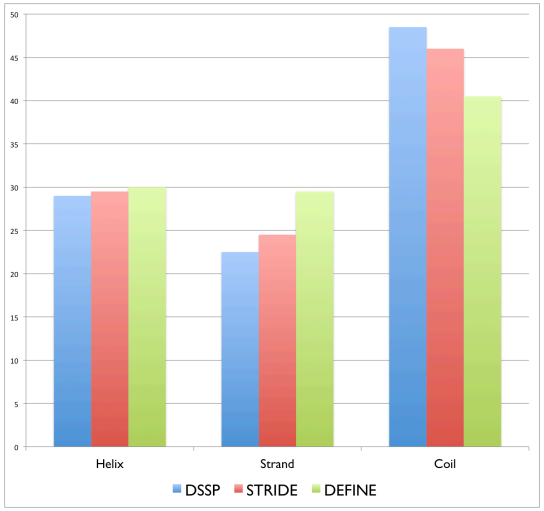
Data

- Set of 494 unique proteins with a PDB structure
  - Less than 20% sequence identity (PISCES)

- Secondary structure already determined by DSSP and STRIDE



Data

- Tab-delimited files for DSSP and STRIDE

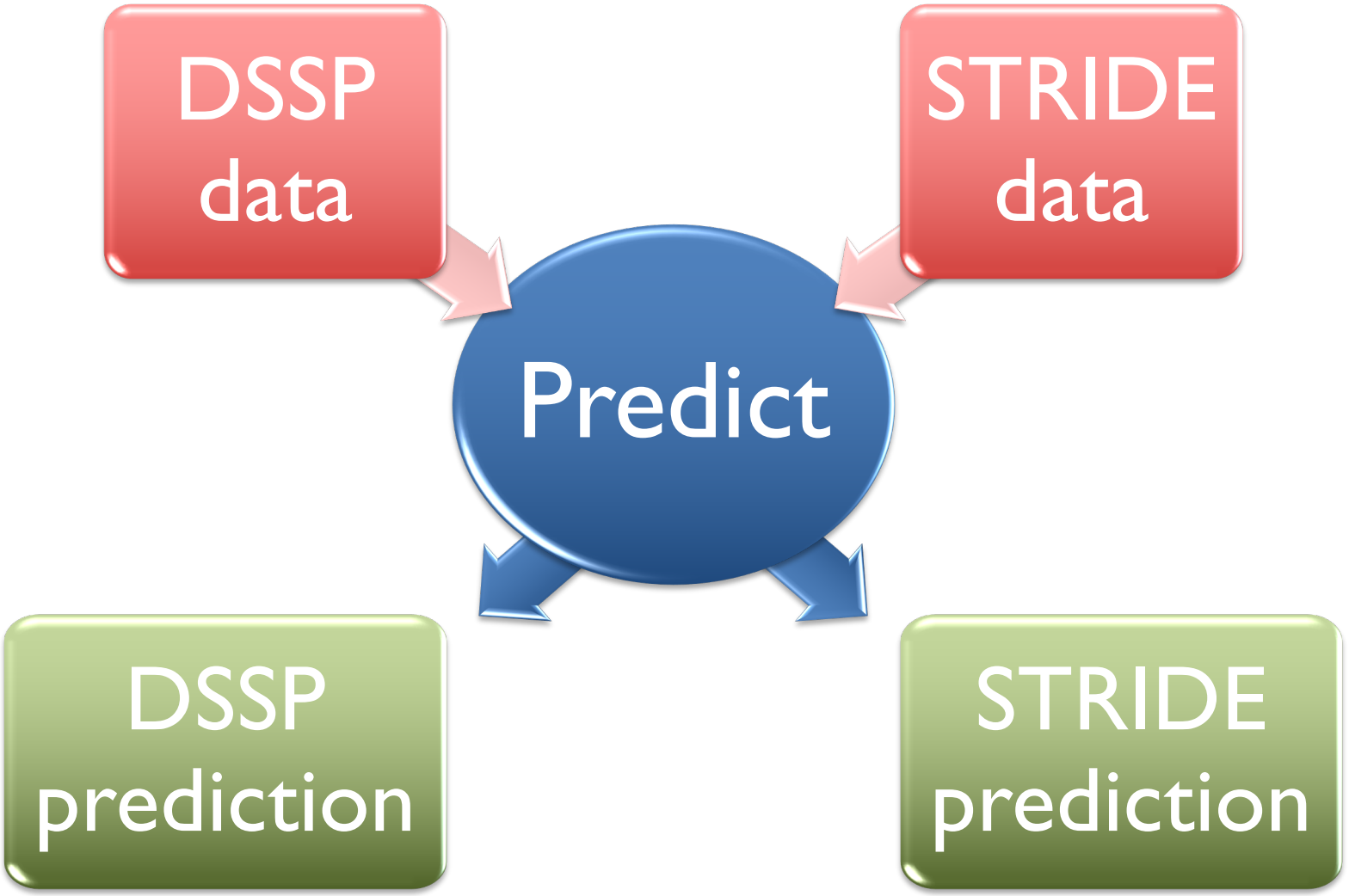
PDB code	PDB chain code	PDB sequence code	Residue	Actual secondary structure
1w0n	A	12	ILE	Other
1w0n	A	13	THR	Beta
1w0n	A	14	LYS	Beta
1w0n	A	15	VAL	Beta
1w0n	A	16	GLU	Beta
1w0n	A	17	ALA	Other
1w0n	A	18	GLU	Other
1w0n	A	19	ASN	Other
1w0n	A	20	MET	Other
1w0n	A	21	LYS	Beta

Data

- Tab-delimited file CATH protein family

PDB code	PDB chain code	Protein family
1w0n	A	Beta
2gpi	A	Alpha/beta
1vbw	A	Alpha/beta
2odk	A	Alpha/beta
2zxy	A	Alpha
2pr7	A	Alpha/beta
2pyq	A	Alpha
1jy2	N	Alpha

Implementation



GOR III

- GOR III

$$I(\Delta S_j; R_1, \ldots, R_n) \approx I(\Delta S_j; R_j) + \sum_{\substack{m=-8 \\ m \neq 0}}^{m=8} I(\Delta S_j; R_{j+m} \mid R_j)$$

Self information

Pair information

$$I(\Delta S_j; R_{j+m} \mid R_j) = \log(f_{S_j, R_{j+m}, R_j} / f_{n-S_j, R_{j+m}, R_j}) + \log(f_{n-S_j, R_j} / f_{S_j, R_j})$$

Implementation

- Predict secondary structure for each of 494 proteins
- Protein to calculate cannot be in data set
  - Implement jackknife/leave-one-out!
- Calculate Q_3 and MCC quality scores
  - Prediction variation between STRIDE and DSSP
  - Both for whole set and per protein family

Implementation

- Q_3 measure

$$Q_3 = \frac{N_{residues\_correctly\_predicted}}{N_{residues\_total}}$$

- Matthews correlation coefficient (MCC)
  - Predicted for each secondary structure state
  - Uses false/true positives and false/true negatives

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Implementation

- Based on the secondary structure prediction, predict the protein family.
  - Can determine own criteria for this step
  - Compare prediction against actual protein family

- Use evolutionary information to improve your results (OPTIONAL)