

- **Computational Intensity:** Demanding molecular dynamics simulations.
- **Incomplete Modeling:** Often neglects complex environmental interactions.

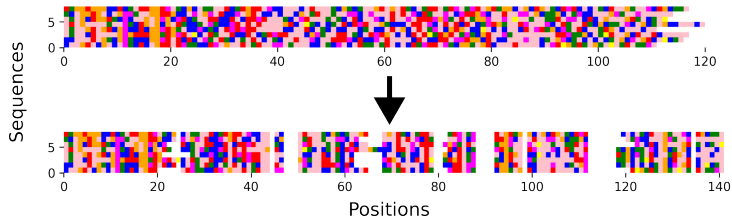
- **Evolutionary optimization:** Natural proteins are optimized to the physical diffusion limits.
- **Homology Modeling:** Borrowing structures from evolutionarily related proteins.
- **Evolutionary Couplings:** Pinpointing residues crucial for function.
- **Advantages:**
 - Sidestep computational hurdles.
 - Tap into nature's tried-and-tested designs.

**Given a target natural function, search for
natural counterparts**

- ① Search natural counterparts for a targeted function.
- ② Extract statistical signature from the collection of natural sequences.
- ③ Use the statistical signature to sample novel sequences.

Multiple Sequence Alignments (MSAs) - the data

- Definition: Aligning multiple sequences to identify regions of similarity.
- Importance in bioinformatics:
 - Studying phylogenetics and evolutionary processes.
 - Identifying protein domains.



- MSA is discrete qualitative data type

How to use them?

Encoding

- One-hot:

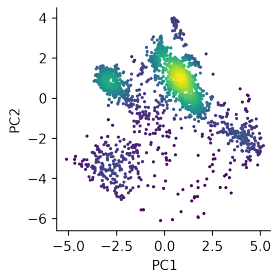
$$A \rightarrow \begin{pmatrix} 1 \\ 0 \\ \dots \\ 0 \\ 0 \end{pmatrix}, C \rightarrow \begin{pmatrix} 0 \\ 1 \\ \dots \\ 0 \\ 0 \end{pmatrix}, \dots Y \rightarrow \begin{pmatrix} 0 \\ 0 \\ \dots \\ 1 \\ 0 \end{pmatrix}, W \rightarrow \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 1 \end{pmatrix}.$$

- Random projection
- Deep learning embeddings

Since we have numerical data, we can also use dimensionality reduction techniques

Singular Value Decomposition (SVD) for Protein Data

- **Application:** Extracting meaningful patterns from vast protein datasets.
- **Dimensionality Reduction:** Simplifies complex data, retaining essential information.
- **Pattern Recognition:** Reveals underlying structures and relationships in protein data.



Introduction

- Fundamental technique in linear algebra.
- Decomposes a matrix into three other matrices.
- Widely used in data compression, noise reduction, and more.

Mathematical Representation

- Given a matrix MSA :

$$MSA = U\Sigma V^T$$

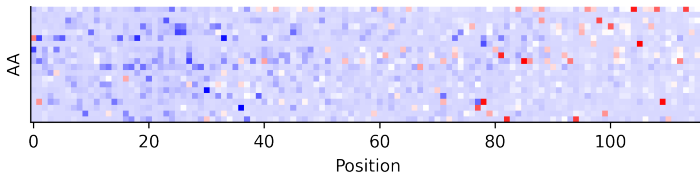
Where:

- U - Left singular vectors (orthogonal matrix).
- Σ - Diagonal matrix of singular values.
- V^T - Transpose of right singular vectors (orthogonal matrix).

What's in those singular vectors ?

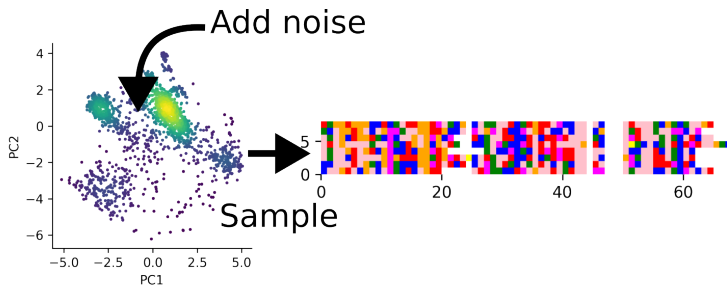
*

- The right singular vectors correspond to compositional motifs (in terms of sequences).



**Sample the compositional motifs observed
in the MSA to form novel sequences**

- Concept of reverse mapping: Generating functional sequences from reduced-dimensional data.



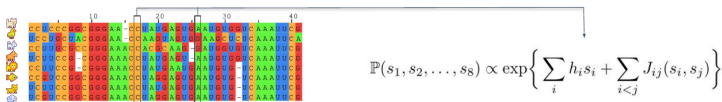
Introduce a Gaussian blank noise to sample the PCs:

$$\tilde{U} = U + \mathcal{N}(0, 1)$$
$$\hat{M}SA = \tilde{U}\Sigma V^T$$

Direct coupling analysis

F. Morcos *et al*, PNAS 2011

- Parametrize a probability distribution describing the distribution of sequences.
- Decompose the complex distribution of sequences into a pairwise potential — the Potts model.



MSA probabilistic model [Morcos *et al*, PNAS, 2011]

- Probability associated to a Sequence given a MSA:

$$P_{\mathcal{H}}(S) \propto \exp\{-\beta \times \mathcal{H}(S)\}$$

- Energy of a sequence (Potts models):

$$\mathcal{H}(S) = \sum_i h_i(S_i) + \sum_{i < j} J_{ij}(S_i, S_j)$$

- Energy parameters:

- $\mathcal{H} = \{h_i; J_{ij}\}$ (lookup table)
- Parameter space: $5 \times L + 5^2 \times \frac{L \times (L-1)}{2} = 464165$

Contact predictions based on coupling terms J_{ij} :

$$F_{ij} = \sqrt{\sum J_{ij}(A, B)^2} \rightarrow F_{ij}^{APC} = F_{ij} - \frac{F_{i.} F_{.j}}{F_{..}}$$

Turn into an optimization procedure

- Fit low-order statistics such as f_i and f_{ij} : Find \mathcal{H} such that:

$$\hat{f}_i(A) = f_i(A) \quad ; \quad \hat{f}_{ij}(A, B) = f_{ij}(A, B)$$

Boltzmann machine learning [Figliuzzi *et al*, Mol. Biol Ev., 2018; Cuturello *et al*, RNA, 2020]

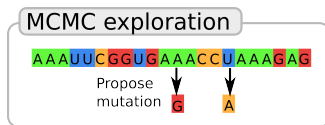
Initialize with a guess for \mathcal{H} (could be zeros)

- 1 Generate a sample given \mathcal{H} (MCMC) and compute \hat{f}_i, \hat{f}_{ij}
- 2 \mathcal{H} parameters are updated following the **log-likelihood**

$$h_i(A) \leftarrow h_i(A) + \eta(\hat{f}_i(A) - f_i(A))$$

- Sampling sequences: sampling new protein variants.
- Ensuring biological relevance: Satisfying coevolutionary constraints.

Perform mutations:



Select using the parametrized distribution:

$$P_{\mathcal{H}}(S) \propto \exp\{-\beta \times \mathcal{H}(S)\}$$