Protein Solubility Prediction

Reese Lennarson

Rex Richard

Project Relevance

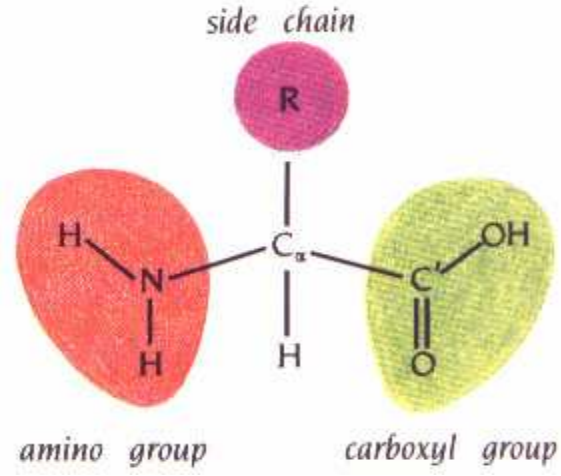
- Recombinant DNA Technology: Insert gene of protein of interest into Escherichia coli accessory DNA
- E. coli uses these new instructions from new DNA and becomes a reactor for the production of the protein of interest
- Proteins not native to E. coli may be soluble or insoluble when expressed
- Insoluble proteins form pellets that are difficult to recover and are not desired in production
- Accurate predictions can save time performing experiments

Project Objectives

- Develop models that can predict whether a protein will be soluble or insoluble when expressed in Escherichia coli based on trends in parameters for collected proteins
- Evaluate different methods for prediction and see which is best
- Identify most important parameters for accurate prediction of solubility

Protein Background: Amino Acids

- Proteins composed of building blocks called amino acids

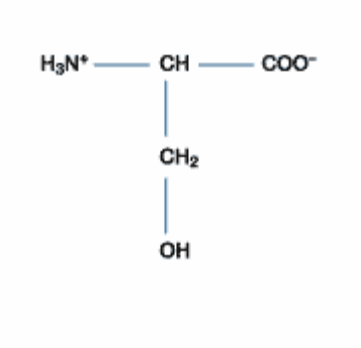


- R groups responsible for protein folding and ultimately function

- 20 amino acids each with different R group

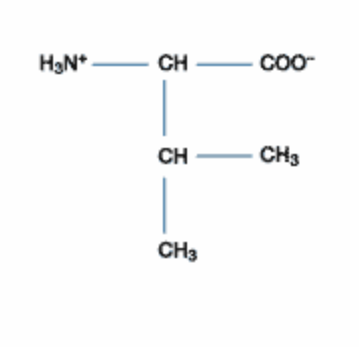
Protein Background: Amino Acids (cont'd)

- R groups characterized by H-bond character, charge, size, shape, hydrophobicity



Serine (hydrophilic)

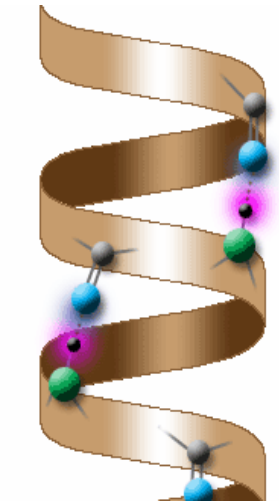
Valine (hydrophobic)



- Sequence of amino acid's R groups (primary structure) determines how protein folds

Protein Background: Secondary Structure

- Secondary structure (local 3-D structure) has three common motifs: α-helix, β-sheet, and turns
- Alpha helix forms stabilizing H-bonds along adjacent coil strands



Alpha helix

B sheet

- Secondary structure can be predicted fairly well with knowledge of amino acid sequence

Creating a Protein Database

- 226 proteins found in research for which solubility status on expression in E. coli is known at set conditions (37 C, no chaperones or fusion partners)
- Amino acid sequences catalogued for each found protein
- 17 parameters based on amino acid sequence and hypothesized to affect solubility calculated for each protein

Protein Parameters

Parameters based on fraction of specific amino acids:

- cysteine fraction
- proline fraction
- asparagine fraction
- threonine fraction
- tyrosine fraction
- combined fraction of asn, thr, and tyr

Parameters based on protein-solvent interaction:

- hydrophilicity index
- hydrophobic residue fraction
- average number of contiguous hydrophobic residues
- aliphatic index
- approximate charge average

Protein Parameters (cont'd)

Parameters based on secondary structure:

- alpha helix propensity
- beta sheet propensity
- alpha helix propensity/beta sheet propensity
- turn-forming residue fraction

Parameters based on protein size:

- molecular weight, total number of residues

Developing a Model that Can Predict Solubility

- Three methods used for prediction: discriminant analysis, logistic regression, and neural network
- Models look for parameter trends from protein to protein in the database
- Each model develops an equation to predict solubility for new proteins

Statistical Analyses

- Discriminant Analysis (DA)
  - Used in all previous solubility studies
- Logistic Regression (LR)
  - More commonly used than discriminant analysis in recent years

SAS (Statistical Analysis System) software used to build models for both methods

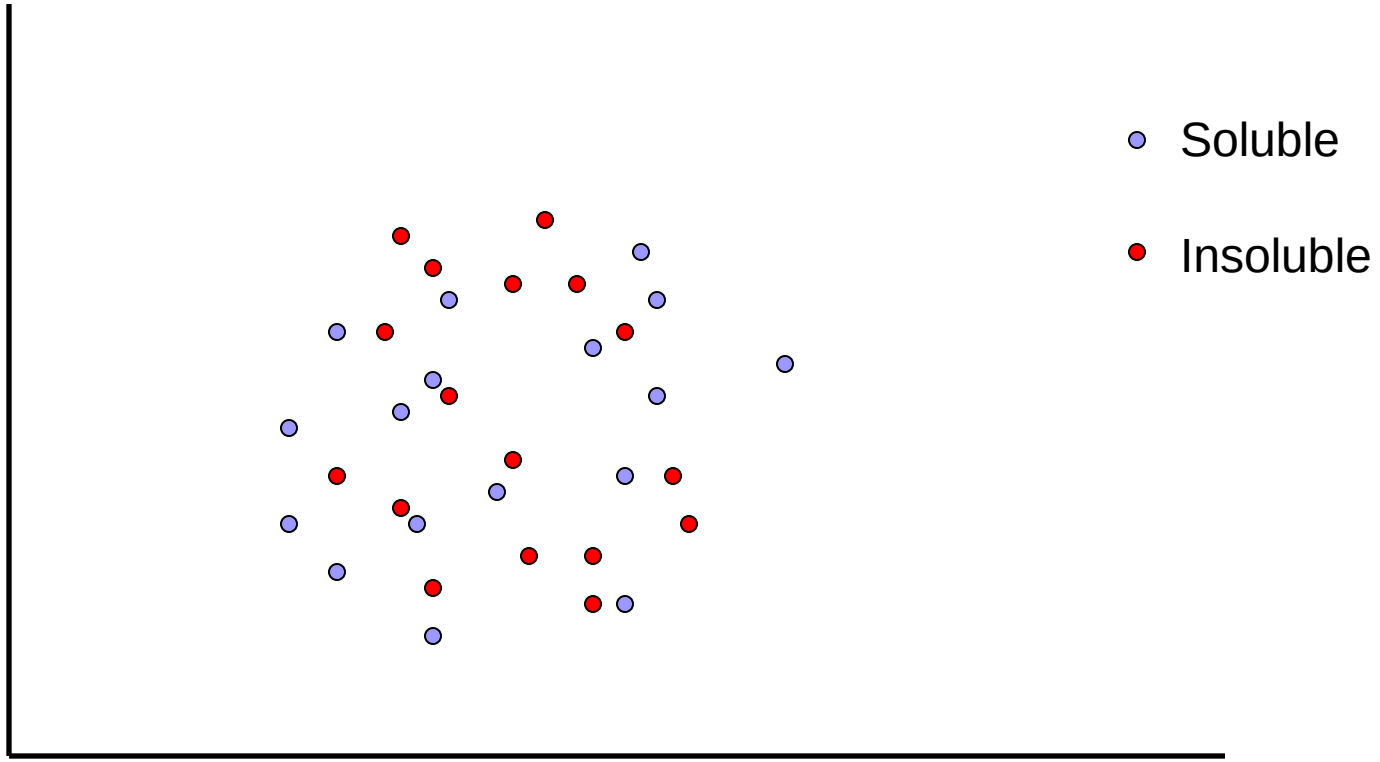
Why investigate logistic regression?

LR fits our system better than DA!

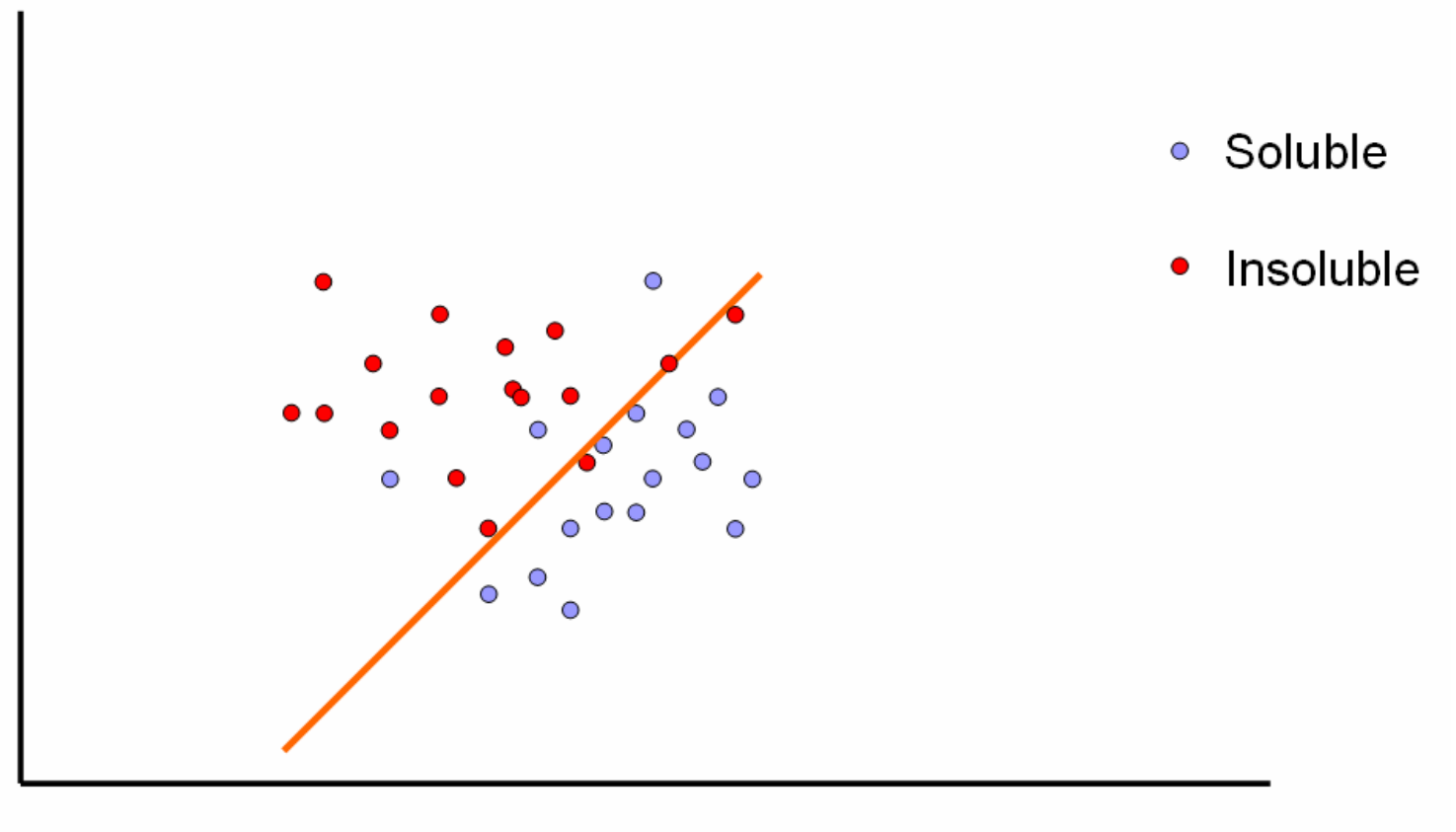
- LR more accurate when there are only 2 dichotomous groups in the dependent variable
- LR more accurate than DA when independent (input) variables are continuous
- DA must assume normal distribution of independent variables
- LR handles unequal group sizes better than DA

→LR can give us a more robust model to make future solubility predictions.

2-D Representation of Statistical Models



2-D Representation of Statistical Models



Discriminant Analysis

- Used to model systems with categorical, rather than continuous, dependent (outcome) variables

- Calculates canonical variable (CV) from parameters for each data point

$$CV = \sum^{n} \lambda_i x_i$$

n = number of parameters

x_i = value of parameter i

λ_i = adjustable coefficient of parameter i

Discriminant Analysis, continued

$$CV = \sum^{n} \lambda_i x_i$$

- DA optimizes λ values to achieve maximum distinction between groups
- Value of discriminant found
- Discriminant is the dividing line between groups for prediction of new data

CV > discriminant; → data belongs to Group 1
CV < discriminant; → data belongs to Group 2

Logistic Regression

Similar in approach to DA, but it transforms the dependent variable via a logit function

$$\log\left[\frac{p_i}{1-p_i}\right] = \alpha + \sum^{n} \beta_i x_i$$

where p_i = probability that data belongs to group 1 (soluble)

and $\log\left[\frac{p_i}{1-p_i}\right]$ = "logit" or "log-odds"

- Maximum likelihood method used to determine α and β values

- $p_i \geq 0.5$ Soluble
- $p_i < 0.5$ Insoluble

Building a DA model in SAS

Step 1: Significant parameters determined in with STEPDISC statement

- Stepwise construction of model
- Parameters evaluated one by one (F to enter, F to remove)
- Parameters with lowest p_r > F value (null-hypothesis test) included in model
- Remaining parameters reevaluated; additional parameters included as necessary
- Parameters may be excluded from the model at any step if F > p value rises above 0.05 (95% confidence)
- Process continues until no more parameters can be added to or removed from model

Building a DA model in SAS

Statistics for Entry, DF = 1, 223

Variable	Partial R-Square	F Value	Pr > F	Tolerance
TotRes	0.0047	1.06	0.3040	0.9999
Mwkda	0.0059	1.31	0.2530	1.0000
CysFrac	0.0023	0.50	0.4784	0.8310
ProFrac	0.0002	0.04	0.8514	0.9476
TurnFrac	0.0000	0.00	0.9880	0.7637
Hphil	0.0008	0.18	0.6736	0.9089
AppChgAvg	0.0040	0.89	0.3457	0.9947
TotHydRes	0.0014	0.31	0.5754	0.9959
ContHydRes	0.0025	0.56	0.4547	0.9986
Aliphatic	0.0010	0.23	0.6339	0.9968
BSheetProp	0.0024	0.54	0.4614	0.9966
ABRatio	0.0023	0.50	0.4783	0.4204
AsnFrac	0.0361	8.34	0.0043	0.9261
ThrFrac	0.0000	0.00	0.9459	0.9178
TyrFrac	0.0000	0.01	0.9410	0.8704
AsnThrTyrFrac	0.0137	3.10	0.0799	0.7476

Variable AsnFrac will be entered.

Variable(s) that have been Entered

AHelixProp AsnFrac

Building a DA model in SAS

Step 2: Coefficients determined with CANDISC statement

- Provides raw and weighted coefficients for parameters

Step 3: Model evaluated with DISCRIM statement

- Provides accuracy of predictions for insoluble proteins, soluble proteins, and overall database

Building a LR Model in SAS

- Model built in reverse-stepwise fashion
- All parameters included at first, run with LOGISTIC statement
- Parameter with highest null-hypothesis probability removed
- Model run again, next parameter deleted
- Process continues until remaining parameters have null-hypothesis probability ≤ 0.05 (95% confidence)
- Intercept (α) and coefficient estimates (β) generated as output

Building a LR Model in SAS

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	72.7053	96.6155	0.5663	0.4517
TotRes	1	-0.00007	0.0334	0.0000	0.9982
Mwkda	1	-0.1369	0.2233	0.3757	0.5399
CysFrac	1	-21.0924	11.6130	3.2989	0.0693
ProFrac	1	-5.6537	12.2846	0.2118	0.6454
TurnFrac	1	-4.5721	5.0099	0.8328	0.3615
Hphil	1	3.4453	1.8176	3.5929	0.0580
AppChgAvg	1	-11.3526	5.2969	4.5935	0.0321
TotHydRes	1	0.0454	0.0220	4.2656	0.0389
ContHydRes	1	-0.1560	0.4043	0.1490	0.6995
Aliphatic	1	-0.0145	0.0589	0.0608	0.8052
AHelixProp	1	54.9303	93.2530	0.3470	0.5558
BSheetProp	1	-65.0460	94.9835	0.4690	0.4935
ABRatio	1	-59.2664	93.2351	0.4041	0.5250
AsnFrac	1	-23.9408	11.6146	4.2488	0.0393
ThrFrac	1	-7.5168	10.7525	0.4887	0.4845
TyrFrac	1	7.8597	10.1721	0.5970	0.4397
AsnThrTyrFrac	0	0	.	.	.

Evaluating the Models

- *Post hoc* (training set) evaluations
  - All proteins used to build model
  - Same proteins plugged into model
  - Model solubility predictions compared to actual solubility of proteins
  - Result reported as percentage accuracy

- *A priori* (test set) evaluations
  - Some proteins used to build model
  - Remaining proteins plugged into model
  - Provides more realistic evaluation of how well models will predict solubility for new proteins

Discriminant Analysis Results

- Important parameters:
  - Previous research:
    - Wilkinson-Harrison: charge average, turn-forming residue fraction
    - Idicula-Thomas: aliphatic index, molecular weight, net charge
  - Current work:
    - Asparagine fraction, α-helix propensity

Discriminant Analysis Results

- Parameter Coefficients:

Parameter	Standardized Coefficient	Raw Coefficient
α-helix Propensity	0.68	18.12
Asparagine Fraction	-0.64	-31.02

- *Post hoc* accuracy:

Soluble	Insoluble	Overall
70.7%	62.3%	66.5%

Logistic Regression Results

Removal of parameters from model:

Parameter	p_r in Removal Step
Total Number of Residues	0.858
αβ Propensity Ratio	0.839
Aliphatic Index	0.810
β-sheet Propensity	0.794
Average # of Contiguous Hydrophobic Residues	0.692
Proline Fraction	0.653
Threonine Fraction	0.628
Combined Asn, Tyr, Thr Fraction	0.628
Turn-Forming Residue Fraction	0.416
α-helix Propensity	0.398
Cysteine Fraction	0.155

Logistic Regression Results

- Parameters included in model:

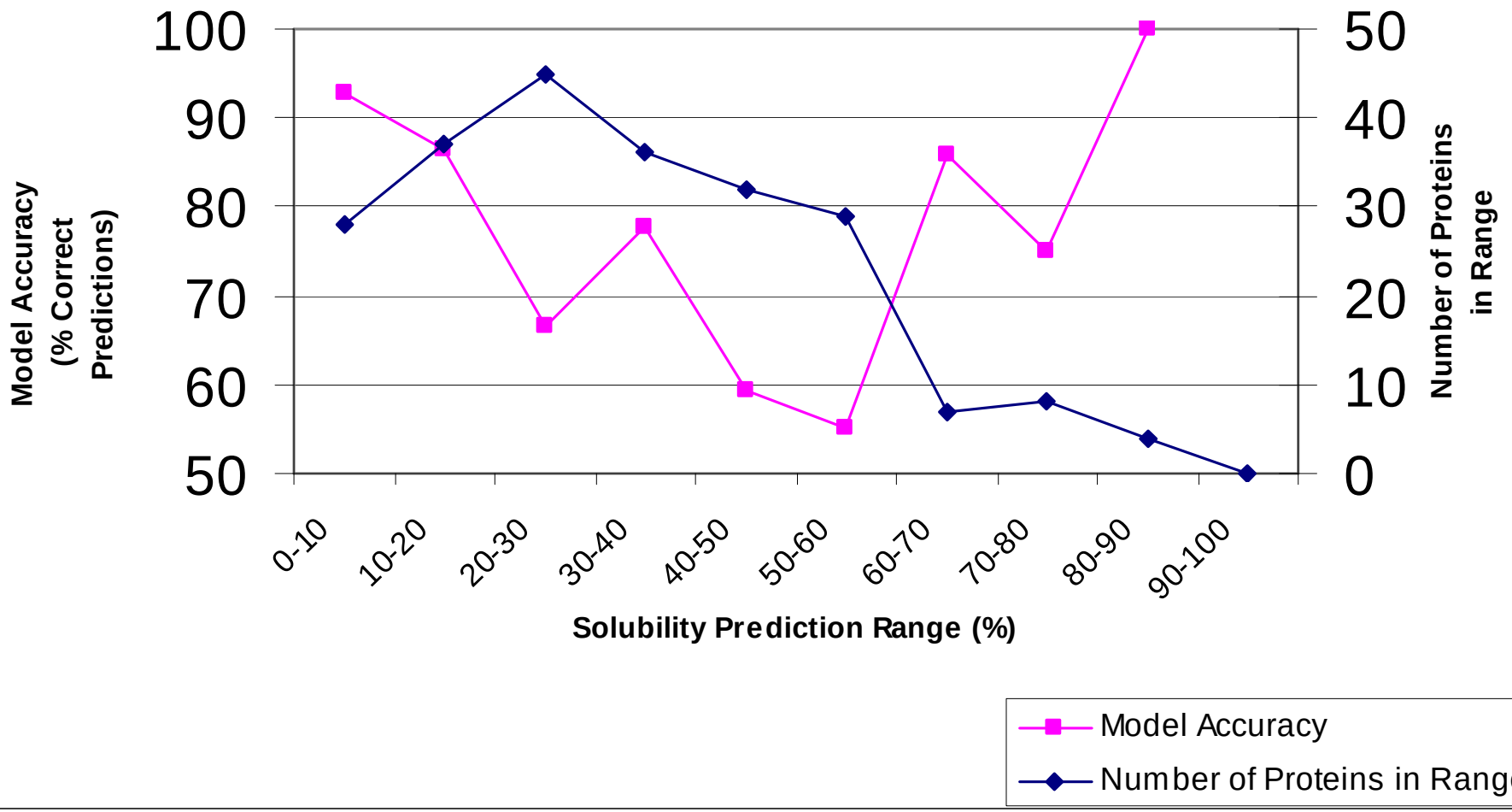
Parameter	p_r	Relative Weight	Estimated Coefficient
Molecular Weight (kDa)	<0.0001	1.00	-0.1693
Total # of Hydrophobic Residues	<0.0001	0.95	0.0600
Hydrophilicity Index	0.0002	0.02	4.9629
Approximate Charge Average	0.0192	0.05	-12.3538
Asparagine Fraction	0.0325	0.11	-20.4259
Tyrosine Fraction	0.0511	0.07	15.1898

- *Post hoc* accuracy

Soluble	Insoluble	Overall
42.7%	89.4%	73.9%

Logistic Regression Model Accuracy over Prediction Ranges

(*Post hoc* analysis of entire database)



LR *A Priori* Analysis

- Database randomized eight times
- Data split into training and test sets of the following ratios:
  - 80/20
  - 85/15
  - 90/10
  - 95/5
- For each ratio, accuracies using the eight randomized data sets were averaged

Logistic Regression Results

Accuracy averages for test sets:

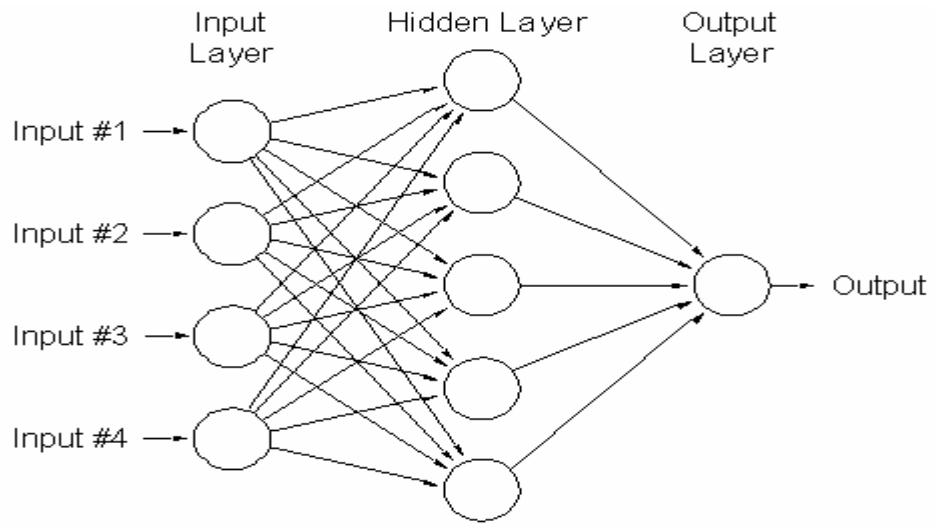
	Training-Set Accuracy (%)			**Test-Set Accuracy (%)**		
Test-Set Size (percent of overall database)	*Soluble*	*Insoluble*	*Overall*	*Soluble*	*Insoluble*	*Overall*
5%	43.7	87.1	72.4	25.3	100.0	88.6
10%	45.2	88.1	74.3	17.0	98.5	78.7
15%	47.2	86.7	73.1	19.5	98.5	78.7
20%	45.9	87.1	72.9	21.7	98.1	76.1

Statistical Analysis Summary

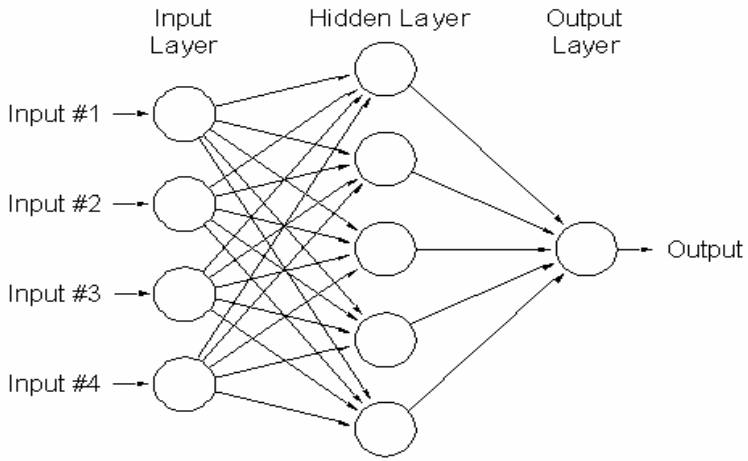
- Discriminant analysis models overpredict solubility
- Logistic regression models overpredict insolubility
- LR models demonstrate better *post hoc* accuracy than DA models
- LR models very accurate (>90%) for solubility probabilities nearing 0% and 100%

Neural Network (NN) Theory

- Neural networks essentially learn by decreasing error through iterations
- For this project, a feedforward network is used with backpropagation
- The most common neural network consists of one input layer, one hidden layer, and one output layer with two connection layers

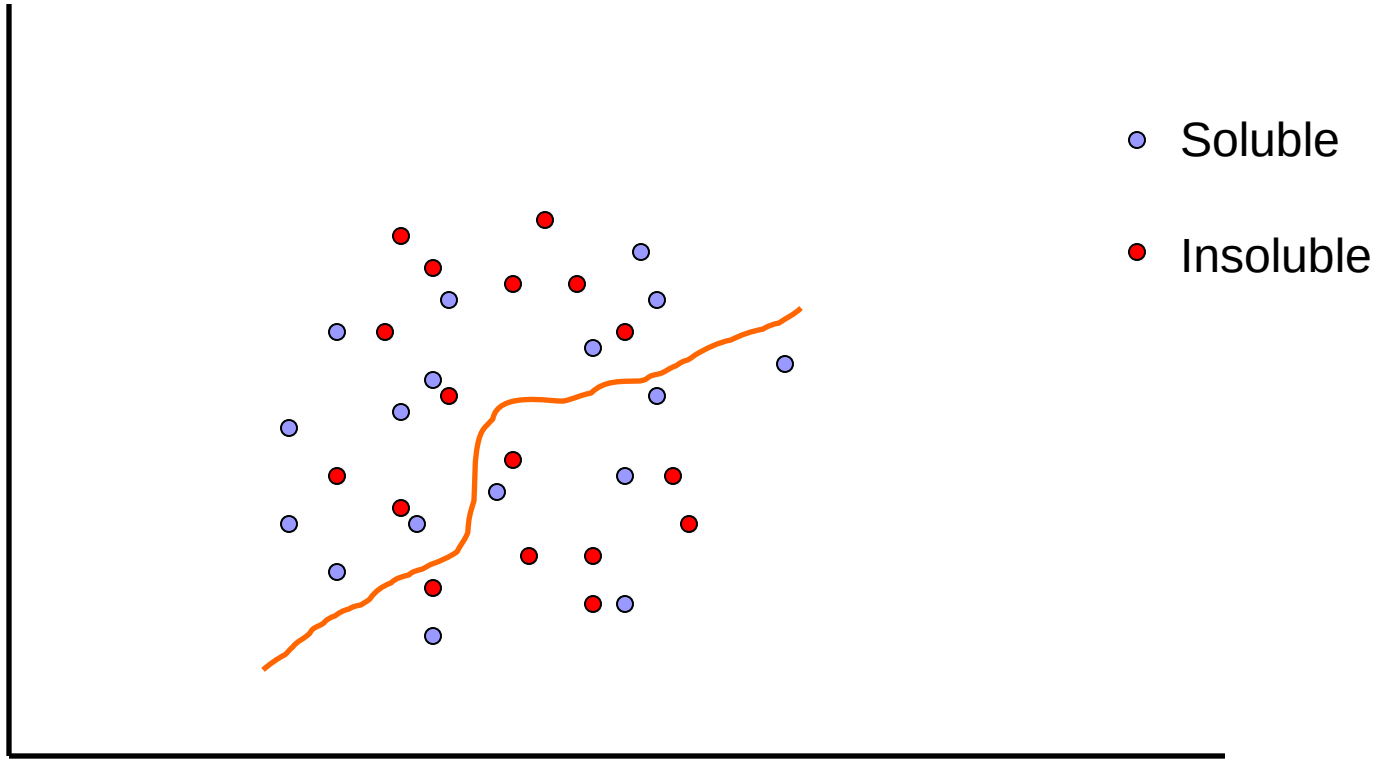


Feedforward NNs with Backpropagation

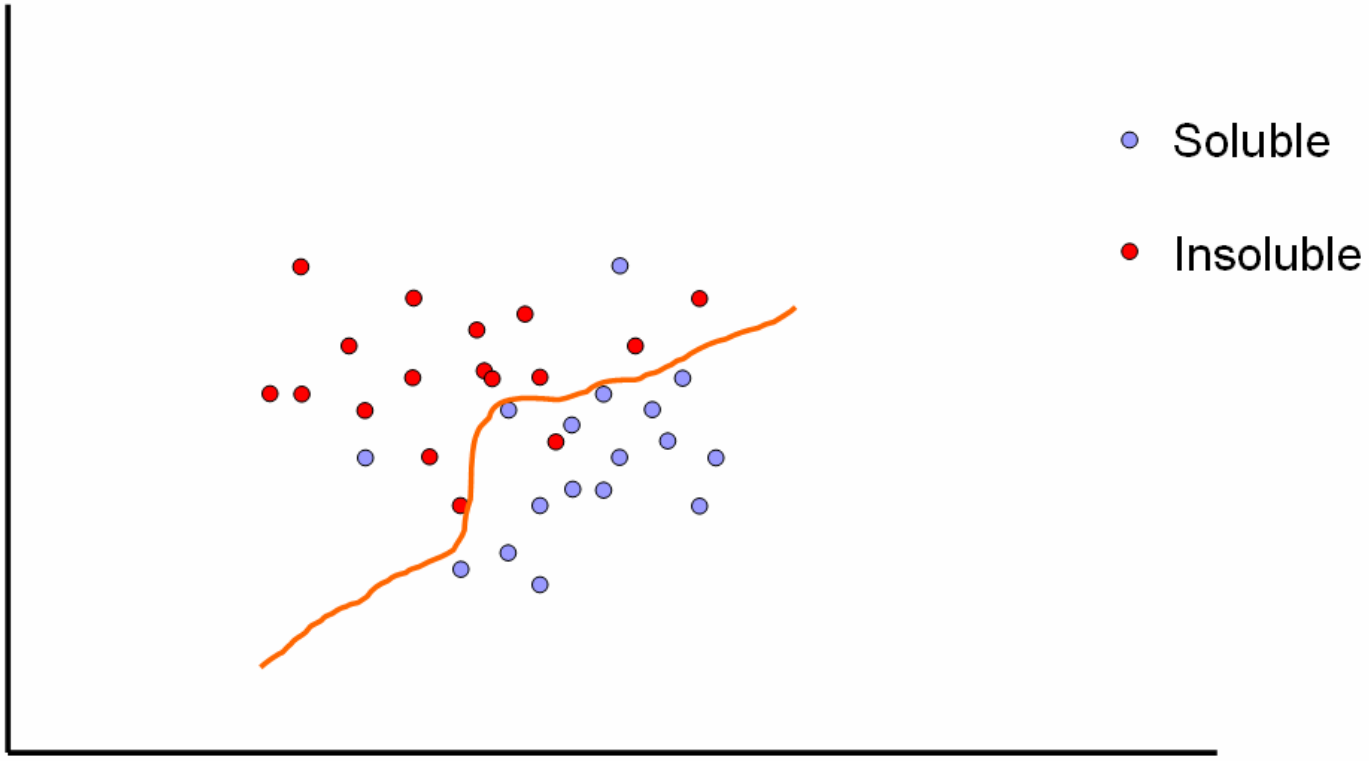


- Data flows from input layer to hidden layer to output layer for each iteration (epoch); output given as value between 0 and 1, where number higher than 0.5 rounded up, numbers lower than 0.5 rounded down
- Error signal is calculated and sent back to first connection layer to update weights for next iteration
- Each connection layer supplies weights (initially randomized) from input layer to hidden layer and from hidden layer to output layer
- All data is normalized

2-D Representation of Neural Network Models



2-D Representation of Neural Network Models



Neural Network Data Analysis:Training/Test Set Randomization

- First, eight randomized training/test set combinations with each in the ratio of 80%/20% were made

- The randomized training/test set combo with the highest test set accuracy was chosen for the optimization of number of nodes

- Set Parameters

  Number of nodes: 4

  Number of iterations: 25,000

  Hidden Layer Step Size: 0.5

  Output Layer Step Size:0.05

Training/Test Set Randomization Results

	Training Accuracy(%)			Test Accuracy(%)		
Random Set	Soluble	Insoluble	Overall	Soluble	Insoluble	Overall
1	67	97	86	78	89	87
2	97	94	95	50	90	78
3	82	98	93	29	65	53
4	90	98	95	29	77	62
5	84	98	95	32	54	38
6	82	97	92	46	81	71
7	80	93	88	40	63	58
8	80	98	92	47	60	56

Neural Network Data Analysis: Node Optimization

- Number of nodes varied from 3 to 9 using optimum training/test combo from before
- Number of iterations and step sizes kept same as before
- Number of nodes giving highest test set accuracy considered optimum

Node Optimization Results

	Training Accuracy(%)			Test Accuracy(%)		
Number of Nodes	Soluble	Insoluble	Overall	Soluble	Insoluble	Overall
3	84	91	89	65	84	78
4	67	97	86	78	89	87
5	83	96	91	55	84	74
6	95	98	97	60	82	74
7	94	99	97	60	79	72
8	95	99	98	60	76	71
9	94	99	97	50	74	66

Neural Network Model Using All Proteins

- Final model included all 226 proteins giving the following training accuracy.

Training Accuracy (%)		
Soluble	Insoluble	Overall
80	96	91

- Almost 90% of outputs in this analysis fell in the ranges of 0-0.1 and 0.9-1

- Can we get a better idea of what kind of accuracy one can expect when this model is used on new proteins?

Neural Network Data Analysis: Varying the Training Set Size

- Same procedure used for logistic regression
- Seven new randomized training/test set combos added to the one used in node optimization
- This was done for 80/20, 85/15, 90/10, and 95/5 ratios

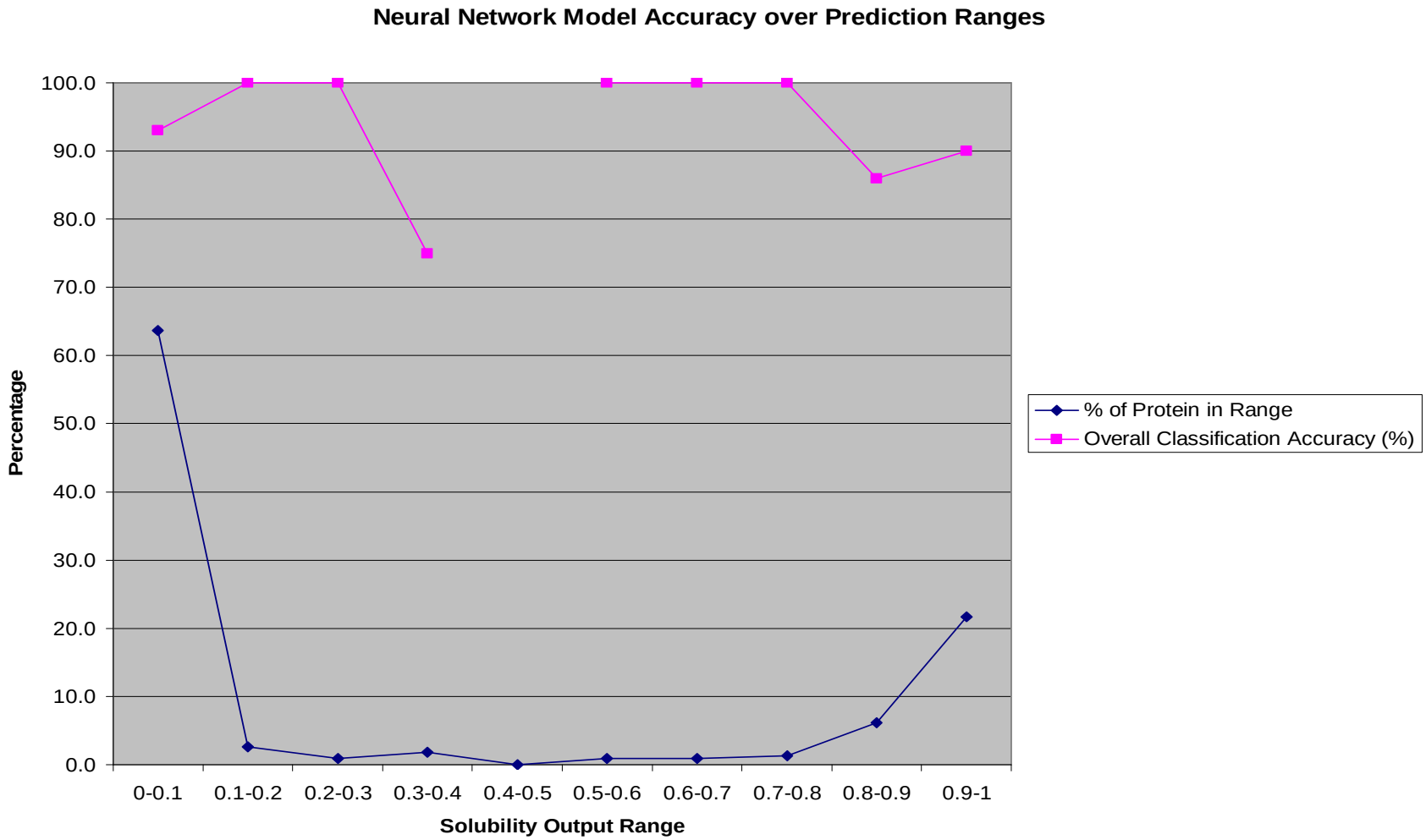
Results of Varying the Test Set Size

	Training Accuracy(%)			Test Accuracy(%)		
% Training Set Proteins/% Test Set Proteins	Soluble	Insoluble	Overall	Soluble	Insoluble	Overall
80/20	83	96	92	44	72	63
85/15	86	95	92	54	76	69
90/10	84	96	92	54	72	66
95/5	89	92	91	82	77	80

Trend indicates that on average, prediction accuracy on new proteins will be worse (possibly 15 to 25%) than training accuracy given post hoc

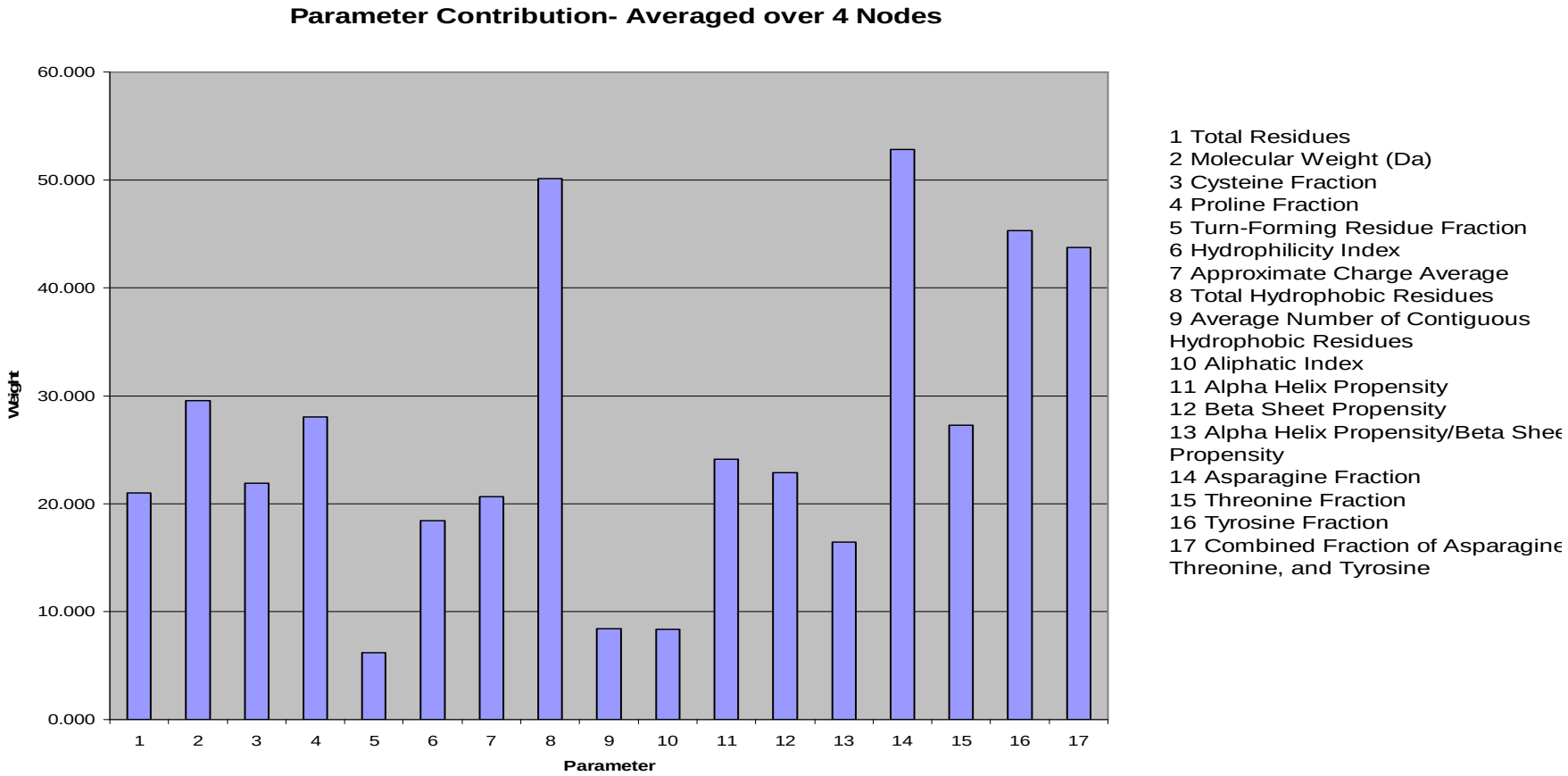
Also indicates that predictions for soluble and insoluble proteins are fairly well-balanced

Variation of Accuracy with Output



Evaluating the Most Important Parameters

- The higher a parameter's weight, the higher the significance
- Asparagine, Tyrosine, and Total Hydrophobic Residues Most Important



Comparing the Methods

Method	***Post hoc* accuracy (for entire database)**
Discriminant Analysis	66.5%
Logistic Regression	73.9%
Neural Networks	91.0%

Comparing the Methods

Method	***A priori* accuracy (10% of database for testing)**
Logistic Regression	≥78.7%
Neural Networks	≥66.0%

Model Trends

- Neural network has the highest *post hoc* accuracy, while logistic regression has the highest accuracy when predicting new proteins
- Logistic regression model very accurate for high and low probability *post hoc* predictions
- Neural network better than statistical methods at classifying soluble proteins correctly

Comparing Three Methods

- Asparagine common to NN and DA;
  Hydrophobic residues common to NN and LR
- Asparagine only parameter found significant in all three models
- Prediction of solubility from amino acid sequence and primary structure extremely difficult
- Secondary structure data would be very useful, but information is limited
- Neural networks represent the most promising method for solubility prediction with the available data

Recommendations for Further Study

- Examine other parameters
  - Secondary structure
  - Second virial coefficient
- Investigate parameter interactions
- Utilize all models in concert
- Incorporate more proteins from other host organisms

Acknowledgements

- Dr. Miguel Bagajewicz
- Dr. Roger Harrison
- Armando Diaz
- Zehra Tosun