Unsupervised learning of AA sequences

- Multiple sequence alignment [3] is an algorithm that uses many sample sequences of related proteins to infer residue contact
- Residue pairs that are consistent across sequences indicate that those residues may be in close contact (evolutionary covariation)
- Conversely, residues pairs that are uncorrelated are unlikely to be in contact

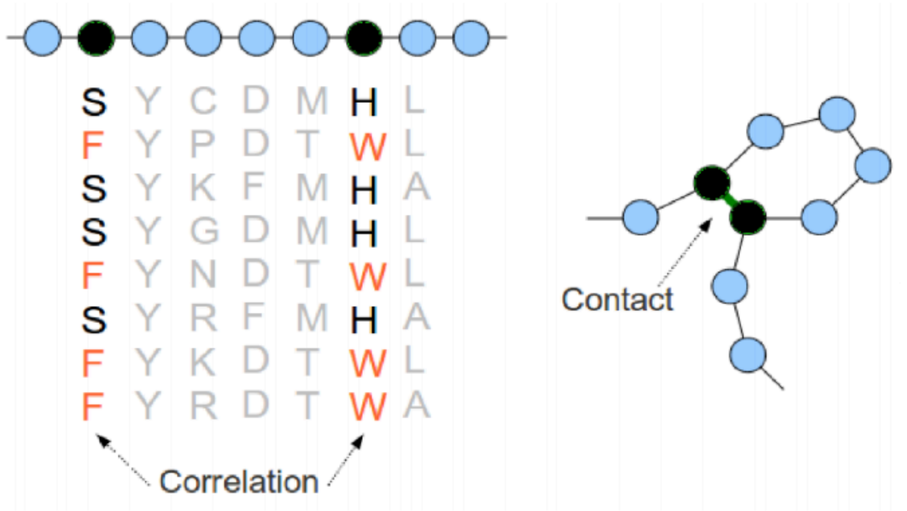

FIG. 1. (Color online) Left panel: small MSA with two positions of correlated amino-acid occupancy. Right panel: hypothetical corresponding spatial conformation, bringing the two correlated positions into direct contact.

[3] Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models

Supervised learning for predicting inter-residue distance

- Use databases of protein sequences with known structures (ie. known inter-residue distances)
- Build a supervised learning model that learns the relationship between amino acid sequences and inter-residue distance
- Predict inter-residue distance for protein sequences with unknown structures