Mr. Residue's Neighborhood: Using Correlated Mutations, Mutual Information Statistics, and Neural Networks in Residue-Residue Contact Predictions

George Shackelford

Protein Structure Prediction

The goal is to predict the structure of a protein when folded from the protein sequence.

- 1D Methods
  - Secondary Structure Prediction
  - Hydrophobicity
- 3D Methods
  - Structure-Structure Alignment
  - Undertaker

What about 2D?

Residue-Residue Contacts

Given a protein sequence we say that two residues, indexed as i and j, are in contact if the distance between their respective C_β atoms is less than 8 Å.

- Nothing to do with Van der Waals distance
- This definition is arbitrary!
- They help with the tertiary structure
- We define separation as $|i - j|$

How do we find these contacts?

Correlated Mutations

When a residue in a protein structure mutates, there is a possibility that an nearby residue will mutate.

- salt bridges
- other sidechain-sidechain interactions
- functional regions
- possible size fittings

How can we detect these correlated mutations?

Using Mutual Information

Given residue indices, i and j, and the pair of columns, $columns_{i,j}$, under i and j we define the *mutual information*, MI, as

$$MI(i,j) = \sum_{(k,l) \in columns_{i,j}} p(k,l) \log_2 \frac{p(k,l)}{p(k)\,p(l)}$$

where $p(k,l)$ is the joint probability of the corresponding residues k and l from the two columns and $p(k), p(l)$ are the marginal probabilities.

Using Mutual Information

- High values of MI indicate a correlation.
- When the columns are independent, MI is 0.
- When the columns are perfectly conserved, this value is also 0.

Problems:

- Likely to over-estimate when sample is small
- Having many recently evolved sequences can skew MI

Small Sample Correction

We hold the marginal probabilities fixed, and randomly re-arrange the joint probabilities, re-calculating the MI each time. Then we plot these values as a histogram, and fit a Gamma distribution to it. Using this distribution and the orginal MI, we calcuate a 'corrected' MI by subtracting the mean of the distribution.
Also we can calculate an e-value from the distribution.

Thinning

To correct for the skew from recently-evolved sequences in the alignment, we thin the set of sequences in the alignment by removing those that have less than a specified percent of identity with at least one other sequence in the set. Then we re-calcuate the MI and e-values using the thinned set.

Scoring Methods

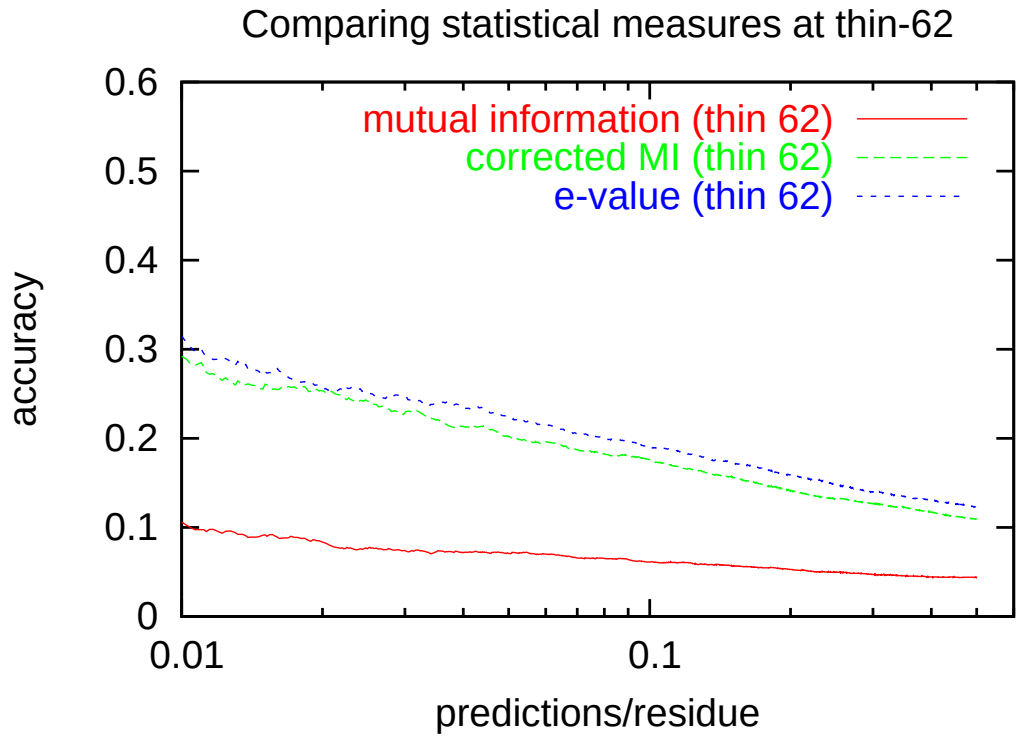
$S(i,j)$ represents our score for a contact between i and j and given some threshold, t:

$$\frac{\sum_{S(i,j)>t} Contact(i,j)}{|S(i,j) > t|}$$

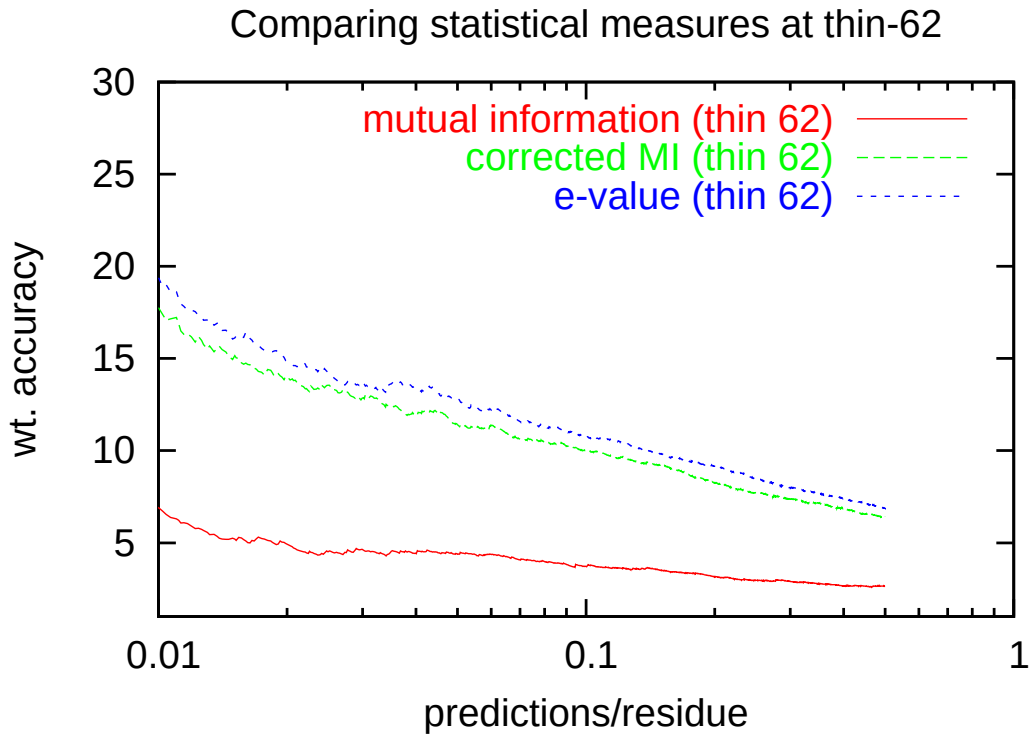
Weighted accuracy vs. contacts/residue – to compensate for high separations that have lower probability of contact

$$\frac{\sum_{S(i,j)>t} \frac{Contact(i,j)}{Prob_contact(|i-j|)}}{|S(i,j) > t|}$$

Results


Comparing statistical measures at thin-62
mutual information (thin 62)
corrected MI (thin 62)
e-value (thin 62)
accuracy
0.6
0.5
0.4
0.3
0.2
0.1
0
0.01
0.1
predictions/residue

Results


Comparing statistical measures at thin-62
mutual information (thin 62)
corrected MI (thin 62)
e-value (thin 62)
wt. accuracy
30
25
20
15
10
5
0.01
0.1
1
predictions/residue

Neural Networks

Given a large set of examples and a carefully selected set of inputs, they can converge to a useful predictor.

Unfortunately they are “black boxes” which tell us noth-

ing conceptually.

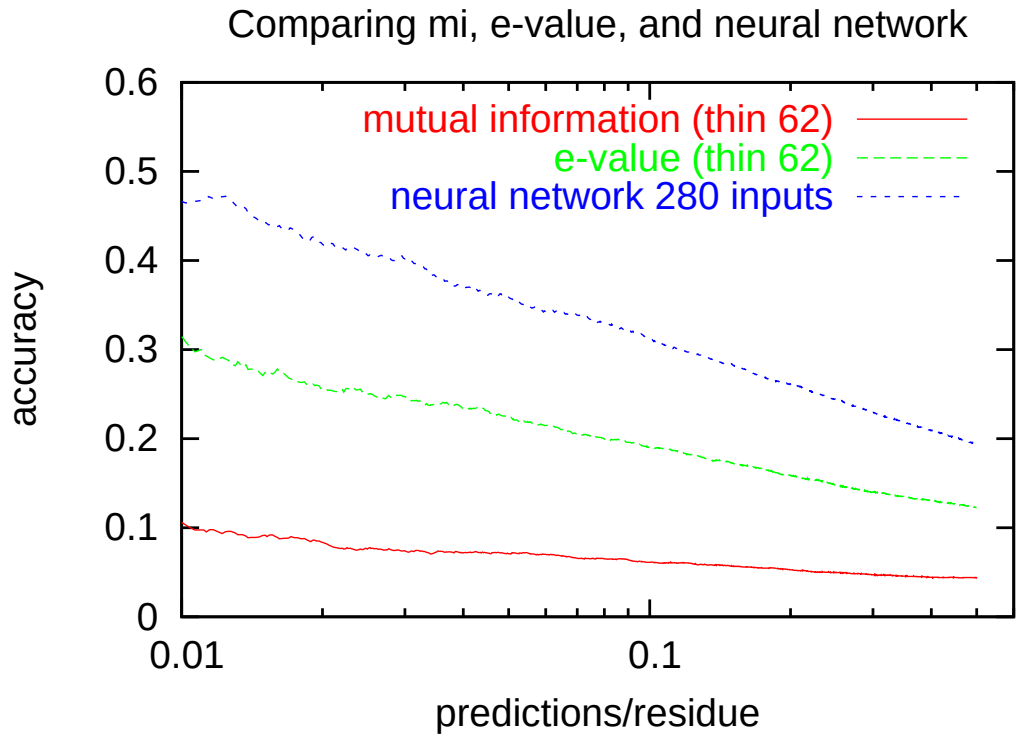
280 Inputs

Uses a window over i-1,i,i+1,j-1,j,j+1

- Corrected MI, e-values, and the no. of pairs to determine them when thinned to 62%,40%,35%,30%
- Amino acid distribution within the columns
- Seconary structure and burial predictions
- Entropy
- Sequence length, and separation of i,j

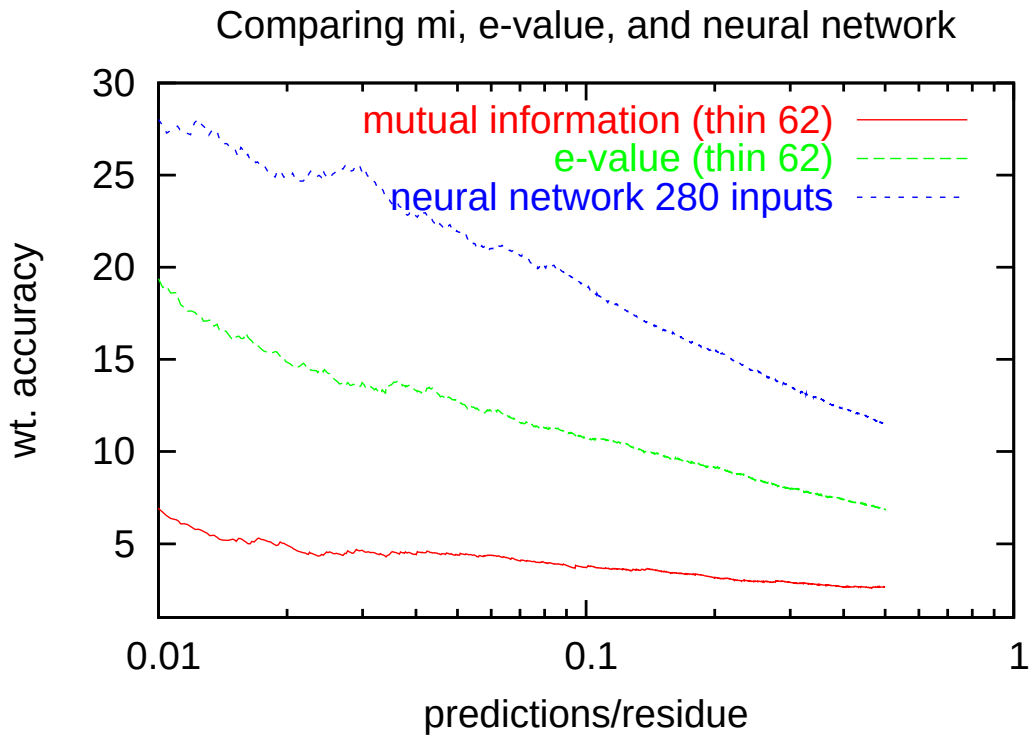
Results

Accuracy vs. contacts/residue



Results

Weighted accuracy vs. contacts/residue



Conclusions

- Small sample correction helps.
- Thinning shows mixed results.
- The neural networks can improve classification predictions significantly.

Future Work

- Consider side-chain distance in place of backbone distance.
- Analyze the neural net to determine the most signifcant inputs.
- Include thinnings of 80%,70%,50%.
- Determine a function for adjusting e-values based on thinnings.
- Add inputs concerning the distribution of the sequences in the alignment.

Thanks to...

Kevin Karplus
Martina Koeva
Sol Katzman
Jenny Draper
Bret Barnes
Marcia Soriano
Carol Rohl