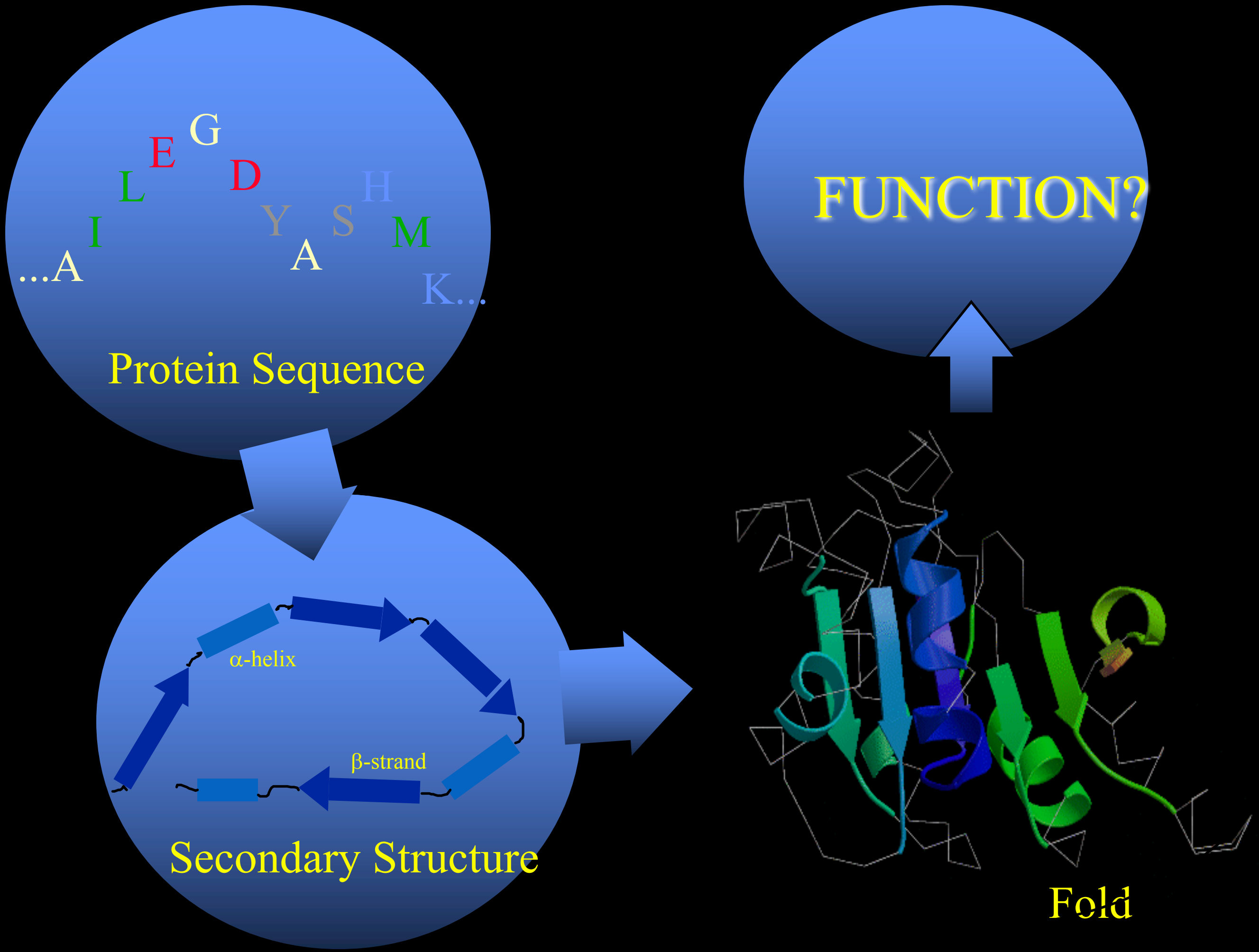
JPred and Jnet: Protein Secondary Structure Prediction

www.compbio.dundee.ac.uk/jpred



Protein Sequence

Secondary Structure

Fold

FUNCTION?

What is the difference between JPred and Jnet???

- JNet refers to the prediction “engine” that does the work. The current version of this is Version 2.3.1

- JPred refers to the website. This uses JNet and other tools to do predictions and present them in different ways.

History of JPred/JNet

- 1987: Zpred: First predictor that used multiple sequence alignment

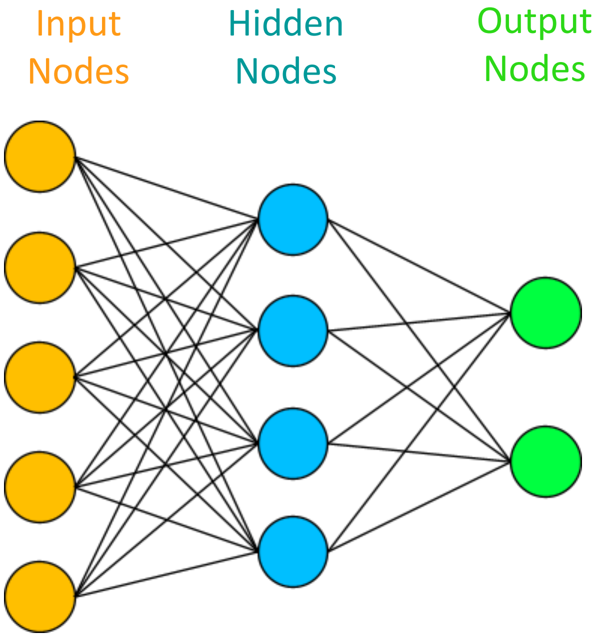
- 1999: Jpred 1: Did prediction by combining prediction methods developed by different groups that worked from multiple sequence alignments

- 2000: JNet 1: Multiple neural network predictor replaced all other methods in JPred
- 2002: JPred 2: Retraining JNet – improved accuracy
- 2009: JPred 3: Retraining JNet, algorithm improvements to Jnet and website refresh
- 2015: Jpred 4: Retraining JNet, major website improvements

Neural Network???

Machine learning method

Neural Networks



- Inductive method of learning
- Supervised learning
  - Inputs and outputs provided
- Dependent on ‘quality’ of observations
  - Representative
  - Unbiased (non-redundant)

Training and Testing JPred4/Jnet 2.3.1

- You need training data – where you know the answer.
  - We use a set of PDB domains of known structure from the SCOP domains database

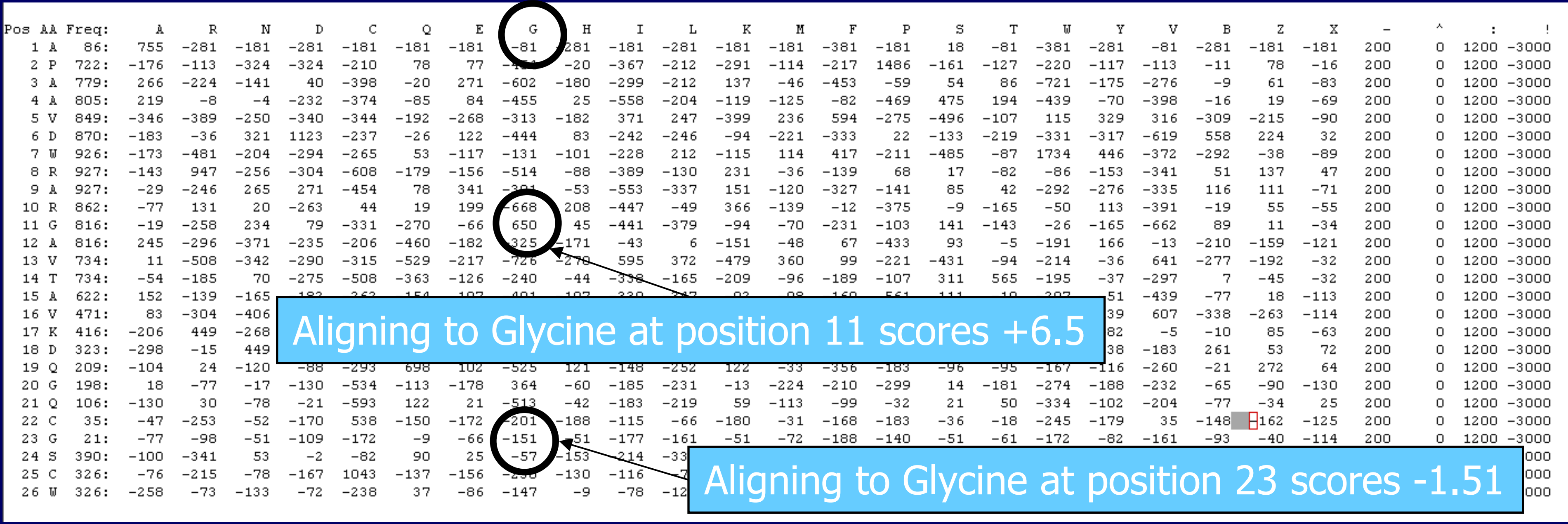
- Testing
  - 1. Cross-validation on 1208 domains
  - 2. Blind test on 150 domains not used in training

Neural Network Inputs

- Generate alignments for each sequence by searching UniRef90 with PSI-BLAST
- Make profiles:
  - Position-Specific Scoring Matrix (PSSM)
  - Hidden Markov Model (HMMer3)

- Earlier versions of JNet/Jpred had more inputs.

Profiles give position-specific scoring



```
Pos AA Freq:     A     R     N     D     C     Q     E     G     H     I     L     K     M     F     P     S     T     W     Y     V     B     Z     X     -     ^     :     !
  1 A   86:    755  -281  -181  -281  -181  -181  -181   -81  -281  -181  -281  -181  -181  -381  -181    18   -81  -381  -281   -81  -281  -181  -181   200     0  1200 -3000
  2 P  722:   -176  -113  -324  -324  -210    78    77 [illegible] -20  -367  -212  -291  -114  -217  1486  -161  -127  -220  -117  -113   -11    78   -16   200     0  1200 -3000
  3 A  779:    266  -224  -141    40  -398   -20   271  -602  -180  -299  -212   137   -46  -453   -59    54    86  -721  -175  -276    -9    61   -83   200     0  1200 -3000
  4 A  805:    219    -8    -4  -232  -374   -85    84  -455    25  -558  -204  -119  -125   -82  -469   475   194  -439   -70  -398   -16    19   -69   200     0  1200 -3000
  5 V  849:   -346  -389  -250  -340  -344  -192  -268  -313  -182   371   247  -399   236   594  -275  -496  -107   115   329   316  -309  -215   -90   200     0  1200 -3000
  6 D  870:   -183   -36   321  1123  -237   -26   122  -444    83  -242  -246   -94  -221  -333    22  -133  -219  -331  -317  -619   558   224    32   200     0  1200 -3000
  7 W  926:   -173  -481  -204  -294  -265    53  -117  -131  -101  -228   212  -115   114   417  -211  -485   -87  1734   446  -372  -292   -38   -89   200     0  1200 -3000
  8 R  927:   -143   947  -256  -304  -608  -179  -156  -514   -88  -389  -130   231   -36  -139    68    17   -82   -86  -153  -341    51   137    47   200     0  1200 -3000
  9 A  927:    -29  -246   265   271  -454    78   341 [illegible] -53  -553  -337   151  -120  -327  -141    85    42  -292  -276  -335   116   111   -71   200     0  1200 -3000
 10 R  862:    -77   131    20  -263    44    19   199  -668   208  -447   -49   366  -139   -12  -375    -9  -165   -50   113  -391   -19    55   -55   200     0  1200 -3000
 11 G  816:    -19  -258   234    79  -331  -270   -66   650    45  -441  -379   -94   -70  -231  -103   141  -143   -26  -165  -662    89    11   -34   200     0  1200 -3000
 12 A  816:    245  -296  -371  -235  -206  -460  -182  -325  -171   -43     6  -151   -48    67  -433    93    -5  -191   166   -13  -210  -159  -121   200     0  1200 -3000
 13 V  734:     11  -508  -342  -290  -315  -529  -217  -726  -270   595   372  -479   360    99  -221  -431   -94  -214   -36   641  -277  -192   -32   200     0  1200 -3000
 14 T  734:    -54  -185    70  -275  -508  -363  -126  -240   -44  -338  -165  -209   -96  -189  -107   311   565  -195   -37  -297     7   -45   -32   200     0  1200 -3000
 15 A  622:    152  -139  -165                                                                                        51  -439   -77    18  -113   200     0  1200 -3000
 16 V  471:     83  -304  -406                                                                                        39   607  -338  -263  -114   200     0  1200 -3000
 17 K  416:   -206   449  -268                                                                                        82    -5   -10    85   -63   200     0  1200 -3000
 18 D  323:   -298   -15   449                                                                                        38  -183   261    53    72   200     0  1200 -3000
 19 Q  209:   -104    24  -120   -88  -293   698   102  -525   121  -148  -252   122   -33  -356  -183   -96   -95  -167  -116  -260   -21   272    64   200     0  1200 -3000
 20 G  198:     18   -77   -17  -130  -534  -113  -178   364   -60  -185  -231   -13  -224  -210  -299    14  -181  -274  -188  -232   -65   -90  -130   200     0  1200 -3000
 21 Q  106:   -130    30   -78   -21  -593   122    21  -513   -42  -183  -219    59  -113   -99   -32    21    50  -334  -102  -204   -77   -34    25   200     0  1200 -3000
 22 C   35:    -47  -253   -52  -170   538  -150  -172  -201  -188  -115   -66  -180   -31  -168  -183   -36   -18  -245  -179    35  -148  -162  -125   200     0  1200 -3000
 23 G   21:    -77   -98   -51  -109  -172    -9   -66  -151   -51  -177  -161   -51   -72  -188  -140   -51   -61  -172   -82  -161   -93   -40  -114   200     0  1200 -3000
 24 S  390:   -100  -341    53    -2   -82    90    25   -57  -153  -214   -33                                                                                       000
 25 C  326:    -76  -215   -78  -167  1043  -137  -156 [illegible] -130 -116   -7                                                                                      000
 26 W  326:   -258   -73  -133   -72  -238    37   -86  -147    -9   -78   -12                                                                                       000
```

Aligning to Glycine at position 11 scores +6.5

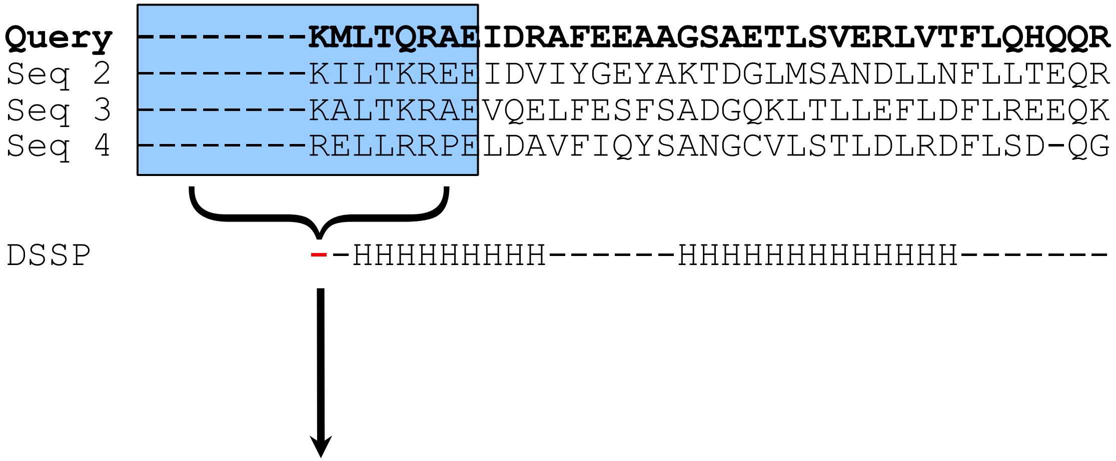
Aligning to Glycine at position 23 scores -1.51

This emphasises position-specific features of the protein family

Compared to Gly-Gly score of 0.6 in the BLOSUM62 matrix.

Neural Network Outputs

- DSSP definitions of secondary structure reduced from 8- to 3-state
  - H: Helix
  - E or B: Extended strand
  - Everything else: coil

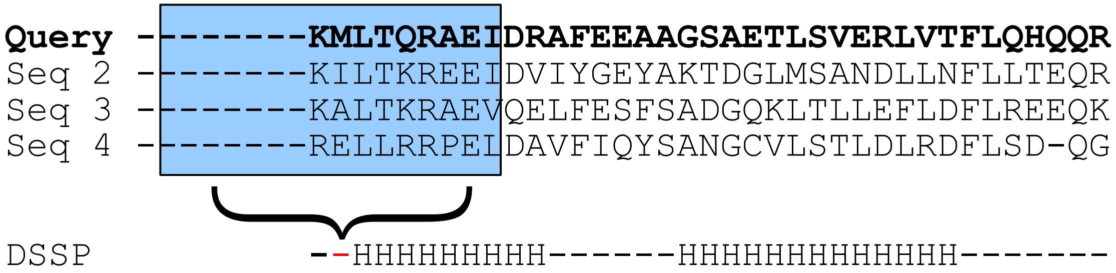


```
Query --------KMLTQRAEIDRAFEEAAGSAETLSVERLVTFLQHQQR
Seq 2 --------KILTKREEIDVIYGEYAKTDGLMSANDLLNFLLTEQR
Seq 3 --------KALTKRAEVQELFESFSADGQKLTLLEFLDFLREEQK
Seq 4 --------RELLRRPELDAVFIQYSANGCVLSTLDLRDFLSD-QG

DSSP          --HHHHHHHHH------HHHHHHHHHHHHH-------
```

```
# Input 1
0 0 0 0 0 0 0 0 0.9820 0.1192 0.28689 ...
# Output 1
1 0 0
```

Each position in the input vector is the score for an amino acid within the window

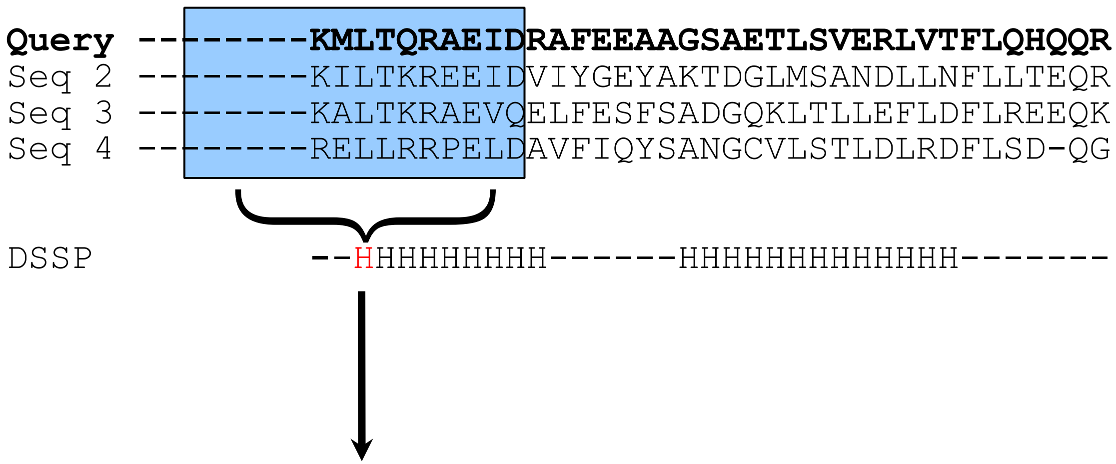


```
Query --------KMLTQRAEIDRAFEEAAGSAETLSVERLVTFLQHQQR
Seq 2 --------KILTKREEIDVIYGEYAKTDGLMSANDLLNFLLTEQR
Seq 3 --------KALTKRAEVQELFESFSADGQKLTLLEFLDFLREEQK
Seq 4 --------RELLRRPELDAVFIQYSANGCVLSTLDLRDFLSD-QG

DSSP          --HHHHHHHHH------HHHHHHHHHHHHH-------
```

```
# Input 2
0 0 0 0 0 0 0 0.9820 0.1192 0.28689 0.0474 ...
# Output 2
1 0 0
```

Each position in the input vector is the score for an amino acid within the window



```
Query ---------KMLTQRAEIDRAFEEAAGSAETLSVERLVTFLQHQQR
Seq 2 ---------KILTKREEIDVIYGEYAKTDGLMSANDLLNFLLTEQR
Seq 3 ---------KALTKRAEVQELFESFSADGQKLTLLEFLDFLREEQK
Seq 4 ---------RELLRRPELDAVFIQYSANGCVLSTLDLRDFLSD-QG

DSSP           --HHHHHHHHH------HHHHHHHHHHHHH-------
```

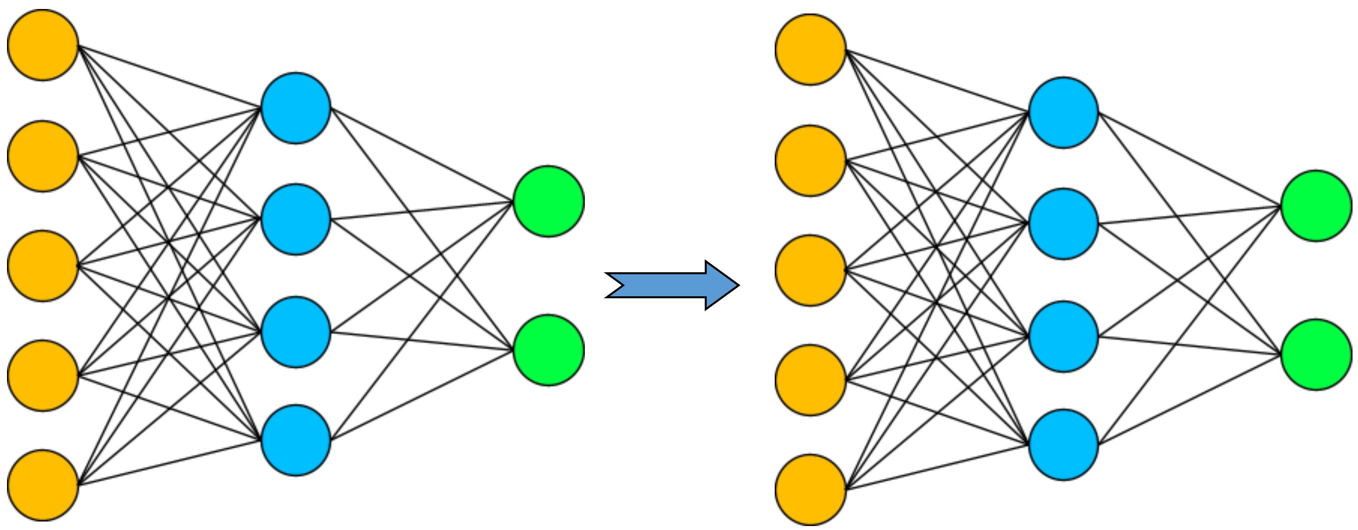
Each position in the input vector is the score for an amino acid within the window

```
# Input 3
0 0 0 0 0 0 0.9820 0.1192 0.28689 0.0474 0.1192 ...
# Output 3
0 0 1
```

Two-Layer Ensemble

Sequence to Structure

Structure to Structure



E-EEEEH-------HHH-HHH-HHH-HE-

--EEEEE-------HHHHHHHHHHHHHH-

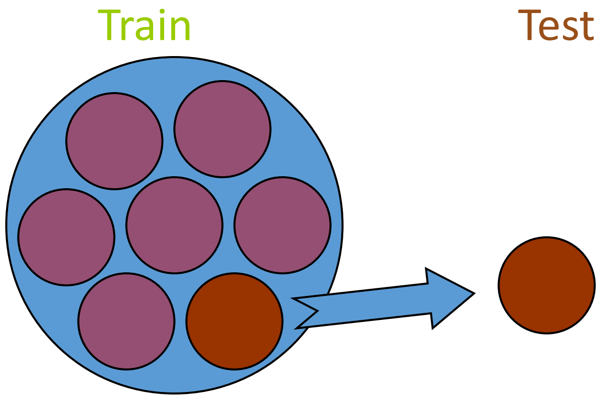
Actually, both have hundreds of inputs and three outputs – only two outputs shown for simplicity

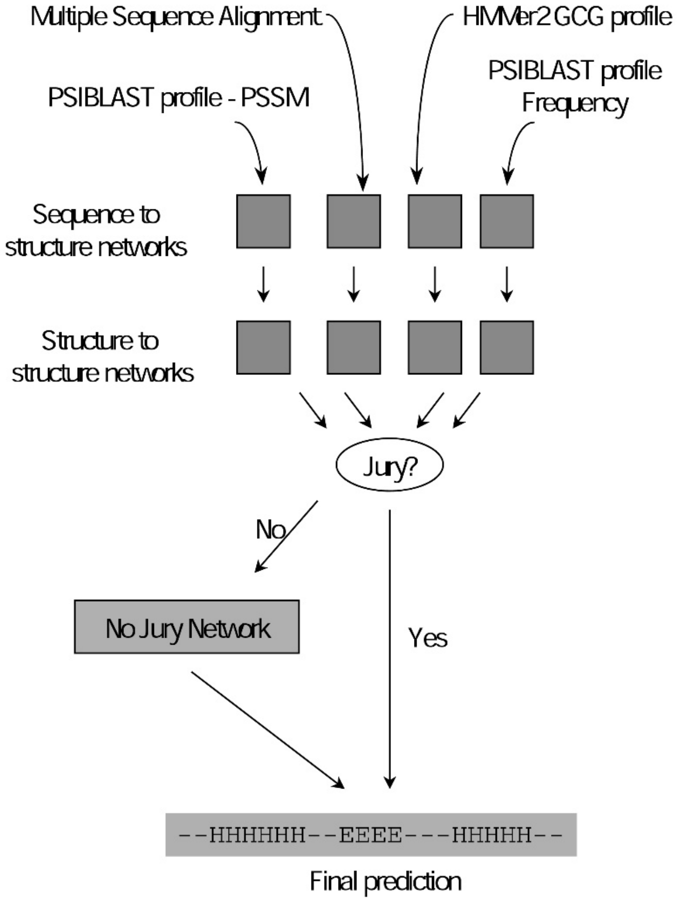
Training and Testing
JPred4/Jnet 2.3.1

- You need training data – where you know the answer.
  - We use a set of PDB domains of known structure from the SCOP domains database
- Testing
  - 1. Cross-validation on 1358 domains
  - 2. Blind test on 150 domains not used in training

Cross-validation training

- *'Blind'* data - removed subset
- *k*-fold Cross-Validation (887 seqs)
  - Divide training data into *k* groups, train on *k*-1 and test on remainder. Do this *k* times.

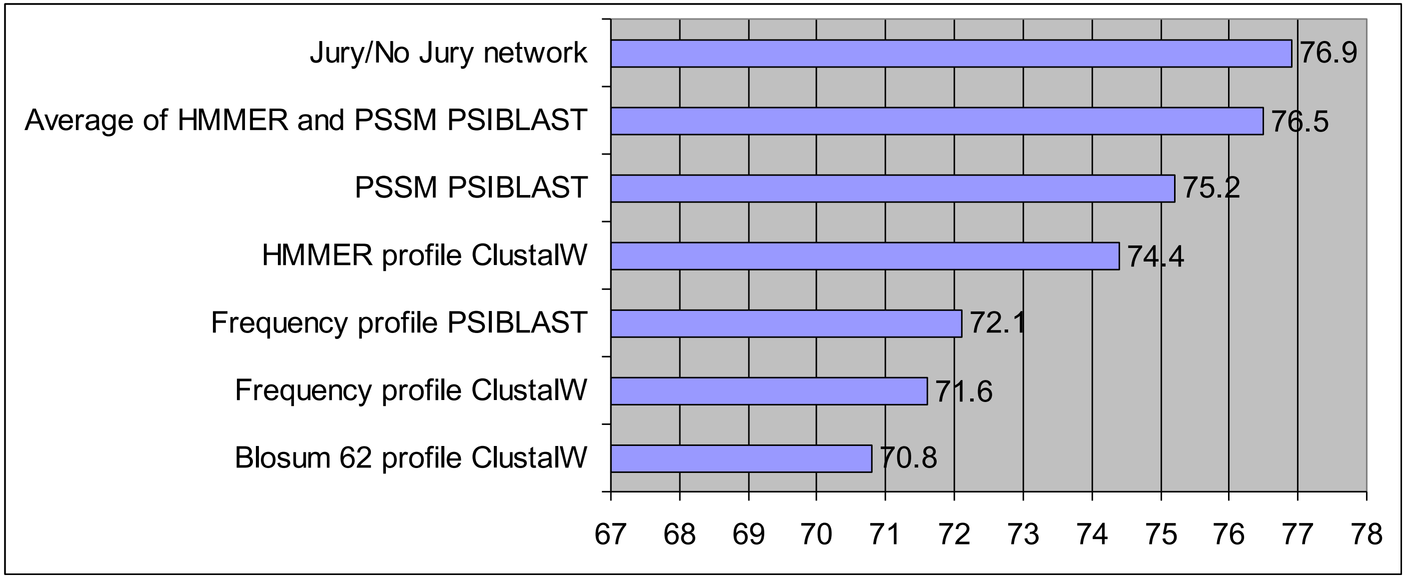




Jnet Version 1: Multiple methods of presenting alignment information.

If alternative networks do not agree, predict with network trained on difficult to predict regions.

JNet Version 1: Effect of Different Alignment Inputs



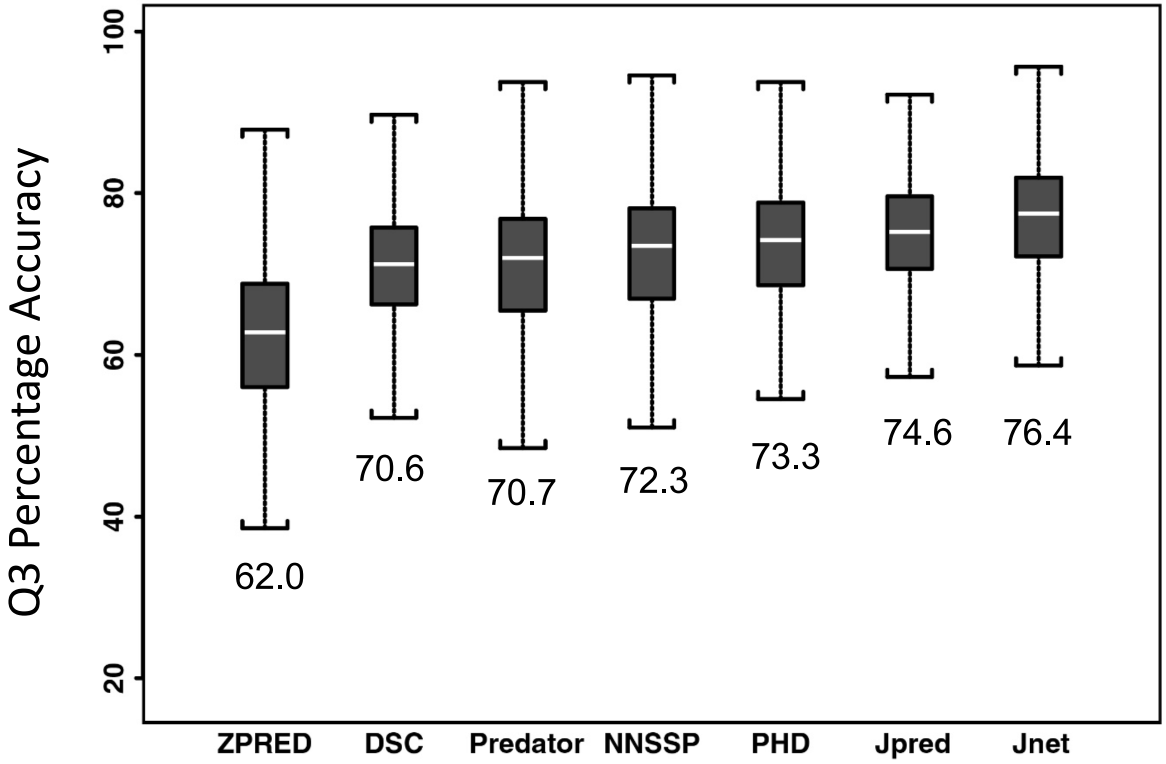
Average Percentage Accuracy
(7-fold cross-validation on 480 proteins)

JNet also accurately predicts whether amino acids will be buried or exposed in the folded structure of the protein.

Blind Test

JNet Version 1

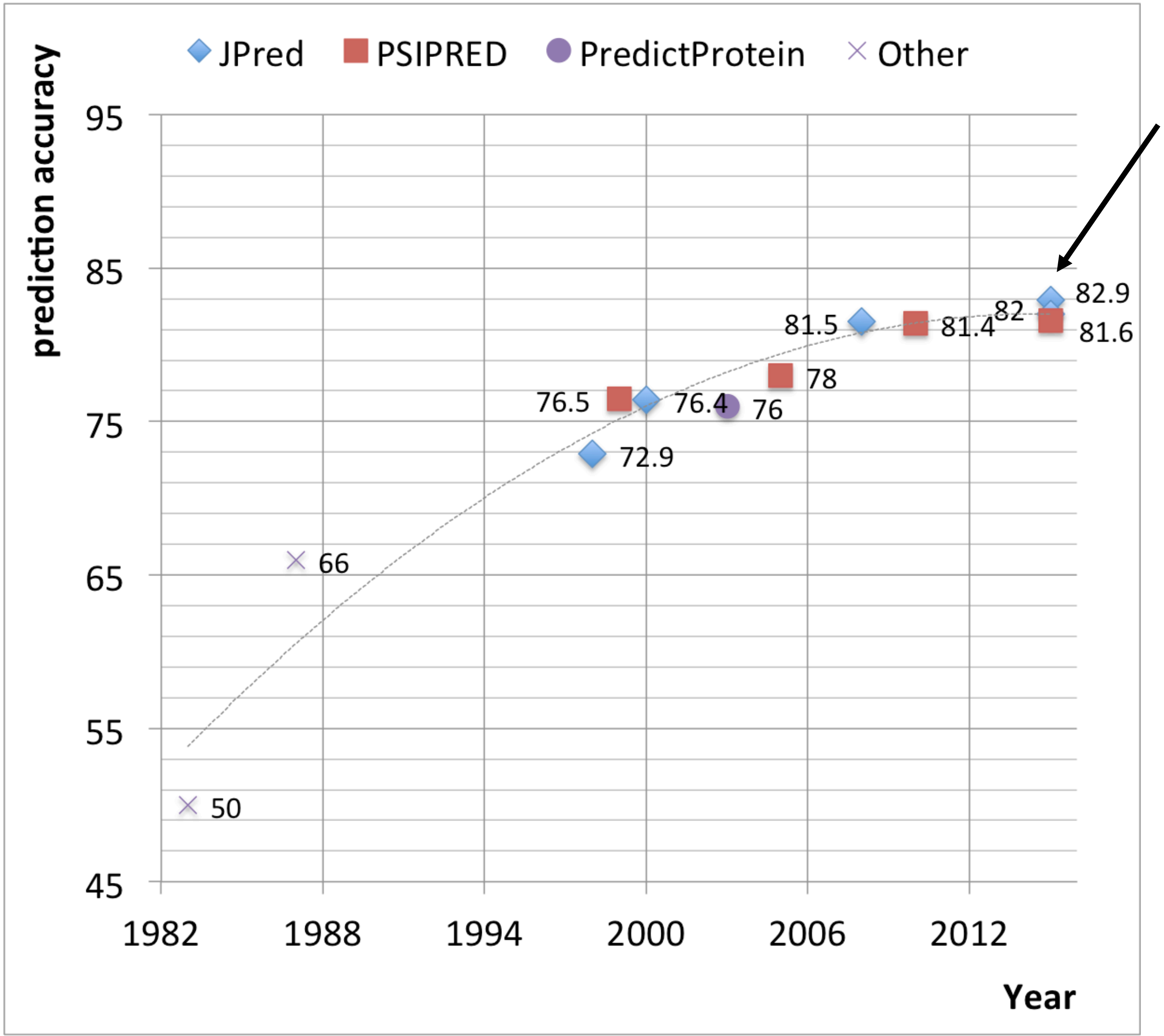
Comparison of JNet Version 1.0 to other Prediction Methods in a Blind Test - (406 proteins)



Cuff, J. A. & Barton, G. J., (2000), *Proteins 40: 502-511.*

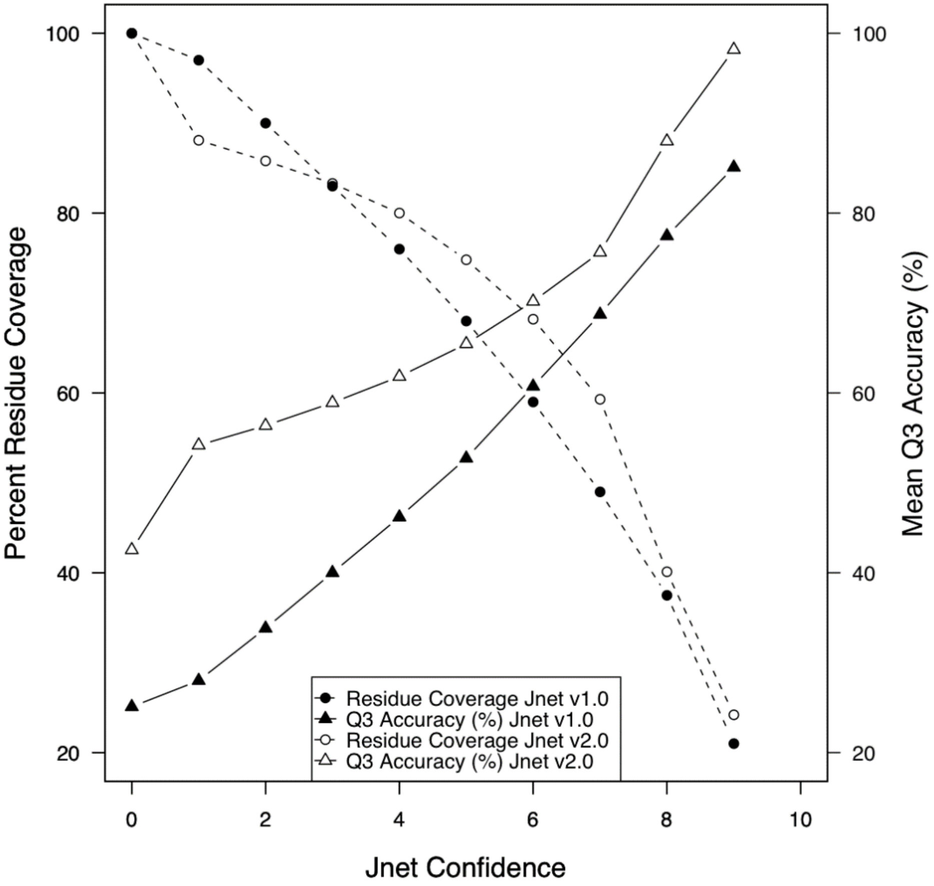
That was in 2000
What has happened since?

Average Prediction Accuracy is Rising, but Flattening off

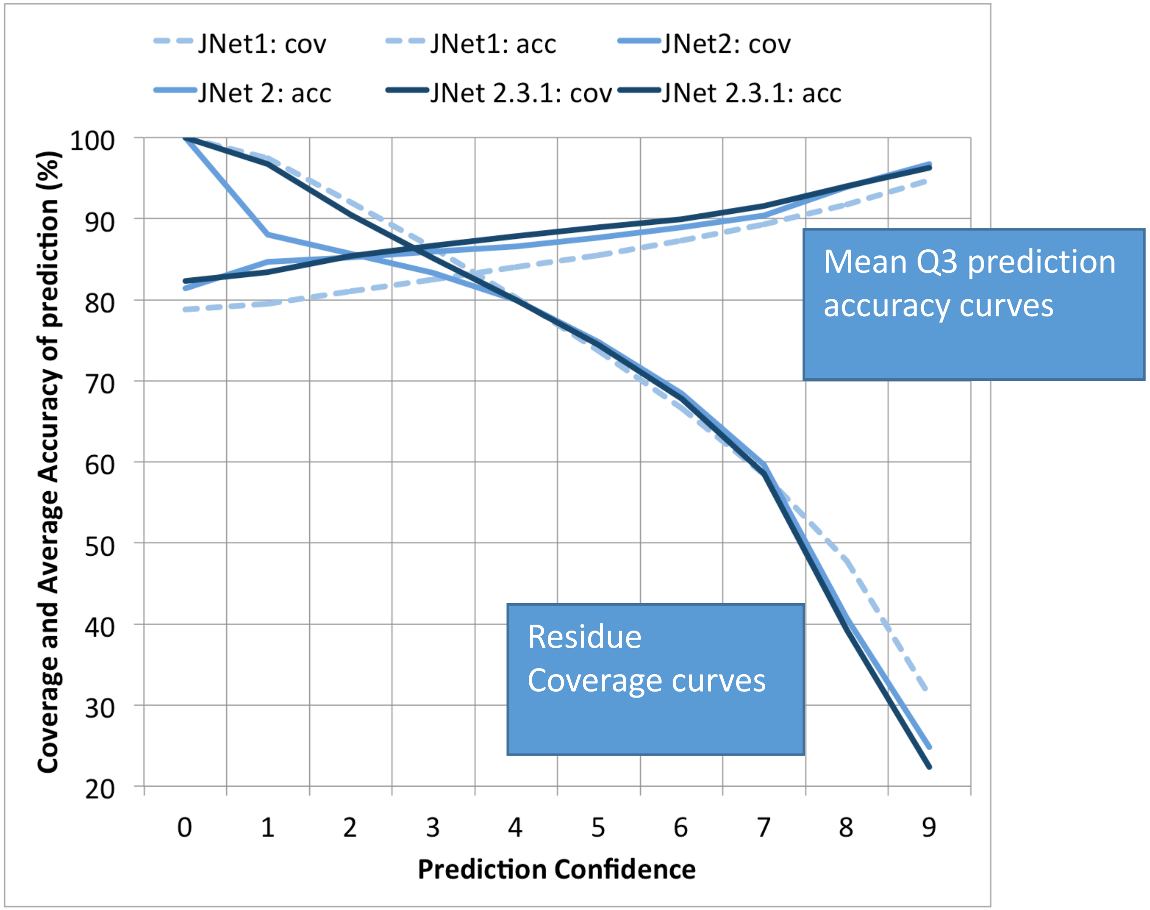


Current JPred

Confidence in prediction. JNet 1.0 vs Jnet 2.0



Latest Jnet…

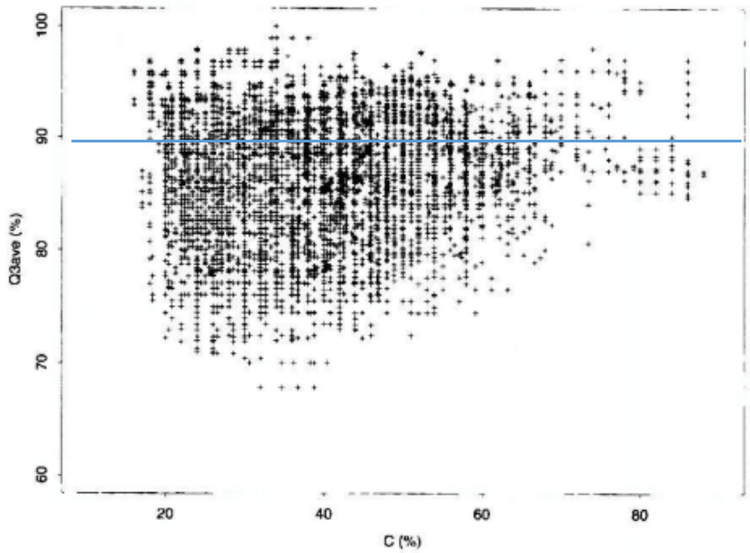


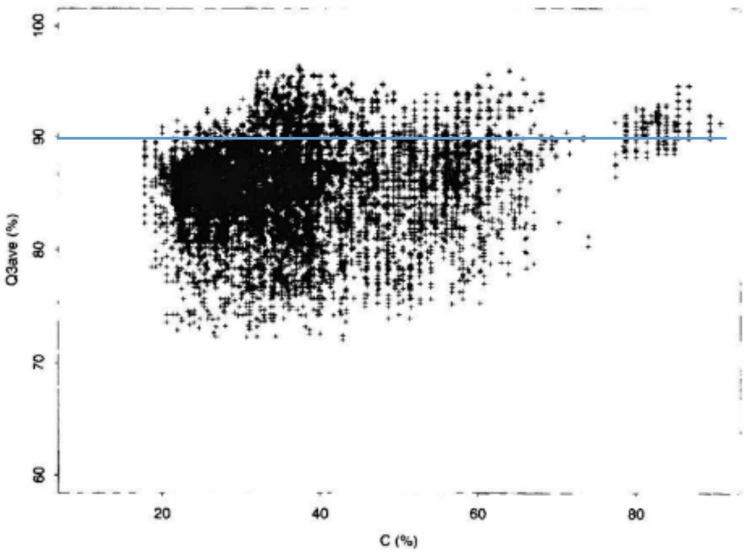
Mean Q3 prediction accuracy curves

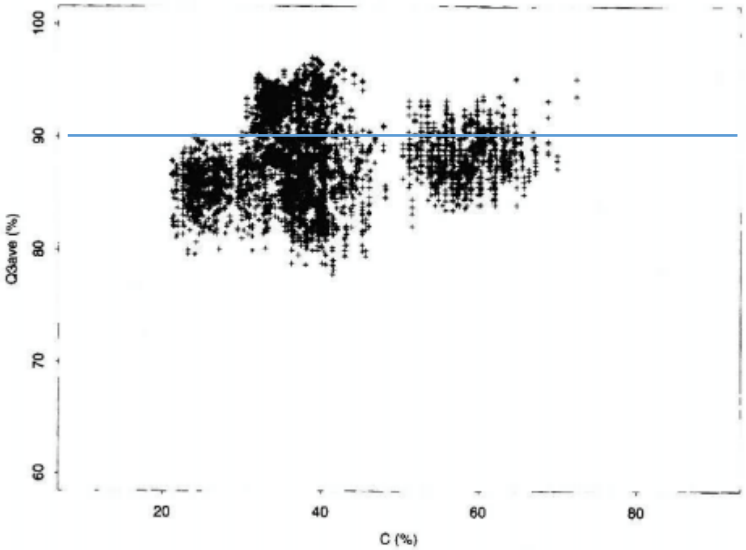
Residue Coverage curves

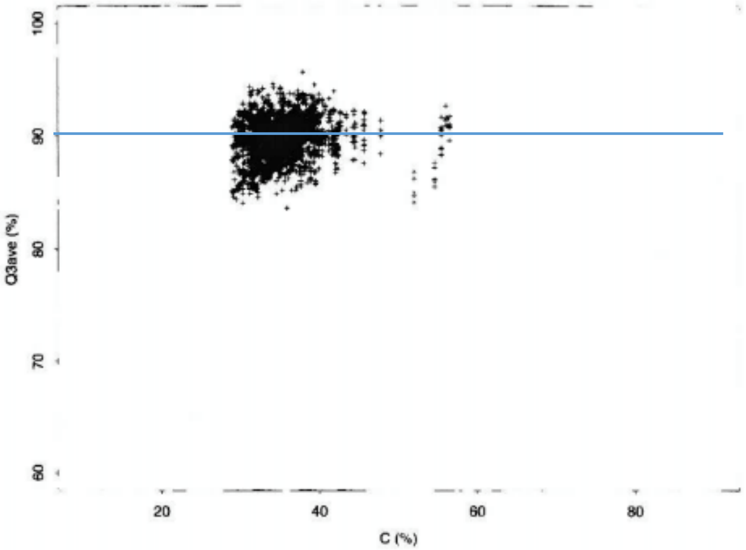
Introducing the Practical

- At last!


a) length <=50


b) length 51 - 100


c) length 101 - 150


d) length > 150