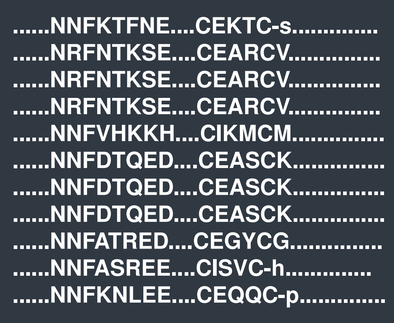
Review of Protein Residue Contact Inference

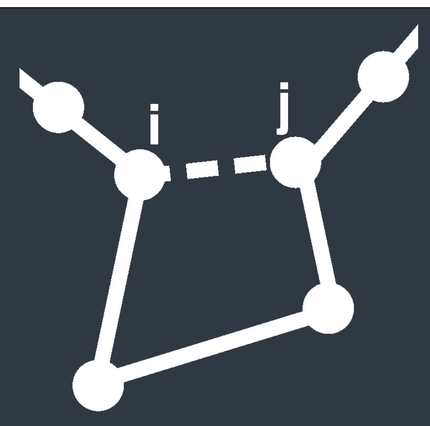
C. Feinauer[1], A. Pagnani[1,2], R. Zecchina[1,2]

1 Department of Applied Science, Politecnico di Torino, Torino, Italy.
2 Human Genetics Foundation, Torino, Italy.

NETADIS
Statistical Physics Approaches to Networks Across Disciplines

PROBLEM





SOLUTIONS

APC[1]

$$I_{ij} = \sum_{a_i,a_j} p(a_i,a_j) \log \frac{p(a_i,a_j)}{p(a_i)p(a_j)} \Rightarrow I_{ij} - \frac{I_{i\bullet} I_{j\bullet}}{I_{\bullet\bullet}}$$

- **Fast**
- Phylogentic correction included
- Cannot resolve indirect dependencies
- **Bad** prediction quality

plmDCA[2]

- Based on inference of a **Potts Model**:

$$P(\vec{a}^k) = \frac{1}{Z}\left(exp(\sum_i^N h(a_i^k) + \sum_{i<j}^{N,N} J_{ij}(a_i^k, a_j^k)\right)$$

- To infer the parameters, a **pseudo-likelihood maximization** is employed

$$\underset{h_\xi, J_{\xi,i}}{argmin}\left(-\frac{1}{M}\sum_k^M \log\left[P(a_\xi^k | \{a_{i\neq\xi}\})\right]\right)$$

- To avoid overfitting a **regularizer (l_2 norm)** is used when minimizing the objective function

$$R(\vec{h},\vec{J}) = \lambda_h \sum_k h_k^2 + \lambda_J \sum_{ij} J_{ij}^2$$

- **Best** prediction quality of the tested methods
- **Slowest** of tested methods

PSICOV[3]

- Computes **Partial Correlation Coefficients** θ_{ij}
- Inversion of the connected correlation matrix $C_{ij}(a_i, b_j)$ necessary
- Due to rank deficiency, **Sparse Inverse Covariance Estimation** is employed
- **Graphical Lasso Method**

$$\underbrace{\sum_{ij} C_{ij}\theta_{ij} - \log(\det\theta)}_{\text{Negative log-likelihood in Gaussian Model}} + \overbrace{\rho \sum_{ij} |\theta_{ij}|}^{Regularizer}$$

- Relatively **slow**
- **Very good** prediction quality (PPV = 1 for the first predicted pair in the tested set)
- Some problems with **convergence**

Bayesian Networks[4]

$$P(D) = \overbrace{\sum_\pi \underbrace{P(D|\pi)}_{\text{Assume Bayesian Tree}} \underbrace{P(\pi)}_{\text{Decomposable Prior}}}^{\text{Compute by generalized KMT}}$$

- Computational complexity: Compuation of a determinant
- **Medium** Prediction Quality

Gaussian DCA *(publication in prep.)*

- Amino acid frequencies and connected correlations are recast as expectation values of binary variables
- Sequence **reweighting**
- The parameters of a multi-variate Gaussian are estimated, treating expectation variables of the binary variables as expectation values of real-valued variables
- A full Bayesian with a **normal-inverse-Wishart prior** is employed
- **Fast**
- **Good** prediction quality

Hopfield-Potts[6]

- The couplings of the Potts-Model become a combination of **patterns** ξ

$$J_{ij}(a_i, a_j) = \sum_\mu^p \xi_{ia}^\mu \xi_{jb}^\mu$$

- Inference is done using a **maximum-likelihood** approach
- The patterns obtained are the **eigenvectors** of a modified version of the **correlation matrix**
- Patterns can be **attractive** (real-valued) or **repulsive** (imaginary)
- Computational complexity: **eigenanalysis**
- Prediction quality is **good**

References

1 Dunn, Stanley D., Lindi M. Wahl, and Gregory B. Gloor. "Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction." Bioinformatics 24.3 (2008): 333-340.

2 Ekeberg, Magnus, et al. "Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models." Physical Review E 87.1 (2013): 012707.

3 Jones, David T., et al. "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments." Bioinformatics 28.2 (2012): 184-190.

4 Burger, Lukas, and Erik van Nimwegen. "Disentangling direct from indirect co-evolution of residues in protein alignments." PLoS Computational Biology 6.1 (2010): e1000633.

6 Cocco, Simona, Remi Monasson, and Martin Weigt. "From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction." arXiv preprint arXiv:1212.3281 (2012).

RESULTS