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

George Shackelford

Goal:Protein Structure Prediction

If nothing else, we could help the crystallographers.

Structure Prediction Methods

- 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

Put an equation here!

Problems with Mutual Information

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

Using Thinning

Using Small Sample Correction

Scoring the Results

$S(i,j)$ represents our prediction of a contact between i and j and given some threshold, t:
Accuracy

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

Enrichment

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

Results of Corrections

Results of Thinning

Neural Nets

Inputs: Alphabets

Inputs: Distributions

Inputs: Neighboring Residues

Results

Conclusions

Future Work