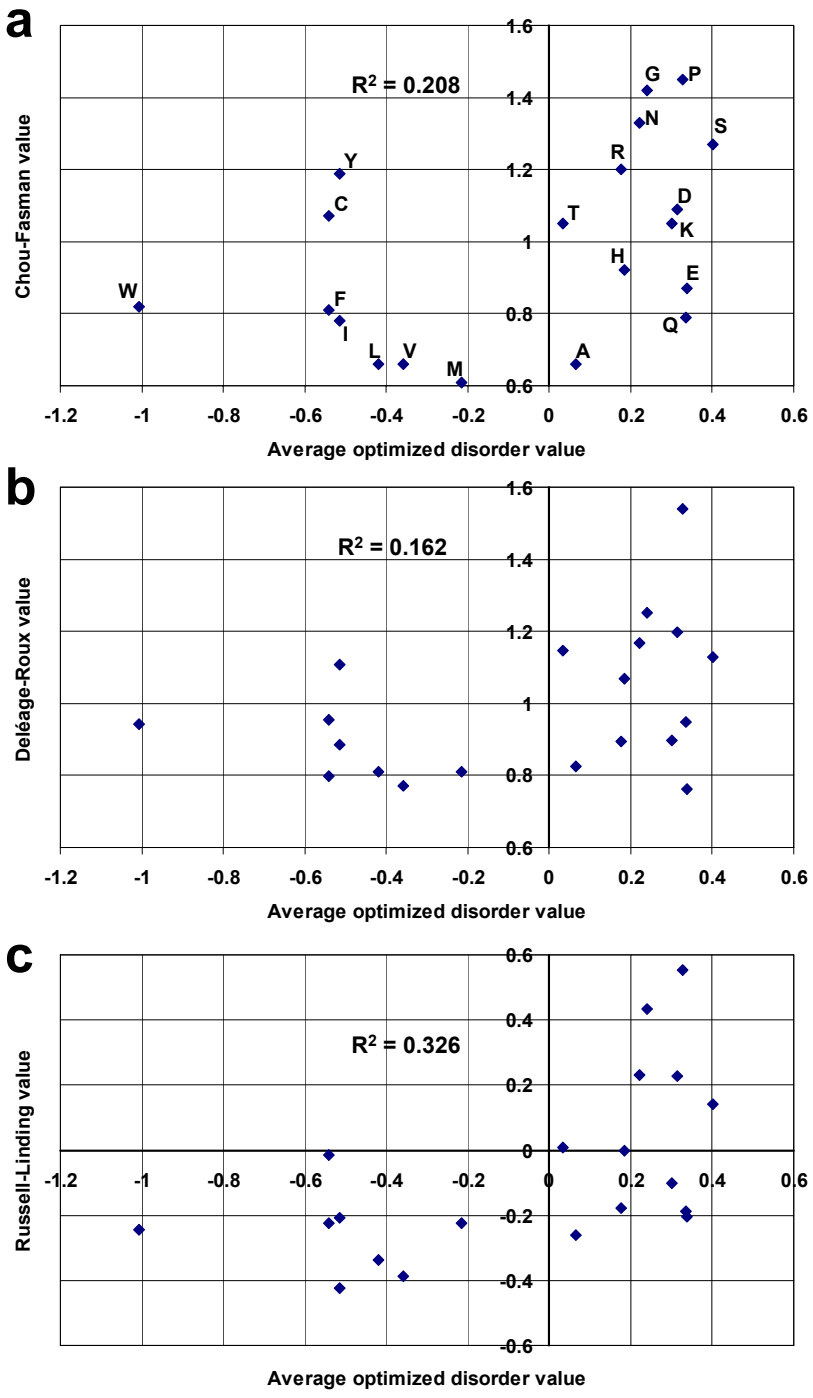

Figure 5.1-1. Correlation plots of coil propensity scales vs. optimized disorder values. a) Chou-Fasman coil propensities vs. optimized disorder values. b) Deléage-Roux coil propensities vs. optimized disorder values. c) Russell-Linding coil propensities vs. optimized disorder values (see Fig. 3.2-8a for labeled version).

5.1.2 Kyte-Doolittle and Hopp-Woods

A number of hydropathy/hydrophobicity-related scales have been constructed. Perhaps the best-known scale in this category is the Kyte-Doolittle hydropathy scale (Kyte and Doolittle 1982) (see section 1.2.3). Optimized disorder values are remarkably different from the Kyte-Doolittle hydrophobicity parameters (Fig. 5.1-2a). One striking contrast is that, whereas tryptophan and serine represent the negative and positive extremes, respectively in the optimized disorder parameters, their Kyte-Doolittle values are almost the same, lying near the median of the Kyte-Doolittle scale. Serine actually has a slightly higher value than tryptophan on the Kyte-Doolittle scale (whose original values have a general positive correlation with hydrophobicity, as opposed to the disorder parameters).

Another hydrophobicity-related scale that was based partly on experiment and partly on human judgment was that of Hopp and Woods (1981) (see section 1.2.2). Hopp and Woods took their scale primarily from one published by Levitt (1976), with minor adjustments. Levitt, in turn, took his "hydrophobic parameters" from values obtained by Nozaki and Tanford (1971) using experimental data, where available, and "When experimental values were not available they were roughly estimated from the relationship between accessible surface area (Lee & Richards 1971) and hydrophobicity (Chothia, 1974)". The association between optimized disorder values and the Hopp-Woods scale is generally quite good, except for strong ionic residues. Strong ionic residues are clearly outliers (see fig. 5.1-2b), where the remaining residues have a decent linear correlation with optimized disorder parameters. Among the non-strong ionic residues, cysteine appears to be the most distant outlier, in the direction that would be expected (its disorder value is lower

than what one would anticipate simply from its estimated hydrophilicity—a change in this direction would be expected due to the stabilizing effect of disulfide bridges.)

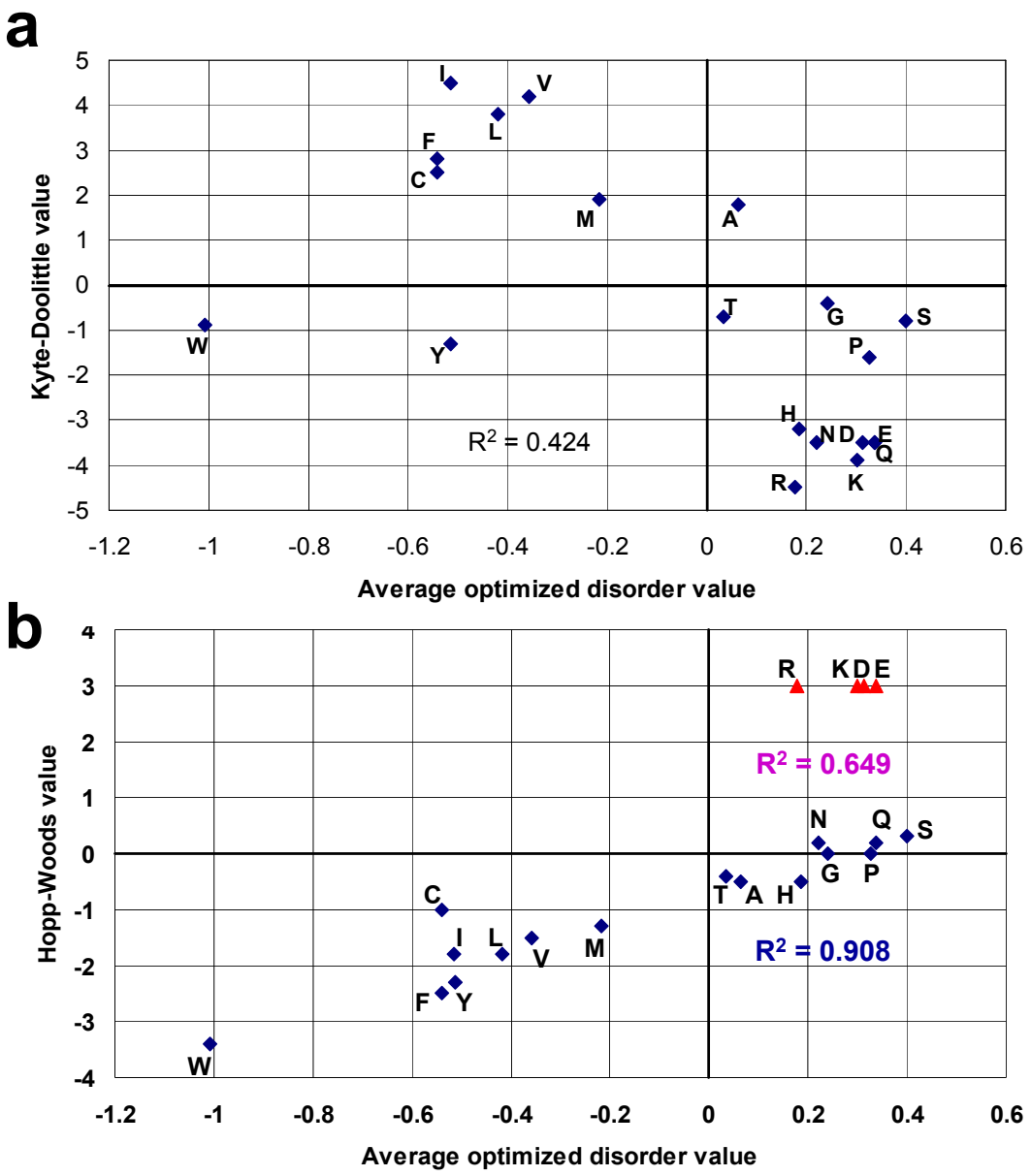

Figure 5.1-2. Correlation of hydropathy/hydrophobicity-related scales with optimized disorder values (sw35_8). a) Kyte-Doolittle. b) Hopp-Woods; red triangles are strongly ionic residues; blue diamonds represent remaining residue types; upper R^2 value (purple) is for all residues, and the lower R^2 value (blue) excludes the strongly ionic values.

5.1.3 Performance of different scales in predicting disorder

When substituting various coil propensity and hydropathy-related scales for the optimized residue disorder parameters in the standard simple sequence predictor (sw35_8), the previously published scales do not perform nearly as well as the optimized predictor (Fig. 3.2 9, Fig. 5.1-3). Although, overall, the hydrophobicity/hydropathy scales appear to perform better than the coil propensity scales in making predictions, one exception is in test set 3 (see section 2.4.1), on which the hydrophobicity scales perform significantly worse than the coil propensity scales (Fig. 5.1-3). This seems to make some sense, given stretches of sequence that imperfectly repeat the sequence, 'PSTPSYS'. As proline and tyrosine both have higher scores relative to other residues in general in the coil propensity scales, in comparison with the hydrophobicity or optimized disorder scales, this is not unexpected. It is also of note that the overall performance of the Kyte-Doolittle values is quite similar to that of the Hopp-Woods values and is markedly better that of the coil propensity values (Fig. 3.2-9, Fig. 5.1-3), which might not be guessed alone from respective R^2 association values. Perhaps this has to do in part with the fact that tryptophan and tyrosine, which are the main outliers, are not the most common residue types. It could be taken as suggesting some possibility of direct association between hydrophilicity and disorder although there does not seem to be such an association (Fig. 3.3-1).

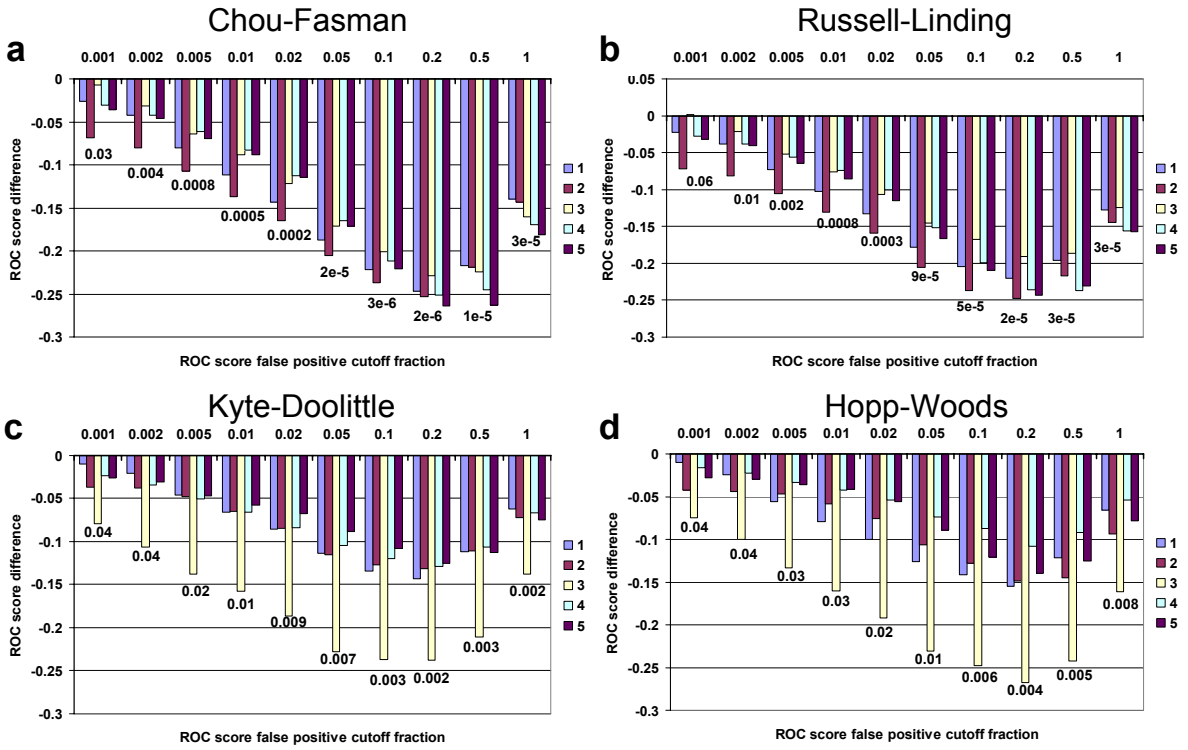

Figure 5.1-3. Differences in performance for simple window predictor using various scales for residue disorder parameters, vs. using optimized disorder values (sw35_8). Number below each set of bars (at each ROC score false positive cutoff fraction) are paired *t*-test *p*-values. a) Chou-Fasman – sw35_8. b) Russell-Linding – sw35_8. d) Kyte-Doolittle – sw35_8. d) Hopp-Woods – sw35_8.

5.2 ASSOCIATIONS WITH HYDROPHOBICITY SCALES

5.2.1 Nozaki-Tanford scale

Experimentally-derived values in the Hopp-Woods scale were those of Nozaki and Tanford (1971). The Nozaki-Tanford values were based on individual amino acid solubilities in water, ethanol, dioxane, or combinations of water and one of the organic solvents. The values shown here are calculated side chain free energies of transfer from water to 100% organic solvent—in some cases, water to ethanol only, and in some cases an average for water to ethanol and water to dioxane, which, according to Nozaki and Tanford "are essentially identical" for hydrophobic side chains. Calculations were based on measurements performed at different water-ethanol or water-dioxane concentrations, and values for solubility in 100% solvent were usually determined through extrapolation. Calculations also included attempts to correct for activity of amino acids at saturation.

Interestingly, Nozaki and Tanford exclude asparagine and glutamine from their scale, observing that: "Asparagine consistently shows more negative ΔFt values than those of glutamine, while one expects the opposite trend, since glutamine has an additional CH_2 group which in our solvent systems should show a negative contribution to ΔFt." (emphasis added). The same is true of the optimized disorder values—asparagine somewhat unexpectedly (given the hypothesis that disorder is mostly a function of hydrophobic residues for non-charged, non-cysteine, non-proline residues) has a lower optimized disorder value than glutamine.

When only the experimentally derived hydrophobicity values from Nozaki and Tanford are compared with their respective optimized disorder values, the correlation is

excellent, with an R^2 value of 0.977 (see fig. 3.2-8c). Regarding reliability of the data, Nozaki and Tanford write, "we believe the reliability of the extrapolated data to be about ±50 cal per mole." They indicate that histidine and leucine are exceptions, and estimate uncertainty for these at ±100 cal/mol. Histidine also represents a special case, because it is the only charged residue type for which they give a value.

The tight correlation of the optimized disorder values with Nozaki and Tanford's calculated values suggests some superiority in Nozaki and Tanford's methods for experimental methods for measuring hydrophobicity and/or methods for calculating hydrophobicity from experimental results over those of others (Radzicka and Wolfenden 1988; Wimley et al. 1996; Wimley and White 1996). Interestingly plots of the Wimley-White scales and 'octanol/water' values (Guy 1985) against optimized disorder values (Fig's 5.2-1 and 3.2-8b, respectively) members of the Nozaki-Tanford subset of residues appear to each agree with the optimized disorder values in different ways. This idea, when added to the strong correlation between the Nozaki-Tanford and optimized values, suggests that the Nozaki-Tanford values may be the most accurate of these different experimental hydrophobicity scales, although not all evidence points toward the superiority of the Nozaki-Tanford scale over the Radzicka-Wolfenden scale (not discussed further here).

5.2.2 Radzicka-Wolfenden/Guy 'Octanol' to water scale

The 'octanol'/water scale displayed by Radzicka and Wolfenden (1988) was an inverted version (that excluded proline) of the Guy (1985) 'octanol'/water scale, which was actually derived from partitioning experiments involving octanol, ethanol, and methanol. Energies from methanol and ethanol were 'normalized' to approximately fit with those of

octanol. Wimley et al. (1996) cite Franks et al. (Franks et al. 1993) as providing X-ray diffraction-based evidence that octanol forms clusters that surround cores where hydroxyl groups and water interact—thus, wet octanol allows both hydrophobic and 'hydrophilic' interactions. The 'octanol' to water transfer energies shows a strong association with disorder, with an R^2 of 0.941, excluding ionic residues (and proline).

5.2.3 Wimley-White scales

The Wimley-White scales use the AcWL-X-LL construct, which complicates matters and appears to yield experimental biases. But taking this into consideration, they lend support to the disorder/hydrophobicity association. Considering the Wimley-White scales in combination can suggest some possible sources of deviation. The slope of the fit for the membrane to water experiment is approximately half that of the octanol to water experiment (Fig. 5.2-1c). The slope presumably relates to the efficiency of hydrophobic protection in the amphipathic phase. A favorable self-interaction in water would be expected to be reflected by negative deviations of approximately the same absolute magnitude in both experiments. Consistent differences in behavior in the organic phase might be expected to have approximately half the effect in the membrane experiment as in the octanol experiment, but this type of effect may be more erratic, due to the complexity in and difference between the organic phases. With a bias in a side chain type's optimized disorder value, rather than its experimental value, the deviation would have approximately half the magnitude in the membrane partitioning plot as in the octanol partitioning plot. Thus, considering deviations in both plots together may support certain plausible explanations for those deviations.

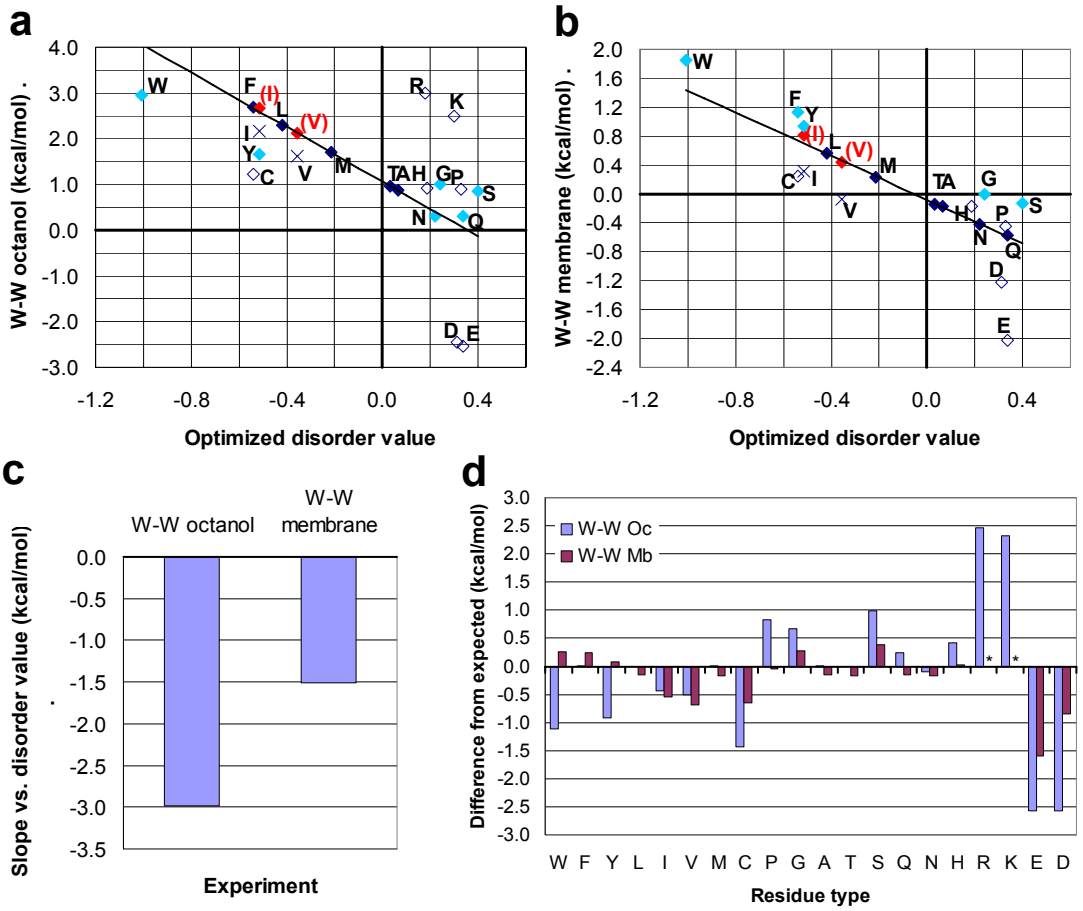

Figure 5.2-1. Correlations of Wimley-White scales with optimized disorder values. Open diamonds mark the residues, H, R, K, D, E, C, and P. Fits are drawn to residues marked by dark blue diamonds. Red diamonds are included for adjusted values for isoleucine and valine, and x's mark unadjusted values in these cases. a) Wimley-White AcWL-X-LL octanol to water transfer energy (Wimley et al. 1996), pH 9, vs. optimized disorder value. b) Wimley-White AcWL-X-LL membrane to water transfer energy (Wimley and White 1996), pH 8, vs. optimized disorder value. c) Slopes of fits. d) Residuals to fits. (*) Asterisks mark missing residuals (due to missing values).

In both Wimley-White scales, the β-branched side chains (isoleucine and valine) have transfer energies that deviate downward from trends (x's in Fig. 5.2-1a, b). The four deviations have roughly similar absolute values (Fig. 5.2-1, c, d), consistent to some degree with a self-interaction that improves solubility in water, as described above. Simple modeling suggests that a β-branched side chain can simultaneously interact with the two adjacent leucines, one leucine with each γ carbon of the β-branched chain, a state where there may be significantly more hydrophobic protection in the aqueous phase than for usual side chain types. With rough corrections for this, several points fall close to straight line trends with relation to optimized disorder values in each scale (Fig. 5.2-1a, b).

The other (non-C, P, ionic) residues that primarily appear to deviate from disorder value in both are W, Y, G, and S. The deviations of tryptophan and tyrosine are not consistent between the two scales, while the deviations of glycine and serine are consistent. Possible explanations are here suggested for these observations. The negative deviations of the polar aromatic residues in the Wimley-White octanol to water scale, with respect to the optimized disorder values suggests that the hydrophilic effect is operating specifically in wet octanol. In wet octanol interfaces (with hydroxyl groups, acyl chains, and water in proximity), the molecules with protruding tyrosine or tryptophan in the middle position may less able to locate in positions where all hydrophilic groups on the backbone and certain side chains of the construct are simultaneously satisfied by polar interactions while still protecting hydrophobic groups well. Glycine and serine may create gaps or regions in the vicinity of the variable side chains that are more conducive to water, also affecting behavior in the organic phase.

Cysteine deviates approximately half as much, in the negative direction, for the membrane experiment as for the octanol experiment. This is consistent with an expected bias in the disorder value for cysteine. An estimated disorder value of –0.08 for non-disulfide bond forming cysteine may be calculated from the Wimley-White experimental values (see section 5.4), although the possibility of special behavior of the construct is likely and makes this value unreliable.

The Wimley-White experiments appear to support the existence of a tight association between hydrophobicity and disorder when special consideration is given to the AcWL-X-LL construct used in these experiments.

5.2.4 Considering other scales

Strengths of associations have been quantified using R^2 values, for which ionic residues, cysteine, and proline are generally reasonably excluded, reasons for which are discussed below. The R^2 value for a qualitative association with disorder may be estimated by assigning each of the more hydrophobic/order-associated residues a value of 1 and each of the less hydrophobic/order-associated residues a value of 0 (like Venanzi (1984) did in associating hydrophobicity with bitterness) and calculating the R^2 value for these values against the optimized disorder values. This method yields an R^2 value of 0.783, or, if placing tryptophan in its own 'super-ordering' category, 0.915 (see Fig. 5.2-2). On the other hand, a perfect quantitative, linear relationship would, of course, have an R^2 value of 1. The Nozaki-Tanford relationship (reflected by R^2 of 0.982) is close to a perfect linear relationship. No good standard statistical test may be applied to distinguish between the strengths of different

associations (Press 1999). But relatively strong associations with hydrophobicity are abundant, and other strong associations are generally lacking.

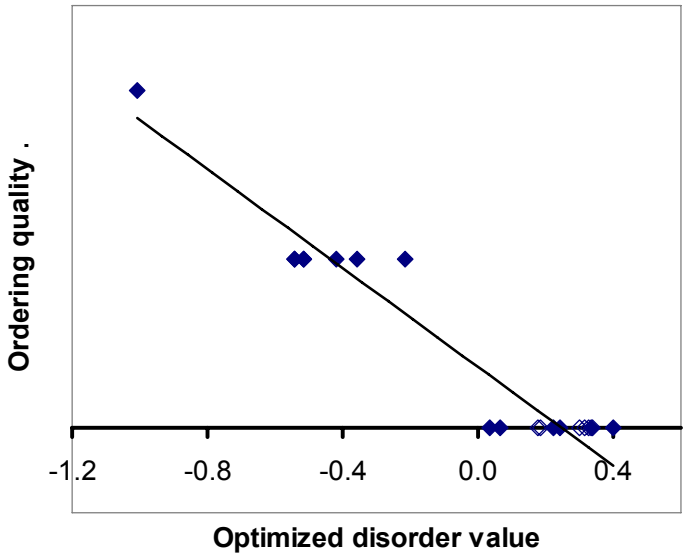

Figure 5.2-2. Ordering quality (discrete values of 0, 1, 2) vs. optimized disorder values. (The residue with the assigned value of two is tryptophan). Compare with Venanzi (1984) bitterness scale.

Williams et al. (2001) demonstrate qualitative relationships between disorder propensities and contact scales, hydrophobicity, flexibility index, and beta-strand propensity. Their best-associated characteristic, 14 Å contact number (Nishikawa and Ooi 1986) (they do not provide R^2 values; $R^2 = 0.830$ with disorder values presented here), is related through its association with hydrophobicity, but also has an association with hydrophilicity (see Fig's 3.3-1, 5.2-3a). (Without going into a full discussion, among other things, the hydrophilicity relationship suggests association between contact number and surface exposure, as would reasonably be expected.) With a clear difference between contact number and disorder, evidenced by difference in deconvolution into hydrophobicity and hydrophilicity, contact

number cannot be used to explain disorder. Williams et al. certainly contributed, but the distinctive relationship between disorder and hydrophobicity was not demonstrated in their work.

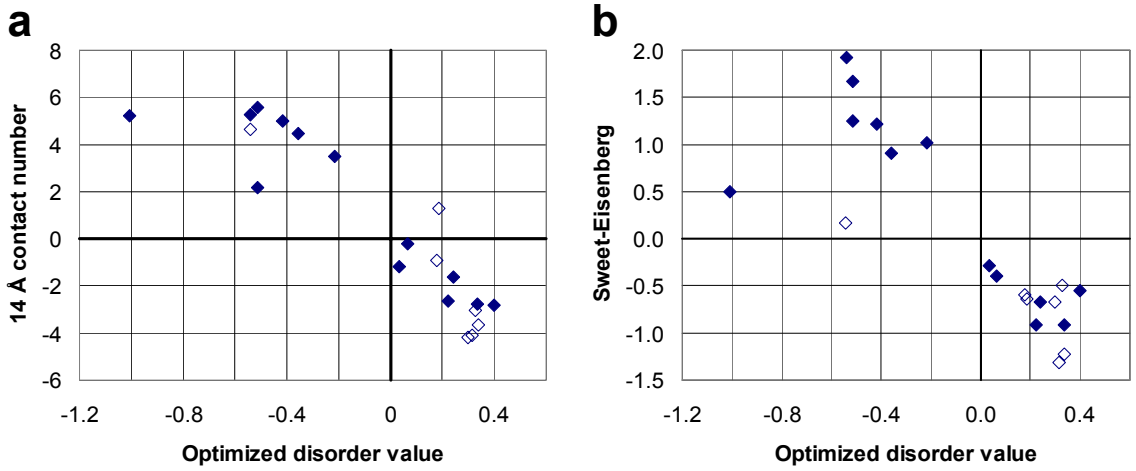

Figure 5.2-3. Certain top scales found by Williams et al. (2001) a) 14 Å contact number (Nishikawa and Ooi 1986). b) Sweet and Eisenberg (1983) 'hydrophobicities'.

The scale that is second-best related to disorder, according to Williams et al. (2001), is the Sweet and Eisenberg (1983) 'hydrophobicity' scale, which was derived using residue substitution data. Dosztányi et al. (2005) found the same scale to be associated with a primary component of a residue-residue interaction matrix, which is used in IUPred, a predictor of 'intrinsically unstructured' protein described in the same paper.

With imperfections in scales, the nature of the hydrophobicity/disorder relationship is not revealed by simple automated searches of scales, either by testing in their ability to make predictions (Williams et al. 2001) (note the Hopp-Woods scale in Fig's 3.2-9 and 5.1-3) or in statistically calculating strengths of associations for all residues. Searches were performed of AAIndex1 (Kawashima et al. 1999), a residue scale database (the version to which local

access was available includes entries with dates no later than 2002), for scales yielding the highest associations of various scales with optimized disorder values as measured by R^2 values—one excluding C, P, and ionic residues in calculation of R^2 values and one including all residues in calculating R^2, the results of which are shown in Tables B-1 and B-2, respectively. When searching by R^2 calculated from all residues, 14 Å contact number (Nishikawa and Ooi 1986), found by Williams et al. to be the characteristic best predictive of disorder, is the second best-associated characteristic, with an R^2 value of 0.840. When excluding the ionic residues, cysteine, and proline, however, the number of scales associated with R^2 better than 0.7 is larger, and significantly stronger associations are observed. All of the thirteen scales with R^2 values greater than 0.9 are apparently hydrophobicity or disorder-related. Four of the top five scales, as measured by R^2, are related to the work of Nozaki and Tanford (1971) and the other is the Guy (1985) 'octanol' to water scale, which also includes influence from ethanol/water partitioning.

5.2.5 Other contributions of methods described here to finding association

Besides the exclusion of C, P, and ionic residues, other aspects of methods used in this work appear to have contributed to discovery and/or allowed improved demonstration of the strength of the disorder/hydrophobicity association, including 1) the size and variety of the data set; 2) the use of predictor optimization rather than simple statistics; and 3) the exclusion in analyses/performance measures of certain cases that would have introduced bias.

Compared with the standard secondary structure categories (helix, strand, coil), disordered residues make up a relatively small fraction of all residues (Table 2.4-1 shows

frequencies of counted residues, ‘missing’ and ‘non-missing’) in crystallographic structures. Even with the large amount of data that were used, there is still some variance in the parameters and other evidence of differences from subset to subset of data (see, for examples, Fig’s 3.2-2, 3.2-3, 4.2-1). Had less data been used, it might be expected that random deviations would be greater. Rather than selecting a single representative from each related group of structures, often several structures were used as representatives from a group, and some evidence (not shown) was observed that this reduced noise in the parameters.

The optimized disorder values for the standard residue types are clearly associated with disorder vs. order log odds ratios, and they thus offer evidence of what might be expected—that log odds ratios (or similar statistics) may function for a linear disorder predictor as near-optimal disorder propensities. However, although they of course showed good correlation with hydrophobicity, log odds ratios (essentially calculated using the same data, balanced in the same fashion) do not correlate as well with the Nozaki-Tanford hydrophobicities as do the optimized disorder values. The optimization process appears to have excluded to some degree some disordering property in the aromatic residues, thus allowing more clear demonstration of the hydrophobicity relationship (see the section 5.3.4 on aromatic residues for more discussion).

Although it cannot be claimed that every source of bias has been removed from the data, significant sources of bias have been removed. When the effects of ‘special cases’ are substantially removed from training, predictions on standard cases may be considered more valid, and optimized values may be better related to characteristics of standard cases. Care was taken to exclude effects of N-terminal methionine and polyhistidine tags. Williams et al.

(2001) show an apparently high value for methionine. Linding et al. (2003a) also provide statistical disorder values where histidine and especially methionine, as they note, appear higher relative to other residue types. Among optimized disorder values shown herein, methionine clearly falls in line with the disorder/hydrophobicity trend, and histidine is close to it (Fig. 3.2-8c).

5.3 RESIDUE-SPECIFIC ISSUES

On the order of simple residue composition, not many factors other than hydrophobicity appear to play a large role in disorder. Nevertheless, with hydrophobicity established as a baseline property, the deviation (or non-deviation) of certain residues from the trend may be informative. Cysteine and proline appear to have disorder propensities that appear to deviate from their hydrophobicities. On the other hand, the discrepancy between disorder propensities and 'octanol/water' (Guy 1985) transfer energies for charged residues may be primarily due to the deviation of experimental values from their true hydrophobicities. These and other residue types are discussed in more detail below.

5.3.1 Cysteine

Cysteine has an obvious special property—that of disulfide bond formation. The disorder propensity for cysteine is lower than that of methionine and yet methionine (which is the other sulfur-containing amino acid) has two more carbon atoms than cysteine, which would suggest that it is more hydrophobic than cysteine (Levitt 1976) and should have a lower disorder propensity than cysteine. Methionine falls in well with the

disorder/hydrophobicity trend. This would suggest that cysteine deviates from the disorder/hydrophobicity trend—having lower disorder propensity than would be expected from hydrophobicity. This is consistent with that idea that disulfide bonds constrain a protein, thus promoting order—an effect opposite to that of the decreased contraint that occurs at protein termini—which are recognized to be more likely to be disordered (Ward et al. 2004) (see Table 2.4-1). The Meek (1980) scale also gives some evidence that cysteine does not follow the hydrophobicity/disorder relationship, although this is not corroborated by Meek and Rossetti (1981).

5.3.2 Glycine

Glycine has decreased steric hindrance of backbone torsion. Though glycine may have a strong influence in producing disorder (Esnouf et al. 2006), glycine follows disorder/hydrophobicity trends (see Fig. 3.2 8a, d), likely reflecting a balance between ordering and disordering properties.

5.3.3 Proline

Proline is not given a value in the Radzicka-Wolfenden ‘octanol/water’ scale (of course, given their use of side chain analogs), nor in the Nozaki-Tanford scale. Proline might significantly alter the behavior of the AcWL-X-LL construct, so the Wimley-White values are not too informative. Some evidence suggesting that proline deviates from the disorder/hydrophobicity relationship may be found in the Meek (1980; fig. 5.3-1) and Meek and Rosetti (1981; not shown) scales, as well as the original Guy (1985) ‘octanol’/water scale

(not shown). Proline may promote disorder not just because it simply disrupts helices and strands (recall that coil propensity is not well-correlated with disorder propensity) but also because it forces an extended backbone conformation. A similar point is that, given its requirement of an extended conformation, proline might be better substituted into regions that tend to be disordered in a given family of proteins. Variation in prolyl isomerization could reasonably play a role in crystallographic disorder.

5.3.4 Charged residues

Some of the biggest scatter in plots of various experimental scales vs. disorder values comes from the ionic residues (Fig's 5.2-1a,b, 3.2-8d). Radzicka-Wolfenden ionic residue values had been adjusted on the assumption that only the uncharged molecules partitioned into the organic phase (Radzicka and Wolfenden 1988). For the Wimley-White octanol to water experimental values given, such adjustments were not made. Measurements were performed under basic conditions, and arginine and lysine reasonably deviate toward greater hydrophobicity and aspartate and glutamate deviate in the opposite direction. With the Meek (1980) HPLC retention coefficients plotted against disorder values in Fig. 5.3-1, the strong ionic residue experimental values do not deviate from optimized values very strongly, and this may be because the hydrophobic phase is related to a surface so that residues interacting with the hydrophobic phase are still interacting with the aqueous phase, with which ionic residues more favorably interact.

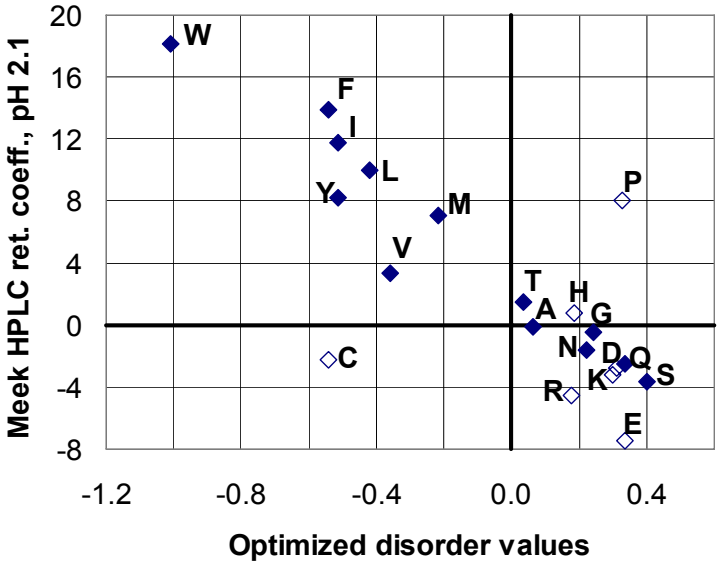

Figure 5.3-1. Retention coefficients, calculated from HPLC retention times of various 'peptides' (Meek 1980). Conditions included a pH of 2.1, 0.1 M NaClO4, and an acetonitrile gradient. It is noted that cysteine does not similarly stand out in a similar set of values (Meek and Rossetti 1981).

Uversky et al. (2000) show that mean net charge can be used in combination with Kyte-Doolittle hydropathies in predicting global disorder. Disorder predictors developed in this work do not utilize net charge. It might be expected that if intermediate-range electrostatic interactions had a substantial net effect in disorder prediction, there would be marked shift in ionic residues' disorder values in short window predictor (sw9_1) optimized disorder values, in contrast to standard optimized values. Considering all charged residues, however, a shift toward disorder, if any, is quite small (see Fig. 5.3-2), although there appears to be less consistency between standard disorder values and log odds ratios (Fig. 3.2-7a). Among the standard disorder values, glutamate and aspartate are similar to asparagine and glutamine (see discussion below on the amide residues). Overall, the evidence suggests that, although charge may play some role in crystallographic disorder, the average effect of

charge is relatively small, and hydrophobicity is the main determinant of the average disorder contributions of charged residues. The charged residue disorder values may reflect their hydrophobicities, but there is not firm evidence for this.

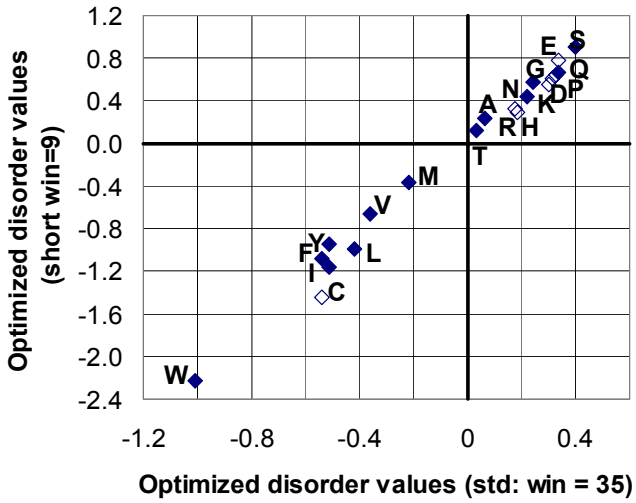

Figure 5.3-2. Short window predictor (sw9_1) residue disorder values vs. standard (sw35_8) optimized disorder values (C, P, and ionic residues are open diamonds).

5.3.5 Amide residues (asparagine and glutamine)

Glutamine has one more CH2 than asparagine and might be expected to be the more hydrophobic of the two, but asparagine's disorder value is markedly less than glutamine's. This might be partly explained by asparagine's ability to form a six membered ring-like shape, with an interaction between the side chain amide NH_2 and its backbone carboxyl/carboxylic acid group. This may have an ordering effect, if the hydrophilic tendency of the backbone carboxyl oxygen to interact with water is reduced. Nozaki and Tanford (1971) mention the troublesome result that the hydrophobicity calculated for asparagine was consistently more than that of glutamine, and this seems to have been why

they excluded these residues from their scale. For other experimental scales, asparagine also unexpectedly has a hydrophobicity less than or equal to that of glutamine (see Fig's 3.2-8d, 5.2-1a, b)—even the Radzicka-Wolfenden scale, derived from experiments with side chain analogs, which have no 'backbone' with which to interact.. Thus, asparagine's self-interaction perhaps only partly explains the unusual relationship between glutamine's and asparagine's disorder values. Glutamine has almost exactly the same disorder value as glutamate (Fig. 3.2-4a). Aspartate has a slightly lower value, and the difference between asparagine and aspartate might roughly represent the degree of asparagine's special behavior. The fact that glutamine and glutamate have almost exactly the same disorder values supports the assertion that hydrophobicity, not hydrophilicity, is tightly connected to disorder, given that glutamate's carboxylic acid group should be more hydrophilic, on average, than the glutamine's amide group.

5.3.6 Aromatic residues (phenylalanine, tyrosine, and tryptophan)

The aromatic residues may demonstrate how, given a tight hydrophobicity/disorder association, deviations from the trend might be informative. Tyrosine, phenylalanine, and tryptophan may deviate toward disorder by a small amount (Fig. 5.3-3). Optimized disorder values for a predictor with a window length of nine, rather than thirty-five (Fig. 5.3-3b) seem to show more apparent deviation toward disorder for phenylalanine and tyrosine, although this is at least in part due to a shift in the relative value for leucine. This shift in aromatic residue disorder values (if real) suggests that they have some disordering quality, apart from their hydrophobicity-related ordering propensity. One might explain this by saying that

tyrosine and phenylalanine are ‘disorder-seeding’ residues. When a $ROC_{1.0}$ score is used to optimize the short window predictor, then there is a right shift in the value for tryptophan also (Fig. 5.3-3c), and the correlation plot against Nozaki-Tanford hydrophobicities looks quite similar to that for log odds ratios. Tryptophan thus perhaps deviates toward disorder when considering regions with low overall disorder propensity (and toward order in regions with overall disorder propensity; not further discussed here).

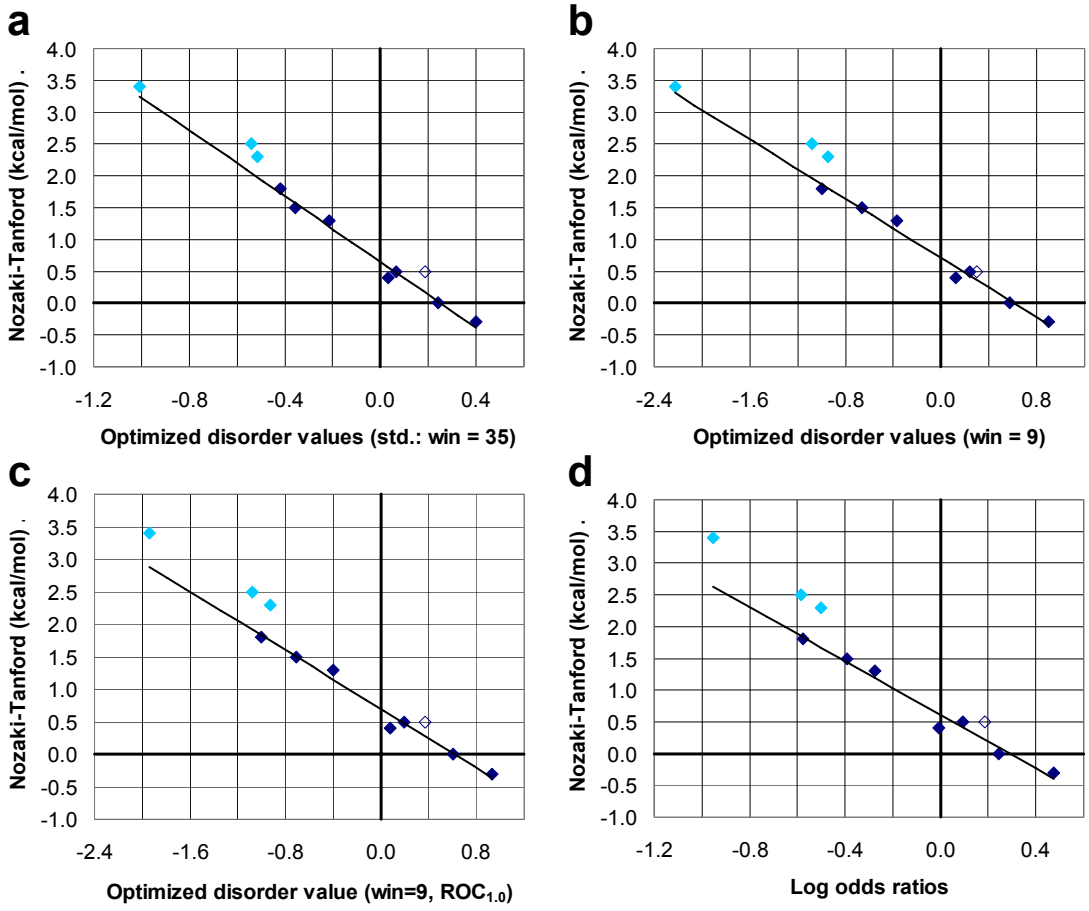

Figure 5.3-3. Correlations with Nozaki-Tanford scale showing possible aromatic residue deviation. a) Nozaki and Tanford (1971) amino acid hydrophobicities vs. optimized disorder values, with fit. b) Nozaki-Tanford hydrophobicities vs. short window optimized disorder values, with fit (using standard $ROC_{0.5}$ score optimization). c) Nozaki-Tanford hydrophobicities vs. short window disorder values optimized using $ROC_{1.0}$ scores (instead of $ROC_{0.5}$ scores), with fit. d) Nozaki-Tanford hydrophobicities vs. average log odds ratio, with fit. (All fits exclude W, F, Y – light blue diamonds). Note that the apparent shift in the light blue residues, however, may be somewhat deceptive, being partly due to a shift in the left-most dark blue residue (Leu).

5.3.7 Methionine

A special disorder value for N-terminal methionine was optimized along with the tail adjustment values for the simple sequence predictor (see Fig. 3.2-3a). The N-terminal methionine disorder propensity is much higher than that of normal methionine and even higher than that for serine (see Table A-1). This is probably primarily explained by two factors: 1) many crystallographers may neglect to include the full chain sequence in a PDB file when termini are disordered; and 2) N-terminal methionine is sometimes cleaved during posttranslational processing (Huang et al. 1987; Boissel et al. 1988). Based on conversation with crystallographers, unobserved terminal regions may often be entirely excluded from their published structure files, notwithstanding PDB instructions to publish the full chain sequence in SEQRES records even if terminal disorder is present (see PDB Format Description Version 2.2, Section 3 – Primary Structure Section, subsection SEQRES, currently available at the PDB website—go to http://www.pdb.org). In structures where no proteolytic cleavage was performed, when N-terminal methionine is present its presence is an indication that the crystallographer has deposited the entire sequence, rather than excluding disordered termini. For this reason, termini would expectedly appear to be disordered more frequently when an N-terminal methionine is present.

In some disorder propensity scales, methionine has a substantially higher relative disorder propensity than in our scale (Williams et al. 2001; Linding et al. 2003a; Weathers et al. 2004). Because of our exclusion of terminal residues during optimization, however, methionine falls in with disorder/hydrophobicity trends (see Fig. 3.2-8c,d), giving evidence that it has no significant special disorder-related properties that are unrelated to its

occurrence at amino termini, which of course tend to be disordered, and supporting the conclusion that disorder is quantitatively related to hydrophobicity. Selenomethionine, which replaces sulfur with the larger selenium atom, also appears to follow the disorder/hydrophobicity relationship, as it is modestly more order-promoting than methionine. Amino-termini are sometimes post- or cotranslationally processed, sometimes with specific removal of methionine by methionine aminopeptidases. Statistics show N-terminal methionine to be missing markedly more frequently when certain ordered residues are in the second position, including S, A, G, P, T, and V (also C with small sample size; see Table 5.3-1) in good agreement with other results (Huang et al. 1987; Boissel et al. 1988) on methionine aminopeptidase processing. Thus, in some cases, the absence of methionine from the coordinates is due to its being altogether absent from polypeptide chains rather than just being disordered.

Table 5.3-1. Statistics on whether or not N-terminal methionine is missing, in cases where second (penultimate) residue is present in the structure. A high relative frequency of missing methionines for a particular penultimate residue type suggests that methionine is often cleaved by an aminopeptidase when that residue type is present in the penultimate position, and thus the N-terminal methionine is entirely missing from the chain rather than simply being disordered.

Penultimate Residue type	Subset log odds ratios Data subset					Whole dataset Relative frequency			log odds ratio
	1	2	3	4	5	Missing	Present	All	
W	N/A	N/A	N/A	N/A	-Inf	0.0000	0.0004	0.0004	-Inf
F	-Inf	-Inf	-Inf	-Inf	-Inf	0.0000	0.0199	0.0199	-Inf
Y	-Inf	-Inf	0.91	-2.50	-Inf	0.0012	0.0149	0.0161	-1.95
L	-1.38	-4.23	-Inf	-2.99	-4.22	0.0015	0.0654	0.0668	-3.28
I	-0.86	-0.17	-1.31	-0.92	-2.24	0.0101	0.0484	0.0584	-0.99
V	1.82	0.33	0.18	1.17	0.32	0.0257	0.0281	0.0538	0.57
M	-Inf	-0.07	Inf	-1.02	-Inf	0.0034	0.0062	0.0096	0.02
C	Inf	N/A	Inf	N/A	N/A	0.0002	0.0000	0.0002	Inf
P	0.86	1.30	1.98	3.83	1.29	0.0369	0.0171	0.0540	1.49
G	1.93	1.89	Inf	3.81	1.37	0.0222	0.0053	0.0275	2.12
A	1.77	2.90	1.97	2.70	2.51	0.0667	0.0141	0.0807	2.37
T	1.56	-0.68	1.46	0.73	1.62	0.0265	0.0207	0.0472	0.92
S	1.43	1.97	2.00	2.00	2.98	0.0698	0.0196	0.0894	2.09
Q	-0.90	-Inf	-1.05	-0.20	0.30	0.0081	0.0345	0.0426	-0.85
N	-0.06	-3.11	0.19	-0.70	0.15	0.0231	0.0599	0.0830	-0.35
H	-2.75	-Inf	-Inf	-Inf	-Inf	0.0000	0.0107	0.0107	-5.29
R	-2.29	-0.77	-Inf	-1.10	-3.14	0.0055	0.0548	0.0603	-1.75
K	-0.97	-1.64	-0.69	-1.04	-1.97	0.0201	0.1159	0.1360	-1.26
E	-0.97	-1.51	-1.46	-1.93	-0.08	0.0129	0.0753	0.0883	-1.22
D	-1.28	-0.89	0.58	-0.33	-0.71	0.0135	0.0409	0.0545	-0.50
Nonstd	N/A	N/A	N/A	Inf	N/A	0.0006	0.0000	0.0006	Inf
(Sum)						0.3480	0.6520	1.0000	

5.4 INTERPRETING THE LINEAR DISORDER/HYDROPHOBICITY RELATIONSHIP

The linear relationship between optimized disorder values and experimental hydrophobicity scales may be at least qualitatively informative, but it should be approached with caution. Using the three sets of linearly related values: log odds ratios, optimized disorder values, and the Nozaki-Tanford hydrophobicities, and considering the score profiles of disordered and ordered residues (see Fig. 3.2-5), values for may be obtained for the following relationship (see Parameter-Energy Workbook):

$$-RT\ln\kappa_{diso,i} = \frac{-RT\ln K_{sc,i}}{d} + b \quad [5.4.1]$$

$K_{sc,i}$ is an equilibrium constant for some side chain type, i, calculated from its free energy of transfer between some organic environment and water. $\kappa_{diso,i}$ is a disorder/order partitioning constant calculated using disorder statistics, adjusted using the scoring behavior of the predictor. d is a dampening factor whose value depends upon the hydrophobicity experiments against which disorder is being compared. b is essentially the average per residue energy of the backbone contribution to local disorder. Given possible sources of substantial error, parameter-derived energies should not be taken to be highly accurate. Thus, there is also little need for debate over what statistical measure (e.g., log odds ratio vs. log probability ratio) is exactly equivalent to an ln K derived from experimental free energy of transfer.

Values for d and b are meaningful. When comparing against the Nozaki-Tanford or Guy octanol to water scales, d is roughly 4 in either case, essentially indicating a 4-fold reduced difference (hydrophobicity-wise) between the disordered and ordered states, on

average, in comparison differences between organic and aqueous states indicated by experiments. This value depends upon the 'temperature' used in the process of converting disorder values to energies. This could reflect a combination of factors. It may primarily reflect less of a difference between average disordered and ordered states than between organic and aqueous environments in model systems—worse protection, on average, of hydrophobic groups in ordered states than in ethanol or dioxane, and less exposure in disordered states than in water. It may reflect that, although on average hydrophobicity is order promoting, it can be disorder-promoting in some circumstances, as mentioned. It presumably reflects the presence of non-hydrophobicity related contributions toward disorder.

The value of *b* depends on determination of where disorder and order have equal 'energy'. This is estimated using the absolute score distributions for ordered and disordered residues. It appears that they should cross over (if a 'bump' in the disorder distribution were smoothed out) not too far from a score of 3.3. Using 3.3 as a score where disordered and ordered regions should occur with equal frequency, the value for *b* is roughly –0.04 kcal/mol (the Nozaki-Tanford values are only precise the 0.1 kcal/mol, and their estimated error range for glycine is ±0.05 kcal/mol—see section 5.2.1), but even with significant deviation in the score at which disordered and ordered conformations should be equally probable, the value for b does not change much relative to the energies of other residue types.

There is a possibility that estimated energies corresponding to tail adjustments may also be calculated, but this has not been done in a manner in which I am confident of the result(s).

CHAPTER SIX
Secondary structure and disorder

6.1 RELATIONSHIPS WITH PSIPRED HELIX, COIL, AND STRAND SCORES

6.1.1 Introduction and Methods

Residue-specific disorder predictors consider only two general states of residues in a protein: disordered and ordered. Traditional secondary structure prediction programs consider three states: ‘helix’, ‘strand’, and ‘coil’.

Jones and Ward (2003) noted improvement in disorder prediction from utilizing secondary structures in the first version of DISOPRED. In reporting the second version of DISOPRED (DISOPRED2; Ward et al. 2004), however, it was shown that secondary structure prediction information helped disorder prediction somewhat when augmenting simple sequence-based prediction, but not when augmenting profile-based prediction.

To investigate relationships between secondary structure prediction and disorder prediction, PSIPRED predictions were successfully performed for all chains included in the general profile dataset (see section 2.4). (PSIPRED prediction results include coil, helix, and strand state scores for each residue, as well as an overall prediction as to which state the residue adopts.) Previous description of the use of data in testing/analysis applies here, including the exclusion of certain residues from analysis (see sections 2.4, 2.5.2). At least eighteen residues were excluded from analysis at each terminus (see section 2.5.2). Disorder scores were calculated using the standard simple sequence predictor (sw35_8).

Interesting relationships were observed between output from PSIPRED, a traditional secondary structure prediction program, and from the simple window predictor. Comparing

PSIPRED results and the simple sequence disorder predictor's results show that the relationships between helix, coil, and strand scores and disorder scores are not monotonic. The relationship between disorder and coil scores may be of particular interest, since some researchers may attempt to use a secondary structure predictor as an ad hoc disorder predictor by looking at coil scores.

6.1.2 PSIPRED study Results

6.1.2.1 Disorder score vs. PSIPRED coil score

Relatively simple, but not monotonic, relationships appear to exist between disorder scores and coil, helix, and strand scores. As expected, disorder scores appear to generally show a positive correlation with coil scores (see fig. 6.1 1). At the high end of coil scores, however, the associated average disorder scores begin to show an inverse correlation. Perhaps this is largely due to high coil scores being associated with easily identifiable features in structures, with more flexible loops being more often associated with scores in the moderately high range.

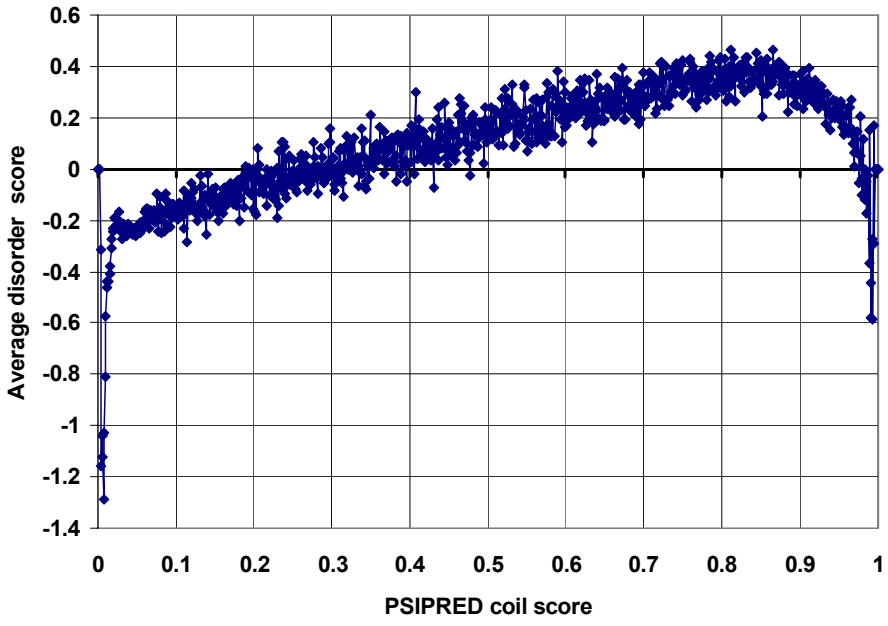

Figure. 6.1-1. Average (standard simple sequence predictor) disorder score vs. PSIPRED coil score.

6.1.2.2 Disorder score vs. PSIPRED helix score

PSIPRED helix scores appear to generally have a mild, inverse relationship with disorder scores (Fig. 6.1-2), as might be expected. Very low helix scores also tend to have lower disorder scores—likely in large part because these tend to correspond with strongly predicted strand (or coil).

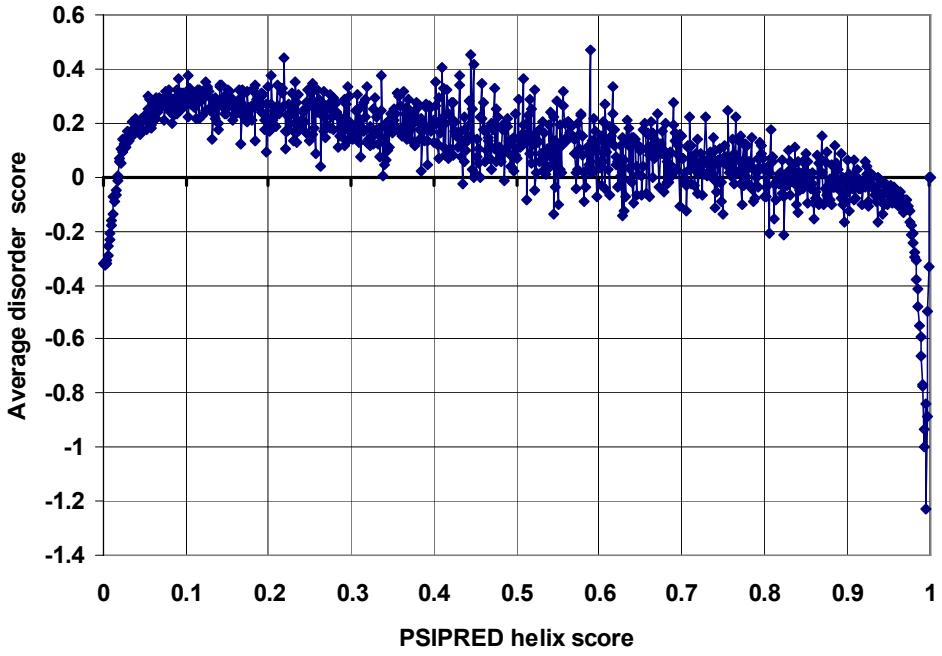

Figure. 6.1-2. Average disorder score vs. PSIPRED helix score.

6.1.2.3 Disorder score vs. PSIPRED strand score

The relationship between PSIPRED strand scores and disorder scores (Fig. 6.1-3) appears to be similar to the relationship between helix scores and disorder scores. One difference is that the curve appears to be more peaked at strand scores close to 0.040. The primary differences appear to be in a generally steeper 'curve', and in generally more negative disorder scores at any given non-extreme secondary structure scores when compared with the helix curve. An association between disorder and β-sheet propensity has previously been noted (Williams et al. 2001).

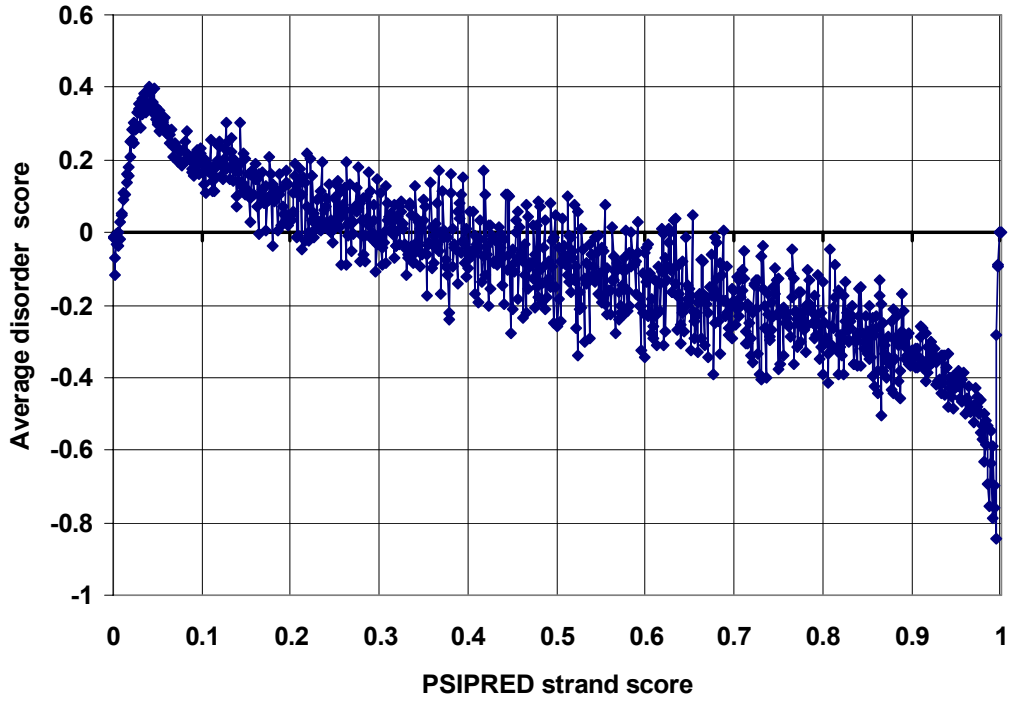

Figure. 6.1-3. Average disorder score vs. PSIPRED strand score.

6.1.2.4 Disorder score vs. sum of PSIPRED coil, helix, and strand scores

The relationship between disorder scores and the sum of the PSIPRED coil, helix, and strand scores (Fig. 6.1-4) appears to be more complex. From a coarse perspective, it appears to have a 'W' shape, and at a finer level, there is a brief dip close to the value of 1, which may reflect certain residues strongly predicted to adopt either helix or strand conformation, to the exclusion of the other. The increase in scatter as scores move farther from 1 in either direction reflects a relative paucity of residues whose sum of scores is very much different from 1. It appears that residues with a very low summed score (in this case, nothing would be very strongly predicted) tend to have higher than average disorder scores, while residues on the higher end, have lower disorder scores.

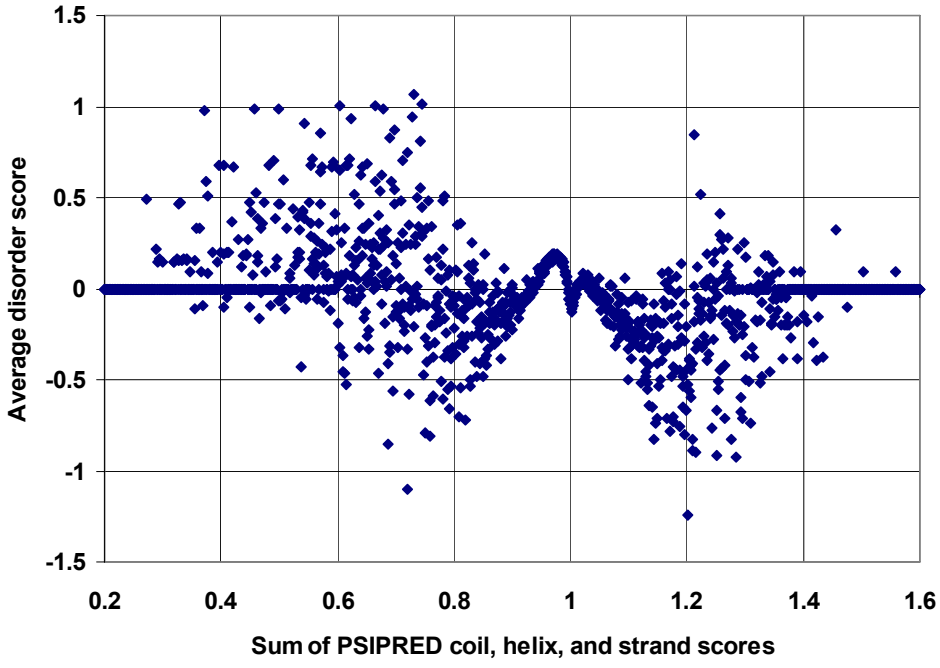

Figure. 6.1-4. Average disorder score vs. sum PSIPRED coil, helix, and strand scores.

6.1.2.5 Disorder in data and various combinations of PSIPRED coil, helix, and strand scores

A different analysis was also done, in which PSIPRED secondary structure predictions were collected into bins for residues that fall within the following moderate range of simple window prediction scores: (0, 1.5). The bins were made according to three variables—PSIPRED helix, strand, and coil scores. Each of these categories was split into five ranges: 0.000 – 0.199, 0.200 – 0.399, 0.400 – 0.599, 0.600 – 0.799, and 0.800 – 1.000. Bins included each combination of ranges for the three secondary characteristic scores (125 in all). For each of the five standard cross-validation profile test sets, the number of residues that fell into each of these bins was tallied for ordered and disordered residues. The ordered and disordered bins were compared, and p values for differences were calculated using a two tailed, pairwise t test. Disordered residues that fall within this score range tend to have higher helix scores and lower strand scores (see Table C-1).

These results may be confounded in relation to the range of scores. Even though the residues were limited to a disorder score range of 0 – 1.5, those residues strongly predicted as strands would likely still have a lower average disorder score than those strongly predicted as helices (see Fig's 6.1-2 and 6.1-3). Narrowing the disorder score range of these residues might reduce this effect but also reduce the statistical power of the sample.

6.1.3 Discussion of PSIPRED/disorder results

No disorder predictor perfectly discriminates between ordered and disordered residues. Of potential interest are disordered residues that do not receive high disorder scores. Combining disorder score and secondary structure prediction may be useful. The results described here suggest a logic that might underlie any contribution that secondary structure

prediction may add to prediction of disorder—for example, a residue with a moderate disorder score, a high helix, and low strand score (and that may be closer to the edge of a helix) may be more likely to be disordered. Whether strands are actually more ordered than helices is another question—distributions of scores for all-alpha and all-beta SCOP classes are quite similar, as discussed below. Nevertheless, it is possible that helices undergo more order-disorder transitions than helices.

6.2 PREDICTION BY SCOP CLASS

The disorder predictors' parameters were developed using the first five classes of protein domains in SCOP—"globular" domains with alpha and/or beta core structure. One might expect that a disorder predictor performs notably well on, say, all-alpha proteins, but poorly on all-beta proteins, or that the expected score distributions for different classes of proteins might be significantly different (as results from the previous section seem to suggest would be the case for all-alpha vs. all-beta proteins). However, there appears to be little noticeable difference between classes for the simple predictor.

Table 6.2-1. Number of families in each SCOP class in SCOP, version 1.67, that were included in the SCOP class data sets and subsets.

SCOP class	Number of families in test set: 1	2	3	4	5	Total families in class
Alpha only	79	79	79	78	79	**394**
Beta only	88	89	89	88	89	**443**
Alpha/Beta	105	104	104	105	105	**523**
Alpha+Beta	101	101	101	100	101	**504**
Multi-domain	9	9	10	10	10	**48**

The first five SCOP classes (alpha and/or beta domains) all have approximately the same overall distribution of disorder scores for non-missing ('ordered') residues (Fig. 6.2-1). However, some differences appear to be present. With respect to disorder score distributions, one class that sets itself apart somewhat is the alpha only class. Its ordered score distribution is noticeably different from the others in its somewhat skewed shape, where it drops off more steeply in the positive direction than in the negative direction (Fig. 6.2-1). Furthermore, the distribution of scores for disordered residues is skewed in the same general way, but more markedly so (Fig's 6.2-2, 6.2-3). In particular, there appears to be a precipitous drop-off between score bin 3.1 and score bin 3.3. If one looks at the shapes of distributions of scores of disordered residues for the five individual alpha-only test sets (Fig. 6.2-4), sets 1, 2, and 5 all show large, steep drops at this location; set 3 shows somewhat of a drop-off there, with a larger one from bin 2.5 to bin 2.9; and set 4, which is odd in its broad, flat shape, seems to have a more subtle drop-off in its average level at this point. But it is still notable that 3 out of 5 sets all have significant drop-offs at the same location, and it suggests, along with the corresponding average histogram shape (Fig's 6.2 2, 6.2-3), that there is a something limiting at a disorder score of around 3.2, beyond which it may be more difficult for alpha-only proteins to maintain stability, or perhaps undergo disorder–order transitions. It is interesting that it appears that the absolute distributions for ordered and disordered residues should cross in the vicinity of score bin 3.3 (see Fig. 3.2-5, section 5.4).

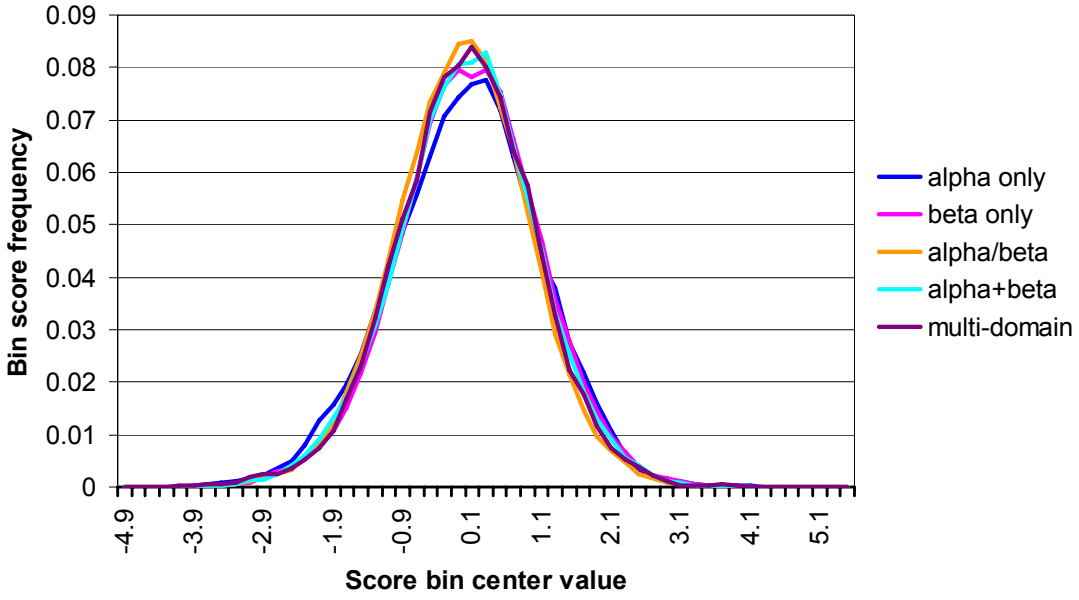

Figure. 6.2-1. Histogram of average frequencies of ordered residues over different scores, for the first five SCOP classes (alpha, beta classes: those used in predictor development). Bin size is 0.2. The x-axis value for a bin is its center value. (e.g., bin 0 – 0.2's value is 0.1).

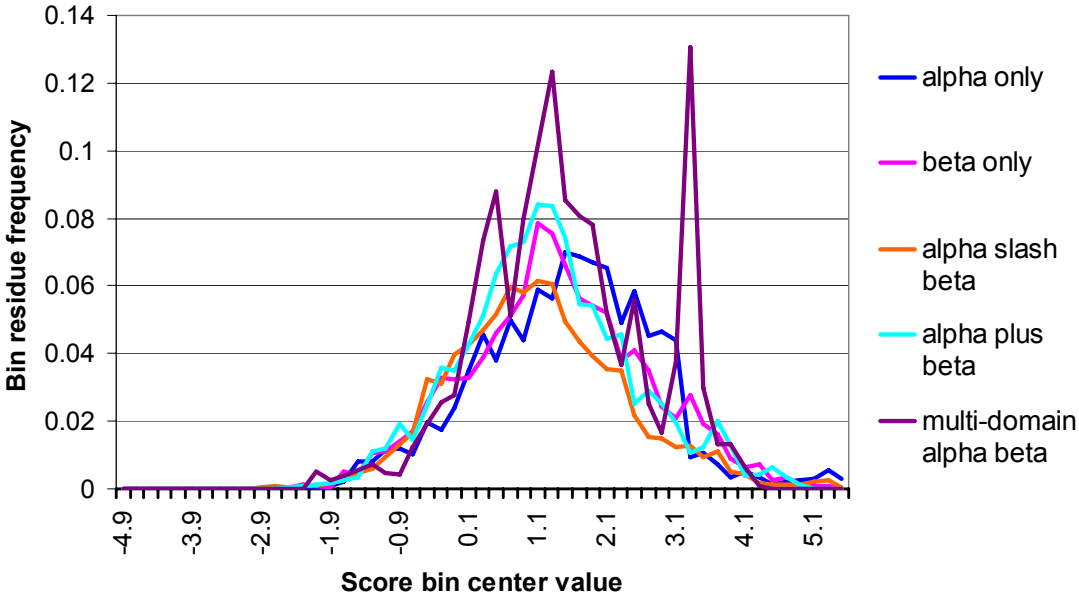

Figure. 6.2-2. Histogram of average frequencies of disordered residues over different scores, for the first five SCOP classes. See explanation of bins in Fig. 6.2-1.

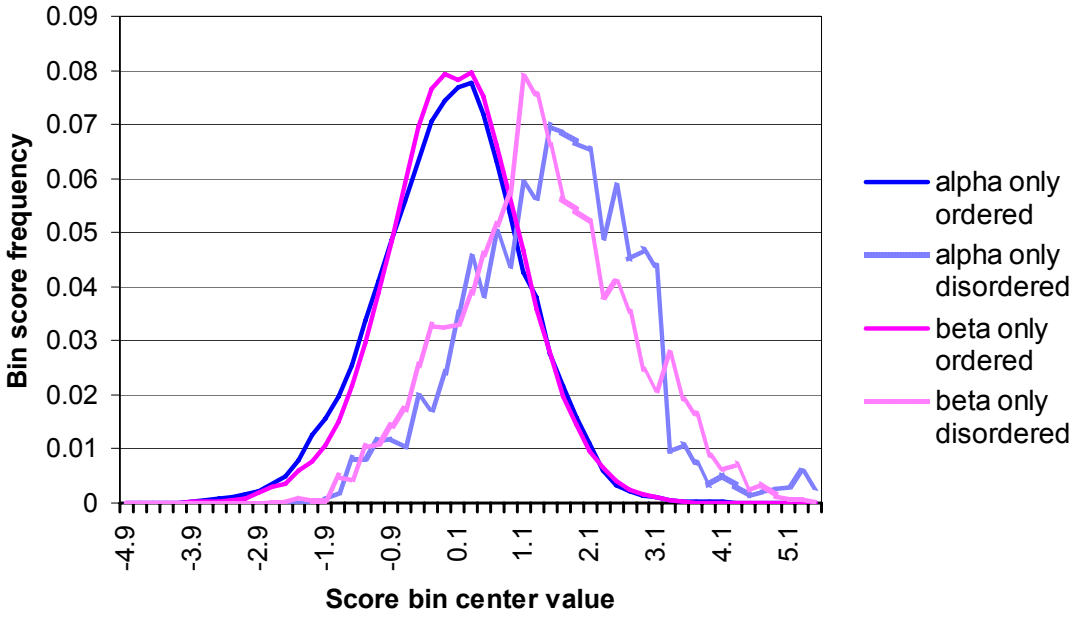

Figure. 6.2-3. Histogram of average frequencies of ordered and disordered residues over different scores, for only the first two SCOP classes.

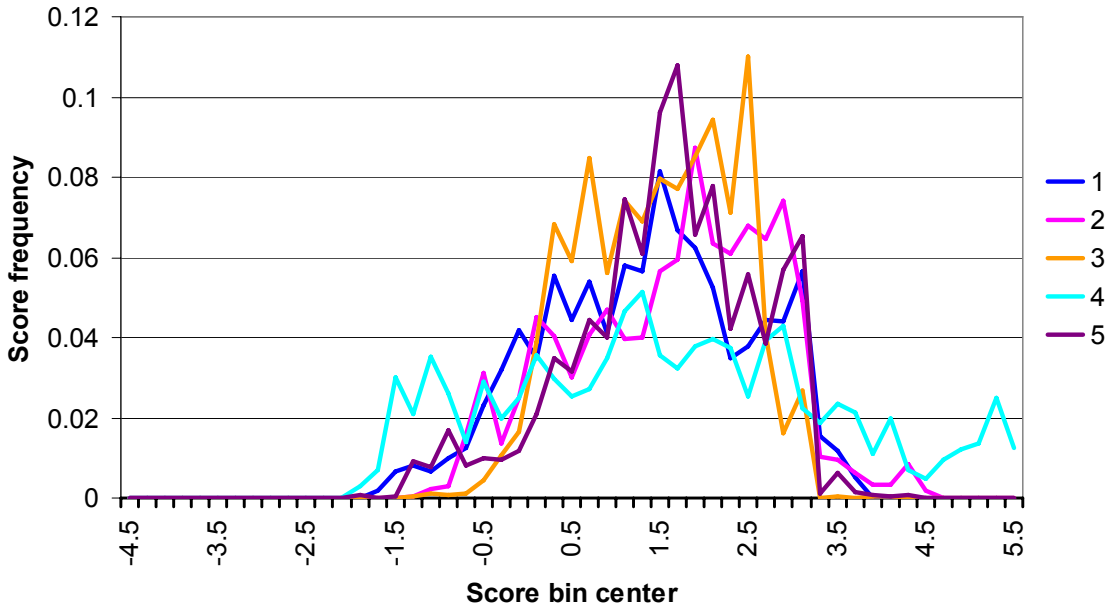

Figure. 6.2-4. Histogram showing frequencies of missing residues for alpha-only proteins. The five individual test sets are shown. Note the precipitous drop, particularly for sets 1, 2, and 5, from score bin 3.1 to score bin 3.3.

CHAPTER SEVEN
Conclusion

Some disorder predictors use parameters that are selected because they seem logical. Other predictors have been data-optimized, but with their complexity, understandably do not have clear explanations for their resulting parameters. We demonstrate a simple, optimized, sequence-based predictor that shows performance similar to a support vector machine, neural network, and profile-based predictor (DISOPRED2). Furthermore, the optimized parameters reveal what is contributing to the prediction. Simple side chain composition may play a stronger role in predicting disorder than in predicting (Rost and Sander 2000) secondary structure. Findings presented here could provide basis for a good 'prior' model that could help to establish whether higher order sequence patterns are significantly related to disorder through establishing whether certain patterns yield a statistically significant increase in predictive power over what would be expected from composition alone. Identifying sequence regions where the predictions provided by our simple sequence-based predictors differ from those of neural network predictors may help to elucidate what additional sequence patterns or factors those predictors are utilizing.

This work demonstrates how data-optimized predictor parameters may be scientifically informative. Although statistics (i.e., log odds ratios) might have been used to demonstrate some of the things demonstrated using optimized parameters, use of the predictor-based approach may be useful in various ways. A predictor, of course, inherently provides means for assessing and utilizing the predictive value of its parameters. Relatively simple predictors provided, overall, predictive performance similar to that of DISOPRED2

(see Fig. 3.2-1). A predictor may provide a simple means of looking at multiple factors in combination. Nonlinear terms may easily be introduced, and compared with linear ones. And even for linear predictors, predictor parameters may in some instances differ substantially from potentially related statistics, as with our profile-based predictor (see Fig. 3.2-7). Optimization may be based on a performance measure that yields more appropriate results than simple statistics. In our case, use of a fractional ROC score reduced the influence of 'low-specificity cases' of disorder (i.e., low-scoring disordered segments) on optimized parameters (see Fig. 5.3-3). Such an approach might be useful if low specificity cases included more artifactual or atypical types of disordered regions.

From a practical standpoint, prediction of disorder may be useful to crystallographers. Detection of disordered regions by NMR and subsequent removal can improve protein crystallizability (personal communication from J. Rizo-Rey). Using predictors such as ours might augment efforts to remove disordered regions that are preventing crystallization (Esnouf et al. 2006). In contrast, selective modification of hydrophilic side chains (Derewenda 2004) may improve crystallization not through eliminating local disorder but through removing a hydrophilic penalty for the formation of crystal contacts.

This work enhances understanding of relationships between disorder and the primary traditional secondary structure classifications (helix, strand, and coil). Although strand propensity for a position appears to show a better inverse relationship with disorder tendency than helix propensity, there does not appear to be strong difference between score distributions for domains in the all-alpha and all-beta classes (Fig. 6.2-3). There is a lack of strong correlation between disorder propensities and coil propensities (Fig. 3.2-8a),

confirming previous observation of such a difference (Linding et al. 2003a). Simple sequence-based predictors that use coil propensities in place of optimized disorder propensities show substantially worse performance in predicting disorder (Fig. 3.2-9). There is substantial non-monotonicity of the relationship between disorder scores and PSIPRED coil scores, with the highest coil scores being associated with lower disorder scores, on average, than moderately high coil scores (Fig. 6.1-1). These lines of evidence with respect to the relationship between 'disordered' and 'coil' regions give reason to view the 'average' coil and disordered states as different and to use a disorder predictor, rather than a secondary structure predictor's coil scores, to predict disorder.

Many factors may affect whether residues will be missing from deposited structures or not, including the presence or absence of ligands, other experimental conditions, and personal interpretation of data. Nevertheless, large quantities of crystallographic data have yielded a clear trend in disorder parameters. Near-optimal parameters for an effective disorder predictor may be largely reduced to relatively few simple components: side chain hydrophobicity, special parameters for cysteine, proline, (and perhaps asparagine), and general window weight and tail adjustment curve descriptions. This work demonstrates, however, how simply using an experimental hydrophobicity scale for disorder scores could be problematic. An experimental model may work relatively well for aliphatic and polar residues, but not for ionic residues. Experimental values may not be obtained for all twenty residues, and certain values may be inaccurately estimated. An example of this lies in the Hopp-Woods scale, which is generally quite well correlated with optimized disorder values but contains substantial outliers—particularly the strong ionic residues, for which values

were not directly based on experiments. There is a substantial difference in performance of the simple sequence method using the Hopp-Woods scale, in comparison to optimized disorder values (see Fig 3.2-9).

Evidence has been previously provided for an association between 'hydrophobicity' and disorder (Williams et al. 2001; Dosztányi et al. 2005). However, in these cited instances, a specific model under which hydrophobicity is best related to disorder is not clearly defined, and the strengths of disorder/hydrophobicity relationships are not directly determined—i.e., the tight nature of the disorder/hydrophobicity relationship is not clearly demonstrated. The tight linear association between disorder and hydrophobicity has significant implications not strongly supported by previous associations of disorder with 'hydrophobicity', among other properties. It may simplify and clarify understanding of what differentiates residues in their tendencies to be disordered. It gives specific evidence for what might be expected—that crystallographically disordered loops tend to adopt states that involve more indiscriminate interaction between side chain and solvent than solvent-exposed ordered regions, which may maintain more significant hydrophobic interactions while still also entertaining hydrophilic interactions if polar moieties are present (an idea that may have been discussed by Wertz and Scheraga 1978—see Radzicka and Wolfenden 1988).

The relationship between order vs. disorder and hydrophobicity contrasts with the relationship between residue exposure vs. burial and hydrophilicity (Fig. 3.3-1; Radzicka and Wolfenden 1988; Wertz and Scheraga 1978). The overall hydrophobicity of residues on the surface and the strength of hydrophobic interactions appears to play a role in stabilizing a protein's surface, while hydrophilicity is likely quite important in determining how a protein

folds (with relation to which residues may be fully buried), and polar/ionic residues on the surface may play important roles in preventing aggregation and conferring specificity upon binding due to the cost of burying a polar moiety without providing compensating interactions. Perhaps more care should be taken when referring to ‘hydrophobic’ residues. For example, if a position in a large, diverse sequence alignment is observed where all or the great majority of residues are non-polar (for example, I, L, V, and F), then this suggests that this position is sometimes or always completely buried from solvent—this might be best referred to as an ‘apolar’ or ‘non-polar’ pattern rather than a ‘hydrophobic’ pattern. Tryptophan is the most hydrophobic residue, and generally confers stability through hydrophobic interactions, even though it is likely to typically be partially exposed to solvent rather than being buried in a protein or in the middle of a lipid bilayer, due to its also being a polar residue. Tyrosine is similar to tryptophan.

Different types of interactions (hydropathic and otherwise) can be of various importance in different situations. The simple observation that a structural property is hydropathy-related is not as informative as separating that property into components representing more specific types of interactions. Application of fundamental physical concepts might still be improved in efforts to understand structural issues related to globular protein folding, structure, and stability; membrane protein folding and structure; ligand binding; and, of course, disorder.

Care should be taken not to simply equate disorder, solvated regions, flexible regions, linkers, coil regions, etc. The optimized parameters and the disorder/hydrophobicity relationship were derived from X-ray crystallographic structure data. Thus, conclusions about

the physical nature of disordered regions drawn from this work may be most strongly applied to crystallographic disorder. However, there appears to be similarity in disorder propensities in crystallographic disorder and disorder as measured by other means (Williams et al. 2001). Crystallographically disordered regions are typically local and short, but the similarity in residue disorder values for the standard sequence-based predictor and the high specificity predictor suggests that the hydrophobicity relationship also applies to longer disordered regions. Nevertheless, there are likely significant types of disordered regions other than highly solvated loops. What tends to contribute to the disorder in various types (Uversky 2002) of 'intrinsically unstructured proteins' (i.e., globally disordered proteins) is not entirely clear. Kyte-Doolittle hydropathies have been used, along with net charge, in predicting these proteins with some success (Prilusky et al. 2005) (see section 1.2.12) but not necessarily optimally.

As has been noted, 'stability and conformation are not synonymous' (Rose et al. 2006). In contrast to hydrophobic groups, hydrogen bonding groups do not play a significant role in stability but instead in conformational and interaction specificity, due to their requirement to form specific interactions with other hydrogen bonding groups or to be exposed to an aqueous environment (Wertz and Scheraga 1978). Some residues possess both hydrophobic and hydrophilic character (e.g., tryptophan), and may contribute both to stability and to fold or interaction specificity.

In summary, data-based optimization of simple predictors of disorder yields predictors that are substantially better than if disorder propensities are based on available scales that might be presumably related to disorder. Optimized simple predictors are shown

to be similar in performance to DISOPRED2, but also have the benefit of yielding relatively interpretable data-optimized parameters. Furthermore, a tight relationship has been established between disorder and experimental hydrophobicity. Considering this relationship—and contrasting it with the relationship between hydrophilicity and residue exposure or burial—enhances understanding of how hydrophilic and hydrophobic interactions play different fundamental roles in structurally related behavior of proteins.

APPENDIX A
Additional parameter data

Table A-1. Average standard parameter values. Abbreviations: sM: selenomethionine; Unk: unknown/nonstandard residue type; nM: N-terminal methionine; Xtn: 'ghost' extension residue type. (*) N-terminal methionine value is 0 without tail adjustments; non-zero value optimized for predictor with tail adjustments. (See also Supplemental Workbook).

Residue disorder parameters			Window position weights			Tail adjustment parameters				
							Simple		Profile	
Residue type	Simple	Profile	Position	Simple	Profile	Position	N	C	N	C
W	-1.00739	-2.23636	-17	0.0810959	0.0799477	1	2.92236	2.87292	3.01219	2.75639
F	-0.540414	-0.253414	-16	0.0945469	0.0555576	2	2.73524	2.74768	2.76413	2.5723
Y	-0.513589	-1.02542	-15	0.120127	0.0646663	3	2.67317	2.65749	2.6004	2.45406
L	-0.418184	-0.972307	-14	0.120692	0.0719768	4	2.63611	2.60081	2.46224	2.36556
I	-0.514274	-0.639313	-13	0.164119	0.0955245	5	2.38821	2.35462	2.09909	1.99874
V	-0.358167	-0.814371	-12	0.139548	0.13562	6	2.24921	2.16798	1.86401	1.67034
M	-0.216377	0.512026	-11	0.207438	0.203993	7	2.04159	1.98436	1.61635	1.46931
C	-0.540732	-1.43576	-10	0.243471	0.240458	8	1.94916	1.77682	1.46399	1.29269
P	0.32731	0.930165	-9	0.286714	0.271349	9	1.83816	1.62037	1.30031	1.07822
G	0.241088	0.20741	-8	0.380712	0.362605	10	1.6573	1.46352	1.03936	0.924174
A	0.0642762	-0.558905	-7	0.447297	0.449501	11	1.49829	1.34756	0.848874	0.820892
T	0.0333232	-0.597119	-6	0.547931	0.579961	12	1.36491	1.13033	0.703339	0.691018
S	0.400289	2.03082	-5	0.66937	0.740327	13	1.28004	0.982911	0.630538	0.52237
Q	0.336406	0.3561	-4	0.801724	0.857468	14	1.16945	0.948353	0.617375	0.519134
N	0.221683	0.269169	-3	0.941093	0.964805	15	1.11546	0.886143	0.530369	0.426656
H	0.18568	0.228484	-2	1.05196	1.06002	16	1.01185	0.826289	0.47963	0.365645
R	0.176914	0.587904	-1	1.16318	1.13616	17	0.981126	0.823421	0.485875	0.33697
K	0.300523	0.879637	0	1.20872	1.20596	18	0.903084	0.736005	0.47725	0.214245
E	0.33729	1.0084	1	1.15937	1.19012	19	0.872304	0.716983	0.420956	0.276911
D	0.313504	0.14952	2	1.0893	1.13637	20	0.752184	0.61095	0.416503	0.194493
sM	-0.373603		3	0.990943	1.02883	21	0.748407	0.499084	0.430953	0.111944
Unk	0		4	0.870869	0.907538	22	0.763484	0.461497	0.473511	0.072126
nM	0/0.502239*		5	0.746035	0.762543	23	0.676419	0.330615	0.44771	0.037252
Xtn	0		6	0.642842	0.658838	24	0.618042	0.367348	0.447826	0.066394
			7	0.536048	0.55776	25	0.542703	0.233581	0.239495	0.083206
			8	0.466368	0.47139	26	0.521499	0.273114	0.268665	0.050606
			9	0.396445	0.406485	27	0.471237	0.0484837	0.233608	0.045287
			10	0.329051	0.353803	28	0.421336	0.156365	0.090456	0.040655
			11	0.288911	0.288708	29	0.386133	0.114393	0.159885	0.054771
			12	0.284116	0.277521	30	0.285652	0.0785499	0.143035	0.078479
			13	0.246708	0.173134					
			14	0.232329	0.165124					
			15	0.214341	0.19073					
			16	0.171901	0.17734					
			17	0.164679	0.177871					

Table A-2. Standard simple sequence predictor normalized residue disorder parameters (see abbreviations descriptions in Table A-1 caption). As with Table A-1, the N-terminal methionine values are 0 for the predictor without tail adjustments and the values shown for the predictor with tail adjustments.

Residue type	Training run number					
	1	2	3	4	5	Average
W	-0.981474	-1.15887	-0.890212	-0.998254	-1.00815	-1.00739
F	-0.513251	-0.569435	-0.596236	-0.544599	-0.478549	-0.540414
Y	-0.514543	-0.502256	-0.522074	-0.553623	-0.475447	-0.513589
L	-0.404131	-0.419579	-0.424611	-0.431416	-0.411183	-0.418184
I	-0.526755	-0.501785	-0.455878	-0.532025	-0.554929	-0.514274
V	-0.33608	-0.353305	-0.395778	-0.341857	-0.363813	-0.358167
M	-0.279456	-0.245185	-0.203865	-0.168794	-0.184584	-0.216377
C	-0.614056	-0.594474	-0.532113	-0.480098	-0.482917	-0.540732
P	0.334784	0.319255	0.324822	0.338855	0.318835	0.32731
G	0.251483	0.232549	0.239284	0.231184	0.250941	0.241088
A	0.0649797	0.0776471	0.0558383	0.0579555	0.0649605	0.0642762
T	0.0376378	0.0365068	0.0114282	0.0655553	0.0154877	0.0333232
S	0.382256	0.405499	0.399513	0.410855	0.403322	0.400289
Q	0.327662	0.359927	0.369718	0.325243	0.29948	0.336406
N	0.214824	0.214658	0.228237	0.231275	0.219422	0.221683
H	0.166189	0.146834	0.222083	0.211106	0.182186	0.18568
R	0.182866	0.158461	0.181368	0.173441	0.188434	0.176914
K	0.294989	0.306523	0.289887	0.317157	0.294059	0.300523
E	0.331722	0.355504	0.330986	0.335849	0.332387	0.33729
D	0.309075	0.349132	0.332807	0.28588	0.290624	0.313504
sM	-0.357501	-0.393014	-0.366183	-0.290883	-0.460432	-0.373603
Unk	0	0	0	0	0	0
nM	0/0.375914	0/0.582456	0/0.389022	0/0.480076	0/0.338568	0/0.433207
Xtn	0	0	0	0	0	0

Table A-3. Standard simple sequence predictor (sw35_8) window position weights.

Window position	Training run number					
	1	2	3	4	5	Average
-17	0.115072	0.073077	0.03553	0.144771	0.037029	0.081096
-16	0.016718	0.19195	0.110667	0.083625	0.069775	0.094547
-15	0.144982	0.186976	0.121768	0.07112	0.07579	0.120127
-14	0.112646	0.197966	0.150181	0.127839	0.01483	0.120692
-13	0.15361	0.218307	0.105064	0.174095	0.169521	0.164119
-12	0.151778	0.179491	0.083493	0.178716	0.10426	0.139548
-11	0.18592	0.24006	0.172559	0.227509	0.211142	0.207438
-10	0.218298	0.282251	0.209617	0.255946	0.251243	0.243471
-9	0.275447	0.308516	0.270668	0.304082	0.274856	0.286714
-8	0.378721	0.398619	0.330181	0.421462	0.374579	0.380712
-7	0.443281	0.468518	0.414094	0.462704	0.447889	0.447297
-6	0.54438	0.558788	0.517371	0.572851	0.546265	0.547931
-5	0.65377	0.675059	0.664989	0.675303	0.67773	0.66937
-4	0.801217	0.785366	0.78536	0.817266	0.81941	0.801724
-3	0.936319	0.911268	0.945917	0.945877	0.966082	0.941093
-2	1.05993	1.02476	1.04959	1.04905	1.07647	1.05196
-1	1.18527	1.11776	1.15017	1.15811	1.2046	1.16318
0	1.21853	1.15035	1.21641	1.18992	1.2684	1.20872
1	1.17486	1.09581	1.1787	1.14759	1.19989	1.15937
2	1.10448	1.03068	1.09938	1.09988	1.11207	1.0893
3	1.00961	0.941716	0.99523	0.993208	1.01495	0.990943
4	0.909861	0.818672	0.860986	0.878433	0.886391	0.870869
5	0.769424	0.709889	0.733907	0.757798	0.759156	0.746035
6	0.653292	0.605388	0.641497	0.662462	0.651568	0.642842
7	0.543691	0.521369	0.545176	0.543049	0.526956	0.536048
8	0.477271	0.464462	0.490214	0.458566	0.441328	0.466368
9	0.404665	0.400067	0.411322	0.393871	0.372301	0.396445
10	0.326691	0.333995	0.366713	0.308524	0.309331	0.329051
11	0.291399	0.291402	0.329476	0.273414	0.258866	0.288911
12	0.274706	0.253982	0.32384	0.271497	0.296555	0.284116
13	0.229783	0.244883	0.277473	0.237906	0.243495	0.246708
14	0.215083	0.259011	0.264033	0.18404	0.23948	0.232329
15	0.193541	0.217347	0.234657	0.210553	0.215605	0.214341
16	0.17695	0.190454	0.219947	0.072688	0.199468	0.171901
17	0.148795	0.151788	0.19382	0.146267	0.182725	0.164679

Table A-4. Standard profile predictor normalized residue disorder parameters.

Residue type	Training run number					
	1	2	3	4	5	Average
W	-2.09148	-2.61706	-2.10484	-2.0888	-2.2796	-2.23636
F	-0.272436	-0.23715	-0.383806	-0.369084	-0.00459476	-0.253414
Y	-1.0419	-1.05509	-1.01128	-0.930724	-1.08809	-1.02542
L	-0.978854	-0.953488	-1.02118	-0.870715	-1.0373	-0.972307
I	-0.729807	-0.420555	-0.294466	-1.05833	-0.693406	-0.639313
V	-0.714005	-0.920532	-0.976652	-0.529684	-0.930981	-0.814371
M	0.746268	0.155	0.326881	0.660004	0.671975	0.512026
C	-1.50318	-1.27558	-1.41914	-1.62398	-1.35694	-1.43576
P	0.976002	0.94045	0.878351	0.899822	0.9562	0.930165
G	0.203708	0.177357	0.251211	0.230187	0.174587	0.20741
A	-0.610474	-0.525186	-0.47482	-0.673394	-0.510652	-0.558905
T	-0.49623	-0.53368	-0.71771	-0.666015	-0.571961	-0.597119
S	2.00211	1.88305	1.96218	2.18644	2.12034	2.03082
Q	0.418755	0.335255	0.380311	0.285556	0.360622	0.3561
N	0.153391	0.373362	0.341846	0.305805	0.171442	0.269169
H	0.198718	0.166899	0.23351	0.337146	0.206149	0.228484
R	0.610558	0.621636	0.436463	0.669908	0.600952	0.587904
K	0.864976	0.894584	1.04699	0.798693	0.792939	0.879637
E	0.909475	1.08819	0.878528	1.08422	1.08159	1.0084
D	0.178503	0.137327	0.268034	0.0476701	0.116066	0.14952

Table A-5. Standard profile predictor window position weights.

Window position	Training run number					
	1	2	3	4	5	Average
-17	0.017013	0.088907	0.133231	0.078248	0.08234	0.079948
-16	0.017552	0.030972	0.073181	0.053874	0.102209	0.055558
-15	0.022917	0.047326	0.114234	0.037145	0.10171	0.064666
-14	0.026025	0.120131	0.012032	0.093898	0.107797	0.071977
-13	0.126485	0.181741	0.066548	0.027241	0.075608	0.095525
-12	0.10354	0.104397	0.150097	0.182928	0.137139	0.13562
-11	0.183511	0.230327	0.196159	0.215022	0.194948	0.203993
-10	0.214491	0.25926	0.207021	0.262806	0.258714	0.240458
-9	0.249253	0.316021	0.224052	0.282858	0.28456	0.271349
-8	0.371762	0.40504	0.312477	0.360584	0.363162	0.362605
-7	0.468579	0.449815	0.411092	0.464358	0.453659	0.449501
-6	0.608507	0.563704	0.560012	0.594176	0.573404	0.579961
-5	0.772676	0.724139	0.726149	0.746663	0.732009	0.740327
-4	0.884612	0.855578	0.842851	0.865021	0.839276	0.857468
-3	0.989309	0.948756	0.953995	0.967486	0.964476	0.964805
-2	1.07997	1.04883	1.05692	1.06	1.05438	1.06002
-1	1.16679	1.10768	1.14318	1.12682	1.13632	1.13616
0	1.2278	1.18035	1.2236	1.18579	1.21225	1.20596
1	1.20561	1.16322	1.21238	1.18308	1.18631	1.19012
2	1.15834	1.09531	1.17186	1.12401	1.13231	1.13637
3	1.04427	1.01774	1.04017	1.01535	1.02663	1.02883
4	0.947668	0.900626	0.912686	0.888584	0.888125	0.907538
5	0.785478	0.766792	0.744848	0.764501	0.751096	0.762543
6	0.685634	0.679634	0.654048	0.645084	0.629792	0.658838
7	0.562415	0.556191	0.590561	0.553921	0.52571	0.55776
8	0.479626	0.457896	0.509658	0.465259	0.444511	0.47139
9	0.419713	0.393795	0.431992	0.402685	0.384238	0.406485
10	0.394628	0.312392	0.369269	0.35628	0.336446	0.353803
11	0.308664	0.267835	0.292386	0.289764	0.284893	0.288708
12	0.283267	0.284649	0.273019	0.272334	0.274334	0.277521
13	0.014283	0.196372	0.215059	0.231336	0.208619	0.173134
14	0.179404	0.195753	0.081787	0.171605	0.197073	0.165124
15	0.158007	0.200946	0.205394	0.191813	0.197492	0.19073
16	0.166223	0.151515	0.21592	0.16617	0.186871	0.17734
17	0.175974	0.19636	0.172138	0.1733	0.171585	0.177871

Table A-6. Simple sequence predictor tail adjustment weights (sw35_st30_8).

Amino-terminal					Carboxy-terminal				
1	2	3	4	5	1	2	3	4	5
2.95122	2.92139	2.88618	2.75912	3.09391	2.87765	2.86338	2.94348	2.83384	2.84627
2.7422	2.71433	2.73129	2.57017	2.9182	2.7382	2.72483	2.80311	2.6961	2.77617
2.69202	2.65603	2.66935	2.49931	2.84913	2.66822	2.62347	2.73143	2.61471	2.64963
2.62266	2.64305	2.62332	2.4697	2.82183	2.59164	2.57489	2.67484	2.55255	2.61011
2.39744	2.42338	2.37943	2.22882	2.51196	2.32182	2.3496	2.43951	2.29603	2.36616
2.26809	2.2525	2.22591	2.09992	2.39965	2.15267	2.14715	2.29833	2.11822	2.12355
2.07061	2.08518	2.00394	1.8712	2.177	1.95088	1.9882	2.0996	1.91267	1.97044
2.0109	1.9641	1.86316	1.82485	2.08281	1.75794	1.78645	1.86391	1.69635	1.77947
1.88323	1.88436	1.77469	1.69793	1.95061	1.60245	1.6285	1.72289	1.56429	1.58372
1.73963	1.6344	1.59361	1.54045	1.77843	1.45094	1.48963	1.54105	1.4123	1.42367
1.54142	1.47213	1.46163	1.40738	1.60889	1.34757	1.42583	1.39471	1.28478	1.28489
1.38652	1.3365	1.33124	1.25732	1.51296	1.16484	1.15849	1.18742	1.05833	1.08258
1.28602	1.3041	1.23413	1.15253	1.42344	0.997914	0.975544	1.04956	0.959134	0.932402
1.17155	1.16964	1.19471	0.989852	1.3215	0.965862	0.985113	0.976916	0.90061	0.913264
1.13186	1.11569	1.05547	0.998223	1.27604	0.900207	0.942069	0.940972	0.861653	0.785814
0.999379	0.979297	0.975673	0.939135	1.16579	0.895872	0.908994	0.857644	0.762732	0.706204
1.00977	0.969899	0.895169	0.902013	1.12878	0.897527	0.848942	0.894894	0.783116	0.692626
0.919833	0.882011	0.851562	0.826684	1.03533	0.797916	0.757788	0.776663	0.665118	0.682541
0.841003	0.930447	0.855604	0.744537	0.98993	0.773312	0.728921	0.792134	0.639136	0.651414
0.649349	0.742536	0.762833	0.687687	0.918513	0.678893	0.602246	0.682521	0.642107	0.448984
0.774696	0.722433	0.750266	0.661668	0.832973	0.616518	0.47774	0.500184	0.567448	0.333528
0.860596	0.72213	0.805107	0.63192	0.797667	0.564332	0.495797	0.499904	0.429144	0.31831
0.699155	0.675839	0.755939	0.594616	0.656546	0.402511	0.391994	0.413192	0.295742	0.149637
0.641918	0.578148	0.588893	0.592867	0.688385	0.404681	0.333733	0.426928	0.422549	0.248848
0.525485	0.572256	0.517095	0.504277	0.594403	0.230442	0.253457	0.327271	0.208478	0.148259
0.50425	0.555744	0.545916	0.496856	0.504731	0.24462	0.409565	0.283658	0.078142	0.349584
0.485847	0.542437	0.40723	0.448216	0.472454	0.025462	0.020112	0.024092	0.031243	0.141509
0.434298	0.41043	0.415835	0.444438	0.40168	0.244944	0.128885	0.066935	0.235698	0.105365
0.254242	0.42818	0.306552	0.509931	0.431761	0.251597	0.072864	0.043097	0.042158	0.162251
0.07064	0.402114	0.211555	0.372051	0.371899	0.17296	0.02949	0.053476	0.104693	0.032131

Table A-7. Profile predictor tail adjustment weights (p2w35_st30_6).

Amino-terminal					Carboxy-terminal				
1	2	3	4	5	1	2	3	4	5
3.06922	2.9956	3.05903	2.84037	3.09672	2.85854	2.6035	2.79042	2.72584	2.80366
2.8145	2.72965	2.8403	2.57816	2.85803	2.68281	2.41594	2.58911	2.5211	2.65256
2.64033	2.60113	2.63615	2.46082	2.66358	2.52904	2.29022	2.50284	2.4284	2.51979
2.52308	2.43623	2.50523	2.29731	2.54933	2.47735	2.21144	2.37356	2.3619	2.40356
2.12913	2.10687	2.11374	1.97505	2.17068	2.08683	1.85937	2.04007	2.01537	1.99208
1.8689	1.88651	1.89506	1.7138	1.95576	1.71439	1.56824	1.75922	1.68059	1.62926
1.6201	1.63726	1.61811	1.51545	1.69083	1.51411	1.31585	1.57172	1.52769	1.41719
1.4764	1.49354	1.46297	1.34681	1.54025	1.34689	1.2029	1.38345	1.30875	1.22148
1.35923	1.26778	1.27759	1.23008	1.36686	1.15164	0.966148	1.13833	1.1037	1.03126
1.09527	0.93518	0.989914	1.02264	1.1538	1.02421	0.822445	0.956797	0.957503	0.859916
0.834503	0.790989	0.820054	0.869739	0.929083	0.930647	0.740141	0.842163	0.830073	0.761437
0.698475	0.70142	0.688236	0.650201	0.778365	0.783838	0.642468	0.740091	0.620224	0.668467
0.625319	0.628164	0.583726	0.62087	0.694613	0.615096	0.503477	0.49992	0.523806	0.469551
0.639186	0.654134	0.590606	0.540729	0.66222	0.642644	0.54821	0.506071	0.43214	0.466605
0.515409	0.450939	0.535996	0.488431	0.66107	0.524963	0.446889	0.461965	0.3463	0.353163
0.490839	0.406112	0.483955	0.450352	0.566892	0.530508	0.336137	0.371206	0.210417	0.379955
0.480844	0.483709	0.469996	0.489319	0.505509	0.529777	0.0899301	0.457047	0.256224	0.35187
0.431788	0.533873	0.478843	0.427692	0.514054	0.39582	0.225915	0.0699969	0.192204	0.187288
0.406164	0.464302	0.443014	0.273624	0.517674	0.422884	0.240674	0.389178	0.215653	0.116166
0.405515	0.411242	0.413671	0.41215	0.439937	0.35419	0.173044	0.283571	0.0215711	0.140087
0.514222	0.509831	0.489043	0.113483	0.528185	0.204253	0.0634088	0.172524	0.0651806	0.0543519
0.501189	0.435677	0.530227	0.350939	0.549524	0.0518996	0.0288032	0.141393	0.0343	0.104233
0.498561	0.327244	0.595595	0.252578	0.56457	0.0284103	0.020274	0.0870711	0.0244622	0.0260432
0.442373	0.367486	0.584159	0.325496	0.519615	0.0588513	0.0452898	0.080633	0.089147	0.0580466
0.332665	0.230197	0.413059	0.0523173	0.169236	0.199379	0.0528933	0.0401139	0.0749777	0.0486681
0.201654	0.291364	0.406757	0.25301	0.190541	0.136381	0.0290707	0.0330476	0.031061	0.0234691
0.264329	0.0308322	0.340227	0.299649	0.233004	0.0312199	0.0440925	0.0255211	0.0280191	0.0975845
0.225599	0.122397	0.0209585	0.046923	0.0364001	0.023547	0.0506543	0.0470612	0.0563806	0.0256308
0.151323	0.221083	0.102846	0.11411	0.210065	0.0667974	0.0225231	0.0205237	0.0749558	0.0890538
0.0295684	0.0595531	0.369945	0.0309509	0.225159	0.0488324	0.0558655	0.0930292	0.120292	0.074376

APPENDIX B
AAIndex search results

Table B-1. Associations found with other scales from searches of AAIndex with H, R, K, D, E, C, and P excluded in calculation of R^2. (**Note:** The values given in the AAIndex1 residue tables were used in calculating R^2. In some instances in these tables, residues without actual values are given a default value of 0, or one value is selected if multiple are given for a single residue; this exercise was not key to making our initial findings, and no effort was made to manually exclude residues without actual values or to take into account alternative residue values).

R^2 range	R^2	AAIndex ID	Note
0.95 - 1	0.9775	HOPT810101	Hopp/Woods
	0.9775	LEVM760101	
	0.9675	NOZY710101	Nozaki/Tanford
0.9-0.95	0.9410	RADA880102	Radzicka/Wolfenden oct/wat
	0.9386	JOND750101	
	0.9384	ARGP820101	
	0.9269	TAKK010101	
	0.9237	MEEJ800102	
	0.9183	CIDH920102	
	0.9166	MEEJ810101	
	0.9122	VINM940101	
	0.9068	MEEJ810102	
	0.9004	SIMZ760101	
0.85-0.9	0.8998	VINM940102	
	0.8794	GOLD730101	
	0.8792	OOBM770103	
	0.8775	CIDH920105	
	0.8766	FAUJ830101	
	0.8646	GUOD860101	
0.8 - 0.85	0.8469	PARJ860101	
	0.8409	KARP850101	
	0.8376	FUKS010104	
	0.8314	WOLS870101	
	0.8301	NISK860101	14 Ang contact number
	0.8268	ROSG850101	
	0.8209	EISD860101	
	0.8140	CIDH920104	
	0.8109	VINM940103	
	0.8047	ZIMJ680105	
0.75 - 0.8	0.7981	PARS000101	
	0.7945	BULH740101	

	0.7903	MEEJ800101
	0.7869	LEVM760107
	0.7861	WERD780101
	0.7818	WIMW960101
	0.7809	CIDH920101
	0.7769	ZASB820101
	0.7731	PLIV810101
	0.7725	CHOP780213
	0.7719	LEVM760106
	0.7715	ROBB790101
	0.7693	GRAR740103
	0.7618	BIOV880101
	0.7558	NADH010104
	0.7555	VENT840101
	0.7496	NADH010105
	0.7402	GRAR740102
	0.7397	WEBA780101
	0.7357	BROC820102
	0.7316	BIOV880102
	0.7286	GOLD730102
0.7 -	0.7220	TSAJ990101
0.75	0.7209	BIGC670101
	0.7185	KRIW790101
	0.7133	TSAJ990102
	0.7127	NADH010106
	0.7114	FUKS010102
	0.7046	ROSG850102
	0.7028	KRIW790103

Table B-2. Associations found with other scales from searches of AAIndex including all residues in calculation of R^2. See note in caption of Table B-1.

0.95 - 1			
0.9-0.95			
0.85-0.9	0.8646	VINM940102	
0.8-0.85	0.8395	NISK860101	14 Ang contact number
	0.8386	CIDH920104	
	0.8371	OOBM770103	
	0.8346	MEEJ810101	
	0.8336	CIDH920102	
	0.8314	CIDH920105	
	0.8084	MEEJ810102	
0.75-0.8	0.7934	VINM940101	
	0.7872	FAUJ830101	
	0.7789	NOZY710101	
	0.7729	NADH010104	
	0.7717	WERD780101	
	0.7686	BIOV880101	
	0.7686	PARJ860101	
	0.7614	ROBB790101	
	0.7612	PONP930101	
	0.7597	MIYS850101	
	0.7569	GUOD860101	
	0.7532	KRIW790101	
	0.7500	VINM940103	
0.7-0.75	0.7481	NADH010105	
	0.7384	PARS000101	
	0.7355	NADH010103	
	0.7296	NISK800101	
	0.7252	GRAR740102	
	0.7168	PONP800108	
	0.7136	MEIH800101	
	0.7133	ROSG850102	
	0.7089	CIDH920103	
	0.7084	BIOV880102	
	0.7061	PLIV810101	

APPENDIX C
PSIPRED

Table C-1. For each bin combining PSIPRED coil, helix, and strand score information, any significant tendency toward being ordered or disordered ('Gap' or 'Nongap') is listed, as determined by two-tailed, pairwise *t*-test *p*-values. (This is not intended to be a perfect statistical test.) For any given test set, relative bin frequencies are produced in two different ways—either by dividing by the number of residues within that bin by the total number of residues in all bins, or by dividing by the total number of residues within its given coils category. As an example, the sixteenth entry gives coil category = 1, helix category = 4, strand category = 1; meaning this bin includes a count of all residues that are assigned by PSIPRED a coil score between 0.000 and 0.199, a helix score between 0.800 and 1.000, and a strand score between 0.000 and 0.199. When the frequencies of ordered and disordered residues within this bin for the 5 cross-validation test sets are compared (ordered vs. disordered), it is found that when the frequencies are calculated relative to all residues, that residues in this bin tend to be ordered, with a *p*-value for this tendency less than 0.1, but greater than 0.05. This makes sense, because residues with low coil scores would be expected to tend to be ordered. However, when frequencies are calculated relative to the coil category—comparing only with other residues that receive a low coil score, residues that fall within this particular bin tend to be disordered when compared to other residues with low coil scores, and this difference has a calculated *p*-value of < 0.05. A similar pattern also occurs for the twenty-first entry: coil category = 1, helix category = 5, strand category = 1. Therefore, it appears that for residues with weak coil scores, a strong helix score and a weak strand score makes it more likely to be disordered.

			p < 0.1		p < 0.05	
Coil bin	Helix bin	Strand bin	by all	by coil bin	by all	by coil bin
1	1	1				
1	1	2	Nongap	Nongap		
1	1	3				
1	1	4	Nongap		Nongap	
1	1	5	Nongap	Nongap	Nongap	Nongap
1	2	1				
1	2	2				
1	2	3				
1	2	4	Nongap			
1	2	5				
1	3	1				
1	3	2				
1	3	3				
1	3	4				
1	3	5				
1	4	1	Nongap	Gap		Gap
1	4	2				

1	4	3	Nongap		Nongap	
1	4	4				
1	4	5				
1	5	1	Nongap	Gap	Nongap	Gap
1	5	2	Nongap		Nongap	
1	5	3				
1	5	4				
1	5	5				
2	1	1	Nongap	Nongap		
2	1	2				
2	1	3				
2	1	4	Nongap	Nongap	Nongap	Nongap
2	1	5	Nongap	Nongap	Nongap	Nongap
2	2	1	Nongap	Nongap		Nongap
2	2	2	Gap	Gap	Gap	
2	2	3				
2	2	4				
2	2	5				
2	3	1	Gap	Gap	Gap	
2	3	2	Gap	Gap	Gap	Gap
2	3	3	Nongap	Nongap	Nongap	Nongap
2	3	4				
2	3	5				
2	4	1	Gap	Gap	Gap	Gap
2	4	2				
2	4	3				
2	4	4				
2	4	5				
2	5	1				
2	5	2				
2	5	3				
2	5	4				
2	5	5				
3	1	1	Nongap	Nongap	Nongap	Nongap
3	1	2	Gap			
3	1	3	Nongap	Nongap	Nongap	Nongap
3	1	4	Nongap	Nongap	Nongap	Nongap
3	1	5				
3	2	1	Gap	Gap	Gap	Gap
3	2	2	Gap	Gap	Gap	Gap
3	2	3				
3	2	4				
3	2	5				
3	3	1	Gap	Gap	Gap	Gap
3	3	2				
3	3	3				
3	3	4				

3	3	5				
3	4	1				
3	4	2				
3	4	3				
3	4	4				
3	4	5				
3	5	1				
3	5	2				
3	5	3				
3	5	4				
3	5	5				
4	1	1	Gap	Gap	Gap	Gap
4	1	2		Nongap		Nongap
4	1	3	Nongap	Nongap	Nongap	Nongap
4	1	4				
4	1	5				
4	2	1	Gap	Gap	Gap	Gap
4	2	2				
4	2	3				
4	2	4				
4	2	5				
4	3	1		Nongap		Nongap
4	3	2				
4	3	3				
4	3	4				
4	3	5				
4	4	1				
4	4	2				
4	4	3				
4	4	4				
4	4	5				
4	5	1				
4	5	2				
4	5	3				
4	5	4				
4	5	5				
5	1	1	Gap	Gap	Gap	Gap
5	1	2	Nongap	Nongap	Nongap	Nongap
5	1	3	Nongap	Nongap	Nongap	Nongap
5	1	4				
5	1	5				
5	2	1				
5	2	2				
5	2	3				
5	2	4				
5	2	5				
5	3	1				

5	3	2		
5	3	3		
5	3	4		
5	3	5		
5	4	1		
5	4	2		
5	4	3		
5	4	4		
5	4	5		
5	5	1		
5	5	2		
5	5	3		
5	5	4		
5	5	5		

APPENDIX D
Code

Some of the following code gives an overview of the details of the individual runs (e.g., number of annealing steps, parameter start file, etc.). Minor editing of the original code has been done, not in a way that should actually affect results (e.g., removing extraneous commented out lines).

This section provides code used for training, code used for normalization/production of final parameters, samples of code used for analysis, and a table listing the functions of some of the code used in relation to the datasets (such as processing PDB files and obtaining cross validation sets). Code has sometimes been reformatted/adjusted from that in the original files, with no material changes. This section is not comprehensive in coverage of code used for generating results described in this dissertation. Code may be made publicly available at http://prodata.swmed.edu.

D.1 OPTIMIZATION CODE

sw35_8.cpp: Training standard simple sequence-based predictor

```
#include "annealer1.hpp"
#include "simple_window35.hpp"
using namespace SimpleWindow35;

int main(int argc, char *argv[])
{       typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
                DomainDivisionNeutralizer<(tail_length + 1)/2>) > preprocessor_tp;
        Annealer1<scorer_tp, SCOPDataset, preprocessor_tp, default_seq_reader_tp> annealer;
        int rseed, set_num;

  try
  {     if(argc != 2)
        {       cerr << "Should have run number as an argument." << endl; exit(1);
        }
        set_num = MyString::string2int(argv[1]);
        rseed = set_num * 1000000 + 5000000;
        srand(rseed);   // Just affects Metropolis decisions, I think
        annealer.dataset.preprocessor.set_num_terminal(tail_length + 1);
        annealer.set_set_num(set_num);
        annealer.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1");
        annealer.set_primary_dir(default_primary_dir);
        annealer.set_run_base_name("sw35_8");
        annealer.set_start_file_base_name("simple_window35");
        annealer.read_start_file(1);
        annealer.genome.randomize();
        annealer.evaluator.set_roc_false_pos_fraction(0.5);
        annealer.set_optimization_direction(Maximize);
        annealer.enviro.set_fractions(0.5, 0.2);
        annealer.genome.set_temperatures(0.4);
        annealer.set_temperature(0.002);
        annealer.set_all_increments(1.01395948);        // Temp halves every 50 steps
        annealer.set_cycles_per_step(250);
        annealer.anneal(700);
  } CATCHES
}
```

sw35_7.cpp: Training 'high specificity' sequence-based predictor

```
#include "annealer1.hpp"
#include "simple_window35.hpp"
using namespace SimpleWindow35;

int main(int argc, char *argv[])
{       typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
                DomainDivisionNeutralizer<(tail_length + 1)/2>) > preprocessor_tp;
        Annealer1<scorer_tp, SCOPDataset, preprocessor_tp, default_seq_reader_tp> annealer;
        int rseed, set_num;

  try
  {     if(argc != 2)
        {       cerr << "Should have run number as an argument." << endl; exit(1);
        }

        set_num = MyString::string2int(argv[1]);
        rseed = set_num * 1000000 + 5000000;
        srand(rseed);
        annealer.dataset.preprocessor.set_num_terminal(tail_length + 1);
        annealer.set_set_num(set_num);
        annealer.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1");
        annealer.set_primary_dir(default_primary_dir);
        annealer.set_run_base_name("sw35_7");
        annealer.set_start_file_base_name("simple_window35");
```

```
        annealer.read_start_file(1);
        annealer.genome.randomize();
/**/    annealer.evaluator.set_roc_false_pos_fraction(0.05);
        annealer.set_optimization_direction(Maximize);
        annealer.enviro.set_fractions(0.5, 0.2);        //??
        annealer.genome.set_temperatures(0.4);
        annealer.set_temperature(0.002);
        annealer.set_all_increments(1.01395948);        // Temp halves every 50 steps
        annealer.set_cycles_per_step(250);
        annealer.anneal(700);
  } CATCHES
}
```

p2w35_4.cpp: Training standard profile-based predictor

```
#include "annealer1.hpp"
#include "profile2_window35.hpp"
using namespace Profile2Window35;

int main(int argc, char *argv[])
{       typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
                DomainDivisionNeutralizer<(tail_length + 1)/2>) > preprocessor_tp;
        Annealer1<scorer_tp, SCOPDataset, preprocessor_tp, default_seq_reader_tp> annealer;
        int rseed, set_num;

  try
  {     if(argc != 2)
        {       cerr << "Should have run number as an argument." << endl; exit(1);
        }
        set_num = MyString::string2int(argv[1]);
        annealer.dataset.preprocessor.set_num_terminal(tail_length + 1);
        annealer.set_set_num(set_num);
        annealer.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1_prof");
        annealer.set_primary_dir(default_primary_dir);
        annealer.set_run_base_name("p2w35_4");
        annealer.set_start_file_base_name("profile2_window35");
        annealer.read_start_file(1);
        annealer.genome.randomize();
        annealer.evaluator.set_roc_false_pos_fraction(0.5);
        annealer.set_optimization_direction(Maximize);
        annealer.enviro.set_fractions(0.5, 0.2);
        annealer.genome.set_temperatures(0.4);
        annealer.set_temperature(0.002);
        annealer.set_all_increments(1.01395948);        // Temp halves every 50 steps
        annealer.set_cycles_per_step(250);
        annealer.anneal(700);
  } CATCHES
}
```

sw35_st30_8.cpp: Training simple sequence-based predictor tail adjustments and N-terminal methionine

```
#include "annealer1.hpp"
#include "simple_window35_simple_tail30.hpp"
using namespace SimpleWindow35_SimpleTail30;

int main(int argc, char *argv[])
{ try {
        typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
                DomainDivisionNeutralizer<9>) > preprocessor_tp;
        Annealer1<scorer_tp, SCOPDataset, preprocessor_tp, default_seq_reader_tp> annealer;
        int rseed, set_n;

        if(argc != 2)
        {       cerr << "Should have run number as an argument." << endl; exit(1);
        }
        set_n = MyString::string2int(argv[1]);
```

```
        annealer.dataset.preprocessor.set_num_terminal(0);
/**/    annealer.dataset.preprocessor.set_min_his_tag_terminus_noncounted_length(30);
        annealer.set_set_num(set_n);
        annealer.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1");
        annealer.set_primary_dir(default_primary_dir);
        annealer.set_run_base_name("sw35_st30_8");
        annealer.set_start_file_base_name("simple_window35_simple_tail30");
        annealer.read_start_file(6, set_n);
        annealer.genome.randomize();
        annealer.evaluator.set_roc_false_pos_fraction(0.5);
        annealer.set_optimization_direction(Maximize);
        annealer.enviro.set_fractions(0.5, 0.2);
        annealer.genome.set_temperatures(0.4);
        annealer.set_temperature(0.001);
        annealer.set_all_increments(1.01395948);        // Temp halves every 50 steps
        annealer.set_cycles_per_step(250);
        annealer.anneal(700);
  } CATCHES
}
```

p2w35_st30_5.cpp: Training profile-based predictor tail adjustments

```
#include "annealer1.hpp"
#include "profile2_window35_simple_tail30.hpp"
#include "my_string.hpp"
using namespace Profile2Window35_SimpleTail30_Fast;

int main(int argc, char *argv[])
{ try
  {     typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
                DomainDivisionNeutralizer<(tail_length + 1)/2>) > preprocessor_tp;
        Annealer1<scorer_tp, SCOPDataset, preprocessor_tp, default_seq_reader_tp> annealer;
        Genome preproc_param;
        int dummy = 0, set_num;

        if(argc != 2)
        {       cerr << "Should have run number as an argument." << endl; exit(1);
        }
        set_num = MyString::string2int(argv[1]);
                // Profile2Window35_SimpleTail30::preproc_file_path
        preproc_param.read(preproc_file_path(4, set_num));
        annealer.dataset.preprocessor.import(preproc_param, 0, dummy);
        annealer.dataset.preprocessor.set_num_terminal(0);
        annealer.dataset.preprocessor.set_min_his_tag_terminus_noncounted_length(30);
        annealer.set_set_num(set_num);
        annealer.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1_prof");
        annealer.set_primary_dir(default_primary_dir);
        annealer.set_run_base_name("p2w35_st30_5");
        annealer.set_start_file_base_name("profile2_window35_simple_tail30_tails");
        annealer.read_start_file(2);
        annealer.genome.randomize();
        annealer.evaluator.set_roc_false_pos_fraction(0.5);
        annealer.set_optimization_direction(Maximize);
        annealer.enviro.set_fractions(0.5, 0.2);        //??
        annealer.genome.set_temperatures(0.2);
        annealer.set_temperature(0.001);
        annealer.set_all_increments(1.01395948);        // Temp halves every 50 steps
        annealer.set_cycles_per_step(250);
        annealer.anneal(700);
  } CATCHES
}
```

sw9_1.cpp: Training short window (length = 9) predictor (standard $ROC_{0.5}$ optimization performance measure used)

```
#include "annealer1.hpp"
#include "simple_window9.hpp"
```

```
using namespace SimpleWindow9;

int main(int argc, char *argv[])
{       typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
                DomainDivisionNeutralizer<9>) > preprocessor_tp;
        Annealer1<scorer_tp, SCOPDataset, preprocessor_tp, default_seq_reader_tp> annealer;
        int rseed, set_num;

  try
  {     if(argc != 2)
        {       cerr << "Should have run number as an argument." << endl; exit(1);
        }
        annealer.dataset.preprocessor.set_num_terminal(18);
        set_num = MyString::string2int(argv[1]);
        annealer.set_set_num(set_num);
        annealer.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1");
        annealer.set_primary_dir(default_primary_dir);
        annealer.set_run_base_name("sw9_1");
        annealer.set_start_file_base_name("simple_window9");
        annealer.read_start_file(1);
        annealer.genome.randomize();
        annealer.evaluator.set_roc_false_pos_fraction(0.5);
        annealer.set_optimization_direction(Maximize);
        annealer.enviro.set_fractions(0.5, 0.2);
        annealer.genome.set_temperatures(0.4);
        annealer.set_temperature(0.002);
        annealer.set_all_increments(1.01395948);        // Temp halves every 50 steps
        annealer.set_cycles_per_step(150);
        annealer.anneal(700);
  } CATCHES
}
```

sw9_2.cpp: Training short window (length = 9) predictor, $ROC_{1.0}$ optimization performance measure used

```
#include "annealer1.hpp"
#include "simple_window9.hpp"
using namespace SimpleWindow9;

int main(int argc, char *argv[])
{       typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
                DomainDivisionNeutralizer<9>) > preprocessor_tp;
        Annealer1<scorer_tp, SCOPDataset, preprocessor_tp, default_seq_reader_tp> annealer;
        int rseed, set_num;

  try
  {     if(argc != 2)
        {       cerr << "Should have run number as an argument." << endl; exit(1);
        }

        annealer.dataset.preprocessor.set_num_terminal(18);
        set_num = MyString::string2int(argv[1]);
        annealer.set_set_num(set_num);
        annealer.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1");
        annealer.set_primary_dir(default_primary_dir);
        annealer.set_run_base_name("sw9_2");
        annealer.set_start_file_base_name("simple_window9");
        annealer.read_start_file(1);
        annealer.genome.randomize();
/**/    annealer.evaluator.set_roc_false_pos_fraction(1);
        annealer.set_optimization_direction(Maximize);
        annealer.enviro.set_fractions(0.5, 0.2);
        annealer.genome.set_temperatures(0.4);
        annealer.set_temperature(0.002);
        annealer.set_all_increments(1.01395948);        // Temp halves every 50 steps
        annealer.set_cycles_per_step(150);
        annealer.anneal(700);
  } CATCHES
}
```

D.2 NORMALIZATION CODE

sw35_8_score_dbn_adjust.cpp

```
#include "simple_window35.hpp"
#include "score_distribution_adjustment_client.hpp"

int main()
{ try
  {     using namespace SimpleWindow35;
        typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
                DomainDivisionNeutralizer<(tail_length + 1)/2>) > preprocessor_tp;
        ScoreDistributionAdjustmentClient<scorer_tp, SCOPDataset, preprocessor_tp,
                default_seq_reader_tp> client;

        client.dataset.preprocessor.set_num_terminal(tail_length + 1);
        client.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1");
        client.set_primary_dir("SimpleWindow35");
        client.set_run_base_name("sw35_8");
        client.set_trial_nums(1, 1);
        client.set_step_num(700);
        client.analyzer.set_num_iterations(2);
        client.gather_results();
  } CATCHES
}
```

p2w35_4_score_dbn_adjust.cpp

```
#include "profile2_window35.hpp"
#include "score_distribution_adjustment_client.hpp"

int main()
{ try
  {     using namespace Profile2Window35;
        typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
                DomainDivisionNeutralizer<(tail_length + 1)/2>) > preprocessor_tp;
        ScoreDistributionAdjustmentClient<scorer_tp, SCOPDataset, preprocessor_tp,
                default_seq_reader_tp> client;

        client.dataset.preprocessor.set_num_terminal(tail_length + 1);
        client.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1_prof");
        client.set_primary_dir(default_primary_dir);
        client.set_run_base_name("p2w35_4");
        client.set_trial_nums(1, 1);
        client.set_step_num(700);
/**/    client.analyzer.set_num_iterations(1);
        client.analyzer.set_predictor_type(ScoreDistributionAdjustmentAnalyzer<scorer_tp>::Profile)
;
        client.gather_results();
  } CATCHES
}
```

sw35_st30_8_700_avg_params.cpp

```
#include "genome_averager.hpp"
#include "file_name_handler.hpp"

int step_num = 700;

int main()
{       FileNameHandler fnh;
        GenomeAverager averager;
        vector<Genome *> v_params;
```

```
        Genome params;
        int set_n;

        fnh.set_primary_dir("SimpleWindow35_SimpleTail30");
        fnh.set_run_base_name("sw35_st30_8");
        for(set_n = 1; set_n <= 5; ++set_n)
        {       v_params.push_back(new Genome(fnh.parm_file_path(set_n, step_num)));
        }
        params = averager.average(v_params);
        params.write(fnh.averaged_param_path(700));
}
```

(Final parameter sets for p2w35_st30_6 were obtained manually)

sw9_1_score_dbn_adjust.cpp

```
#include "simple_window9.hpp"
#include "score_distribution_adjustment_client.hpp"

int main()
{ try
  {     using namespace SimpleWindow9;
        typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
DomainDivisionNeutralizer<9>) > preprocessor_tp;
        ScoreDistributionAdjustmentClient<scorer_tp, SCOPDataset, preprocessor_tp,
default_seq_reader_tp> client;

        client.dataset.preprocessor.set_num_terminal(18);
        client.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1");
        client.set_primary_dir(default_primary_dir);
        client.set_run_base_name("sw9_1");
        client.set_trial_nums(1, 1);
        client.set_step_num(700);
        client.analyzer.set_num_iterations(2);
        client.gather_results();
  } CATCHES
}
```

sw9_2_score_dbn_adjust.cpp

```
#include "simple_window9.hpp"
#include "score_distribution_adjustment_client.hpp"

int main()
{ try
  {     using namespace SimpleWindow9;
        typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
DomainDivisionNeutralizer<(tail_length + 1)/2>) > preprocessor_tp;
        ScoreDistributionAdjustmentClient<scorer_tp, SCOPDataset, preprocessor_tp,
default_seq_reader_tp> client;

        client.dataset.preprocessor.set_num_terminal(tail_length + 1);
        client.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1");
        client.set_primary_dir("SimpleWindow9");
        client.set_run_base_name("sw9_2");
        client.set_trial_nums(1, 1);
        client.set_step_num(700);
        client.analyzer.set_num_iterations(2);
        client.gather_results();
  } CATCHES
}
```

D.3 SAMPLES OF CODE FOR OTHER ANALYSES

sw35_8_700_norm_param_report.cpp: Produce a report of parameters in tab-delimited form that can be opened and easily used in spreadsheet form [Results found at SimpleWindow35/sw35_8/NormalizedScoreParams/sw35_8_700_train_norm_scr_params_report_1.txt]

```
#include "parameter_report_maker.hpp"

int main()
{       SingleScorerParmReportMaker<SingleWindowParmReportModule> r_m;
try
  {     r_m.set_step_num(700);
        r_m.set_primary_dir("SimpleWindow35");
        r_m.set_run_base_name("sw35_8");
        r_m.set_norm_param_trial_num(1);
        r_m.set_score_type(FileNameHandler::Normalized);
        r_m.create_report();
  } CATCHES
}
```

sw35_8_test_roc_prof_exc0t_htx30.cpp: Obtaining ROC scores at different cutoffs [note that file name reflects that 0 terminal residues were excluded, except for sequence ends where polyhistidine tags were excluded, in which case, the first thirty residues at that end are excluded; results found at SimpleWindow35/sw35_8/ROC/sw35_8_700_test_norm_exc0t_htx30_prof_roc_all_1.txt]

```
#include "simple_window35.hpp"
#include "roc_score_analysis_client.hpp"

int main()
{ try
  {     using namespace SimpleWindow35;
        typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
                DomainDivisionNeutralizer<9>) > preprocessor_tp;
        ROCScoreAnalysisClient<scorer_tp, SCOPDataset, preprocessor_tp, default_seq_reader_tp>
                client;

        client.dataset.preprocessor.set_num_terminal(0);
        client.dataset.preprocessor.set_min_his_tag_terminus_noncounted_length(30);
        client.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1_prof");
        client.set_primary_dir(default_primary_dir);
        client.set_run_base_name("sw35_8");
        client.set_score_type(FileNameHandler::Normalized);
        client.set_trial_nums(1, 1);
        client.set_step_num(700);
        client.set_report_token("exc0t_htx30_prof_roc");
        client.gather_results();
  } CATCHES
}
```

sw35_8_test_snsp_prof_exc0t_htx30.cpp: Obtaining specificity vs. sensitivity curves, which can be transformed into ROC curves (sens. vs. 1 – spec.) [Results found at SimpleWindow35/sw35_8/SensSpec/sw35_8_700_test_norm_exc0t_prof_snsp_htx30_all_1.txt]

```
#include "simple_window35.hpp"
#include "sens_spec_client.hpp"
```

```
int main()
{ try
  {     using namespace SimpleWindow35;
        typedef SequenceProcessorTuple<TYPELIST_2(default_preprocessor_tp,
                DomainDivisionNeutralizer<9>) > preprocessor_tp;
        SensSpecClient<scorer_tp, SCOPDataset, preprocessor_tp, default_seq_reader_tp> client;

        client.dataset.preprocessor.set_num_terminal(0);
        client.dataset.preprocessor.set_min_his_tag_terminus_noncounted_length(30);
        client.set_dataset_base_name("scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1_prof");
        client.set_primary_dir(default_primary_dir);
        client.set_run_base_name("sw35_8");
        client.set_score_type(FileNameHandler::Normalized);
        client.set_trial_nums(1, 1);
        client.set_step_num(700);
        client.set_report_token("exc0t_prof_snsp_htx30");
        client.gather_results();
  } CATCHES
}
```

sw35_8a_combine_train_score_output.pl: Obtain average begin-of-cycle scores for each optimization step from record files produced during optimization [Results found at SimpleWindow35/sw35_8a/RecordFiles/sw35_8a_train_scr_summ.txt]

```
#!/usr/bin/perl -w

use strict;

my $run_base_name = "sw35_8a";
my $fileprefix = "SimpleWindow35/$run_base_name/RecordFiles/$run_base_name" . "_train";
my $filesuffix = ".rec2";
my @line;
my $c;

sub infilename
{       my $filename;

        $filename = $fileprefix . $_[0] . $filesuffix;
        return $filename;
}

open(INFILE1, infilename(1))
        || die "Cannot open input file";
open(INFILE2, infilename(2));
open(INFILE3, infilename(3));
open(INFILE4, infilename(4));
open(INFILE5, infilename(5));

my $outfilename = $fileprefix . "_scr_summ.txt";
print "Writing to $outfilename.\n";
open(OUTFILE, "> $outfilename");

print OUTFILE "\t1_1\t1_2\t2_1\t2_2\t3_1\t3_2\t4_1\t4_2\t5_1\t5_2\n";
while(<INFILE1>)
{       chomp;
        my @split_line = split(/\t/, $_);
        print OUTFILE $split_line[0];
        output_line($_);
        $_ = <INFILE2>;
        output_line($_);
        $_ = <INFILE3>;
        output_line($_);
        $_ = <INFILE4>;
        output_line($_);
        $_ = <INFILE5>;
        output_line($_);
        print OUTFILE "\n";
```

```
}

sub output_line
{       chomp($_);
        my @split_line = split(/\t/, $_[0]);
        print OUTFILE "\t";
        print OUTFILE $split_line[1];
        print OUTFILE "\t";
        print OUTFILE $split_line[2];
}
```

D.4 CODE RELATED TO DATASETS

Table D.4-1. Description of some (not all) dataset-related code files. (Actual code may be made available at http://prodata.swmed.edu.)

get_scop1.67_chain_names.cpp	Gets set of unique chain names (4-character PDB ID with lower case letters, plus chain identifier character) for chains containing domains from SCOP classes A – H.
make_seqfiles6files.cpp	Processing PDB data—obtaining files with information on chains—listing the type and status (Gap, Nongap, etc.) of each individual residue, the sequence being based on the sequence provided in SEQRES entries in PDB files. Writes successfully obtained 'sequence files' to SeqFiles6/.
produce_scop1.67_fam_all.cpp	Obtains initial set of domains, grouped into families, with revised domain boundaries, in XML form.
find_scop1.67_mult_chain_domain_problems.cpp	Helps find domains that are located in multiple chains and need to be manually revised.
produce_scop1.67_fam_alpha_beta.cpp	Obtains the subset of grouped families of domains that are in the first five families of SCOP classes.
produce_scop1.67_fam_alpha_beta_2000_3.0_min50.cpp	Obtains subset of domains in first five SCOP classes that are dated 2000 or later and have a resolution no worse than 3.0.
produce_scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1.cpp	Divides families into cross-validation sets
produce_scop1.67_fam_alpha_beta_2000_3.0_min50_5xv1_env.cpp	Produces standardized testing data 'subsets' (same domain can represent a family in the 'subset' multiple times).

BIBLIOGRAPHY

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. Science 181: 223-230.

Berman, H.M., Bhat, T.N., Bourne, P.E., Feng, Z., Gilliland, G., Weissig, H., and Westbrook, J. 2000. The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol 7 Suppl: 957-959.

Boissel, J.P., Kasper, T.J., and Bunn, H.F. 1988. Cotranslational amino-terminal processing of cytosolic proteins. Cell-free expression of site-directed mutants of human hemoglobin. J Biol Chem 263: 8443-8449.

Cai, W., and Shao, X. 2002. A fast annealing evolutionary algorithm for global optimization. J Comput Chem 23: 427-435.

Chothia, C. 1974. Hydrophobic bonding and accessible surface area in proteins. Nature 248: 338-339.

Chou, P.Y., and Fasman, G.D. 1974. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 13: 211-222.

Coeytaux, K., and Poupon, A. 2005. Prediction of unfolded segments in a protein sequence based on amino acid composition. Bioinformatics 21: 1891-1900.

Deleage, G., and Roux, B. 1987. An algorithm for protein secondary structure prediction based on class prediction. Protein Eng 1: 289-294.

Denning, D.P., Patel, S.S., Uversky, V., Fink, A.L., and Rexach, M. 2003. Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. Proc Natl Acad Sci U S A 100: 2450-2455.

Derewenda, Z.S. 2004. Rational protein crystallization by mutational surface engineering. Structure (Camb) 12: 529-535.

Dominguez-Vidal, A., Saenz-Navajas, M.P., Ayora-Canada, M.J., and Lendl, B. 2006. Detection of albumin unfolding preceding proteolysis using fourier transform infrared spectroscopy and chemometric data analysis. Anal Chem 78: 3257-3264.

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 347: 827-839.

Drenth, J. 1999. Principles of Protein X-Ray Crystallography, 2nd ed. Springer-Verlag, New York.

Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M., and Obradovic, Z. 2002. Intrinsic disorder and protein function. Biochemistry 41: 6573-6582.

Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., et al. 2001. Intrinsically disordered protein. J Mol Graph Model 19: 26-59.

Esnouf, R.M., Hamer, R., Sussman, J.L., Silman, I., Trudgian, D., Yang, Z.R., and Prilusky, J. 2006. Honing the in silico toolkit for detecting protein disorder. Acta Crystallogr D Biol Crystallogr 62: 1260-1266.

Franks, N.P., Abraham, M.H., and Lieb, W.R. 1993. Molecular organization of liquid n-octanol: an X-ray diffraction analysis. J Pharm Sci 82: 466-470.

Gribskov, M., and Robinson, N.L. 1996. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput Chem 20: 25-33.

Guy, H.R. 1985. Amino acid side-chain partition energies and distribution of residues in soluble proteins. Biophys J 47: 61-70.

Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89: 10915-10919.

Hopp, T.P., and Woods, K.R. 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci U S A 78: 3824-3828.

Hopp, T.P., and Woods, K.R. 1983. A computer program for predicting protein antigenic determinants. Mol Immunol 20: 483-489.

Huang, S., Elliott, R.C., Liu, P.S., Koduri, R.K., Weickmann, J.L., Lee, J.H., Blair, L.C., Ghosh-Dastidar, P., Bradshaw, R.A., Bryan, K.M., et al. 1987. Specificity of cotranslational amino-terminal processing of proteins in yeast. Biochemistry 26: 8242-8246.

Hubbard, S.J., Eisenmenger, F., and Thornton, J.M. 1994. Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. Protein Sci 3: 757-768.

Huber, R. 1979. Conformational flexibility and its functional significance in some protein molecules. Trends Biochem Sci 4: 271-276.

Huber, R., and Bennett, W.S., Jr. 1983. Functional significance of flexibility in proteins. Biopolymers 22: 261-279.

Hurle, M.R., Helms, L.R., Li, L., Chan, W., and Wetzel, R. 1994. A role for destabilizing amino acid replacements in light-chain amyloidosis. Proc Natl Acad Sci U S A 91: 5446-5450.

Iakoucheva, L.M., Brown, C.J., Lawson, J.D., Obradovic, Z., and Dunker, A.K. 2002. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 323: 573-584.

Ingber, L. 1989. Very fast simulated re-annealing. Math Comp Model 12: 967-973.

Jacobs, D.M., Lipton, A.S., Isern, N.G., Daughdrill, G.W., Lowry, D.F., Gomes, X., and Wold, M.S. 1999. Human replication protein A: global fold of the N-terminal RPA-70 domain reveals a basic cleft and flexible C-terminal linker. J Biomol NMR 14: 321-331.

Jones, D.T., and Ward, J.J. 2003. Prediction of disordered regions in proteins from position specific score matrices. Proteins 53 Suppl 6: 573-578.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22: 2577–2637.

Kawashima, S., Ogata, H., and Kanehisa, M. 1999. AAindex: Amino Acid Index Database. Nucleic Acids Res 27: 368-369.

Khare, S., Ding, F, Gwanmesia, KN, Dokholyan, NV. 2005. Molecular origin of polyglutamine aggregation in neurodegenerative diseases. PLOS Comp Biol 1: 230-235.

Kyte, J., and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105-132.

Lee, B., and Richards, F.M. 1971. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55: 379-400.

Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104: 59-107.

Li, X., Romero, P., Rani, M., Dunker, A.K., and Obradovic, Z. 1999. Predicting Protein Disorder for N-, C-, and Internal Regions. Genome Inform Ser Workshop Genome Inform 10: 30-40.

Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. 2003a. Protein disorder prediction: implications for structural proteomics. Structure (Camb) 11: 1453-1459.

Linding, R., Russell, R.B., Neduva, V., and Gibson, T.J. 2003b. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* **31:** 3701-3708.

Lise, S., and Jones, D.T. 2005. Sequence patterns associated with disordered regions in proteins. *Proteins* **58:** 144-150.

MacCallum, R. 2004. Order/disorder prediction with self organising maps. *FORCASP*.

Meek, J.L. 1980. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc Natl Acad Sci U S A* **77:** 1632-1636.

Meek, J.L., and Rossetti, Z.L. 1981. Factors affecting retention and resolution of peptides in high-performance liquid chromatography. *J Chromatogr* **211:** 15-28.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. 1953. Equation of state calculations by fast computing machines. *J Chem Phys* **21:** 1087-1092.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247:** 536-540.

Nishikawa, K., and Ooi, T. 1986. Radial locations of amino acid residues in a globular protein: correlation with the sequence. *J Biochem (Tokyo)* **100:** 1043-1047.

Nozaki, Y., and Tanford, C. 1971. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J Biol Chem* **246:** 2211-2217.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C.J., and Dunker, A.K. 2003. Predicting intrinsic disorder from amino acid sequence. *Proteins* **53 Suppl 6:** 566-572.

Pappu, R.V., and Rose, G.D. 2002. A simple model for polyproline II structure in unfolded states of alanine-based peptides. *Protein Sci* **11:** 2437-2455.

Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K., and Obradovic, Z. 2006. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7:** 208.

Peng, K., Vucetic, S., Radivojac, P., Brown, C.J., Dunker, A.K., and Obradovic, Z. 2005. Optimizing long intrinsic disorder predictors with protein evolutionary information. *J Bioinform Comput Biol* **3:** 35-60.

Press, W., Teukolsky, SA, Vetterling, WT, Flannery, BP. 1999. *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, New York.

Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., and Sussman, J.L. 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21:** 3435-3438.

Punta, M., and Maritan, A. 2003. A knowledge-based scale for amino acid membrane propensity. *Proteins* **50:** 114-121.

Radzicka, A., and Wolfenden, R. 1988. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **27:** 1664-1670.

Raffen, R., Dieckman, L.J., Szpunar, M., Wunschl, C., Pokkuluri, P.R., Dave, P., Wilkins Stevens, P., Cai, X., Schiffer, M., and Stevens, F.J. 1999. Physicochemical consequences of amino acid variations that contribute to fibril formation by immunoglobulin light chains. *Protein Sci* **8:** 509-517.

Romero, Obradovic, and Dunker, K. 1997a. Sequence Data Analysis for Long Disordered Regions Prediction in the Calcineurin Family. *Genome Inform Ser Workshop Genome Inform* **8:** 110-124.

Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J.E., and Dunker, A.K. 1997b. Identifying disordered regions in proteins from amino acid sequence. *Proc IEEE International Conference on Neural Networks* **1:** 90-95.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. 2001. Sequence complexity of disordered protein. *Proteins* **42:** 38-48.

Rose, G.D., Fleming, P.J., Banavar, J.R., and Maritan, A. 2006. A backbone-based theory of protein folding. *Proc Natl Acad Sci U S A* **103:** 16623-16633.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., and Zehfus, M.H. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science* **229:** 834-838.

Rost, B., and Sander, C. 2000. Third generation prediction of secondary structures. *Methods Mol Biol* **143:** 71-95.

Sadreyev, R., and Grishin, N. 2003. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* **326:** 317-336.

Sambashivan, S., Yanshun, L, Sawaya, MR, Gingery, M, Eisenberg, D. 2005. Amyloid-like fibrils of ribonuclease A with three-dimensional domain-swapped and native-like structure. *Nature* **437:** 266-269.

Savitzky, A., Golay, MJE. 1964. Smoothing and differentiation of data by simplified least squares procedure. *Analyt Chem* **36:** 1627-1639.

Sunyaev, S.R., Eisenhaber, F., Rodchenkov, I.V., Eisenhaber, B., Tumanyan, V.G., and Kuznetsov, E.N. 1999. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng* **12:** 387-394.

Sweet, R.M., and Eisenberg, D. 1983. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J Mol Biol* **171:** 479-488.

Tatusov, R.L., Altschul, S.F., and Koonin, E.V. 1994. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci U S A* **91:** 12091-12095.

Tompa, P. 2002. Intrinsically unstructured proteins. *Trends Biochem Sci* **27:** 527-533.

Uversky, V.N. 2002. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* **11:** 739-756.

Uversky, V.N., Gillespie, J.R., and Fink, A.L. 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41:** 415-427.

Venanzi, T.J. 1984. Hydrophobicity parameters and the bitter taste of L-amino acids. *J Theor Biol* **111:** 447-450.

Vucetic, S., Brown, C.J., Dunker, A.K., and Obradovic, Z. 2003. Flavors of protein disorder. *Proteins* **52:** 573-584.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337:** 635-645.

Weathers, E.A., Paulaitis, M.E., Woolf, T.B., and Hoh, J.H. 2004. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. *FEBS Lett* **576:** 348-352.

Wehrens, R., and Buydens, L.M.C. 1998. Evolutionary optimisation: a tutorial. *Trends Analyt Chem* **17:** 193-203.

Wei, G., Liu, G., and Liu, X. 2003. Identification of two serine residues important for p53 DNA binding and protein stability. *FEBS Lett* **543:** 16-20.

Wertz, D.H., and Scheraga, H.A. 1978. Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules* **11:** 9-15

Williams, R.M., Obradovic, Z., Mathura, V., Braun, W., Garner, E.C., Young, J., Takayama, S., Brown, C.J., and Dunker, A.K. 2001. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. *Pac Symp Biocomput* **6:** 89-100.

Wimley, W.C., Creamer, T.P., and White, S.H. 1996. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry* **35:** 5109-5124.

Wimley, W.C., and White, S.H. 1996. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat Struct Biol* **3:** 842-848.

Wright, P.E., and Dyson, H.J. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* **293:** 321-331.

Yang, Z.R., Thomson, R., McNeil, P., and Esnouf, R.M. 2005. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* **21:** 3369-3376.

Zdanov, A., Li, Y., Bundle, D.R., Deng, S.J., MacKenzie, C.R., Narang, S.A., Young, N.M., and Cygler, M. 1994. Structure of a single-chain antibody variable domain (Fv) fragment complexed with a carbohydrate antigen at 1.7-A resolution. *Proc Natl Acad Sci U S A* **91:** 6423-6427.

VITAE

Nathan Brent Holladay was born in Provo, Utah on March 30, 1977, the son of Brent Richins Holladay and Dana Vorwaller Holladay. After graduating from Lake Mary High School in 1995 Lake Mary, FL, he entered Brigham Young University in Provo, Utah. From 1996 to 1998 he served a full time mission for the Church of Jesus Christ of Latter-day Saints and then returned to Brigham Young University. Undergraduate courses were completed in 2000 and requirements for University Honors were completed in 2001, when he officially received his Bachelor of Science in Biochemistry. In 2000 he moved to Dallas, TX, where he joined the Medical Scientist Training Program at the University of Texas Southwestern Medical Center. He joined the Molecular Biophysics Program and performed research for his Doctorate of Philosophy in the laboratory of Nick Grishin, Ph.D. In 1999, he married Marcie Hofmann (Holladay), and they have been blessed with four children since that time.

Permanent (parents') address: 820 Eastgate Trail
Longwood, FL 32750