Protein 3D Structure Computed from Evolutionary Sequence Variation

Debora S. Marks , Lucy J. Colwell , Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, Chris Sander

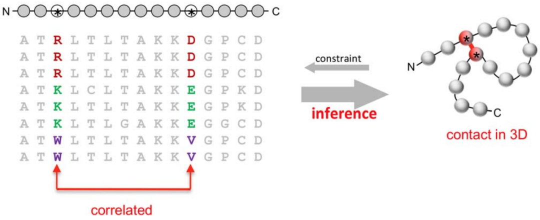
Presenter: Arshdeep Sekhon
https://qdata.github.io/deep2Read

Motivation

- A protein family: group of proteins that share a common evolutionary origin, reflected by their related functions and similarities in sequence or structure.
- Very large space of sequences, only few observed
- conservation of function imposes boundaries on sequence variation and ensures 3D structure similarity

Motivation

- to maintain energetically favorable interactions, residues in spatial proximity may co-evolve across a protein family
- suggests that residue correlations could provide information about amino acid residues that are close in structure



```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

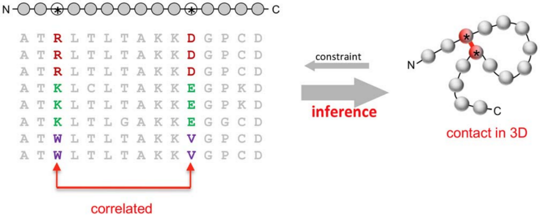
Residue-Residue Correlation

- correlated residue pairs within a protein are not necessarily close in 3D space
- Confounding Correlations:
    - transitivity of correlations: if (i,j),(j,k) correlated, (i,k) also correlated
    - technical noise, oligomerization, protein-protein, or protein-substrate interactions or other spatially indirect or spatially distributed interactions can result in co-variation between residues not in close spatial proximity.

Motivation

This Paper:

Infer evolutionary constraints from a set of sequence homologs of a protein.
Predicting 3D Protein Structure from these evolutionary interactions.



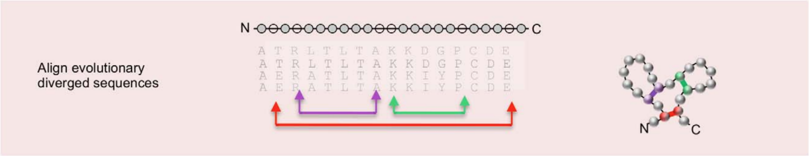
```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

Methods: Pipeline

1. Protein sequence alignment of an iso-structural protein family (from PFAM database) of length L
2. **Residue-Residue Coupling Scores(DI $\in R^{L \times L}$) for all pairs of residues in [1]**
3. Derivation of a ranked set of evolutionarily inferred contacts (EICs) from [2]
4. Prediction of 3D structures by using EICs

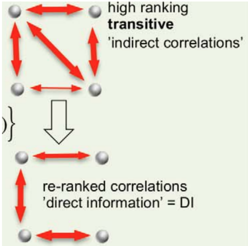
Step 1: Align Evolutionarily Diverged Sequences

Protein sequence alignment for the protein family containing the target protein (from PFAM database)



```
A T R L T L T A K K D G P C D E
A T R L T L T A K K D G P C D E
A E R A T L T A K K I Y P C D E
A E R A T L T A K K I Y P C D E
```

Step 2: Residue Coupling Scores



- For sequence length L for a protein family, a matrix $DI \in R^{L \times L}$ is inferred:

$$MI_{ij} = \sum_{A_i, A_j = 1}^{q} f(A_i, A_j) ln\Big(\frac{f(A_i, A_j)}{f_i(A_i) f_j(A_j)}\Big) \tag{1}$$

$$DI_{ij} = \sum_{A_i, A_j = 1}^{q} P_{ij}^{Dir} ln\Big(\frac{P_{ij}^{Dir}}{f_i(A_i) f_j(A_j)}\Big) \tag{2}$$

- q: types of residues (20)
- L: length of sequence (50-250 in these experiments)

Computing Residue Coupling Scores

- Estimate a $p(A_1, \ldots, A_L)$ such that it maximizes entropy $S = -\sum P(A_1, \ldots, A_L) lnP(A_1, \ldots, A_L)$ subject to the following constraints:

$$P_i(A_i) = \sum_{A_k=\{1,\ldots,q\}, k\neq i} P_i(A_1, \ldots, A_L) = f_i(A_i) \quad (3)$$

$$P_{ij}(A_i, A_j) = \sum_{A_k=\{1,\ldots,q\}, k\neq i,j} P_i(A_1, \ldots, A_L) = f_{ij}(A_i, A_j) \quad (4)$$

- Make empirical correlation matrix

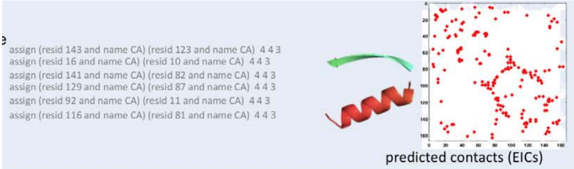
$$C_{ij} = f_{ij}(A_i, A_j) - f_i(A_i) f_j(A_j) \quad (5)$$

- $e_{ij} = C_{ij}^{-1}$

$$P_{ij}^{Dir} = \frac{1}{Z} exp\Big(e_{ij}(A_i, A_j) + h_i(A_i) + h_j(A_j) \Big) \quad (6)$$

3. Derivation of a ranked set of evolutionary inferred contacts (EICs)

- evolutionary inferred contacts (EICs): predicted to be close in 3D space
- Convert the above *DI* matrix into EICs using rules:
  - Remove residue pairs close in sequence
  - consistent with predicted secondary structure: PredictProtein and PsiPred Algorithms
  - ..



```
assign (resid 143 and name CA) (resid 123 and name CA)  4 4 3
assign (resid 16 and name CA) (resid 10 and name CA)  4 4 3
assign (resid 141 and name CA) (resid 82 and name CA)  4 4 3
assign (resid 129 and name CA) (resid 87 and name CA)  4 4 3
assign (resid 92 and name CA) (resid 11 and name CA)  4 4 3
assign (resid 116 and name CA) (resid 81 and name CA)  4 4 3
```

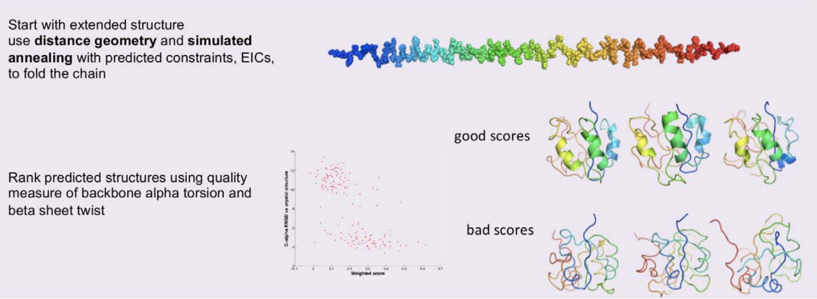
predicted contacts (EICs)

- The first N_c inferred EIC pairs are ranked according to the *DI* scores and used as distance constraints to distance geometry and simulated annealing calculations

Step 4: Prediction of 3D structures

EICs used as input to distance geometry and simulated annealing calculations.

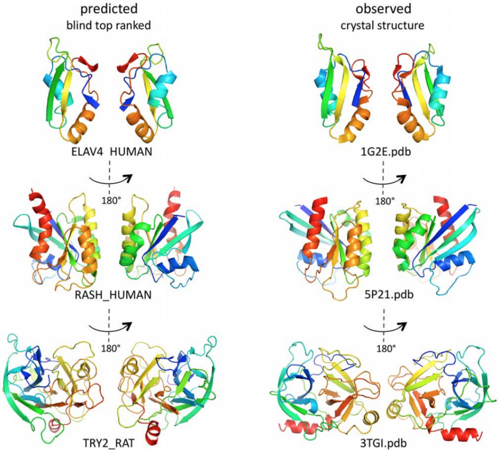
tested on multiple protein families (from PFAM database)with range of Multiple Sequence Alignment of 71/161/223



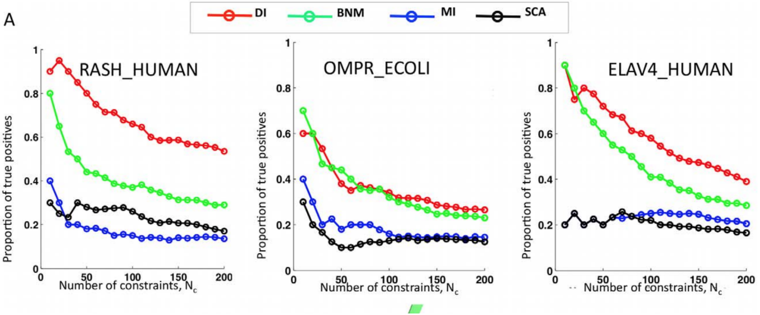
Start with extended structure use **distance geometry** and **simulated annealing** with predicted constraints, EICs, to fold the chain

Rank predicted structures using quality measure of backbone alpha torsion and beta sheet twist

Results:Prediction of 3D structures

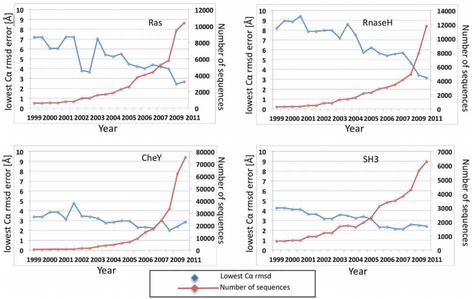


Evaluation of residue-residue contact prediction:



- BNM: Bayesian network model (also global)
- SCA: statistical coupling analysis (local)
- MI: Mutual Information(local) coupling analysis (local)

$C_\alpha - RMSD$ [1] Error as a function of number of sequences



Other factors:

- Which sequences are used/distribution of sequences in the protein family? For example, this algorithm removes sequences with over 70% residue identity to family neighbors are down-weighted
- uneven sampling in the space of natural sequences, due to experimental ascertainment bias during sequencing.

[1] the root-mean-square deviation of atomic positions- average distance between the atoms (usually the backbone atoms) of superimposed proteins.

Conclusion

- pairwise without indirect/confounding interactions for residue-residue contact prediction
- DI based(global) works better than MI based (local)
- Lots of feature engineering: data selection, removal of invalid correlations, etc