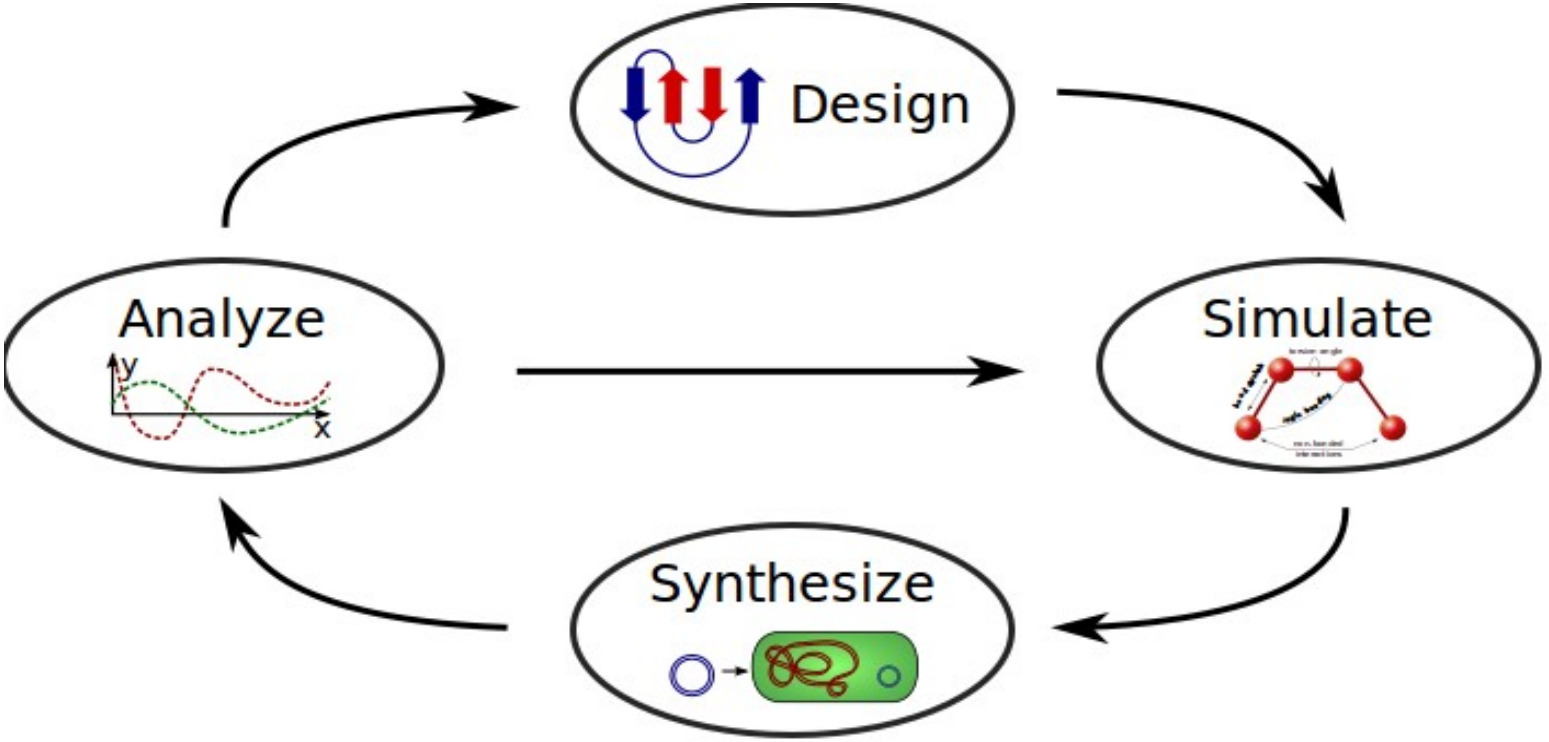
Pareto-Optimal Multistate Protein Design Using Tree Graphical Models

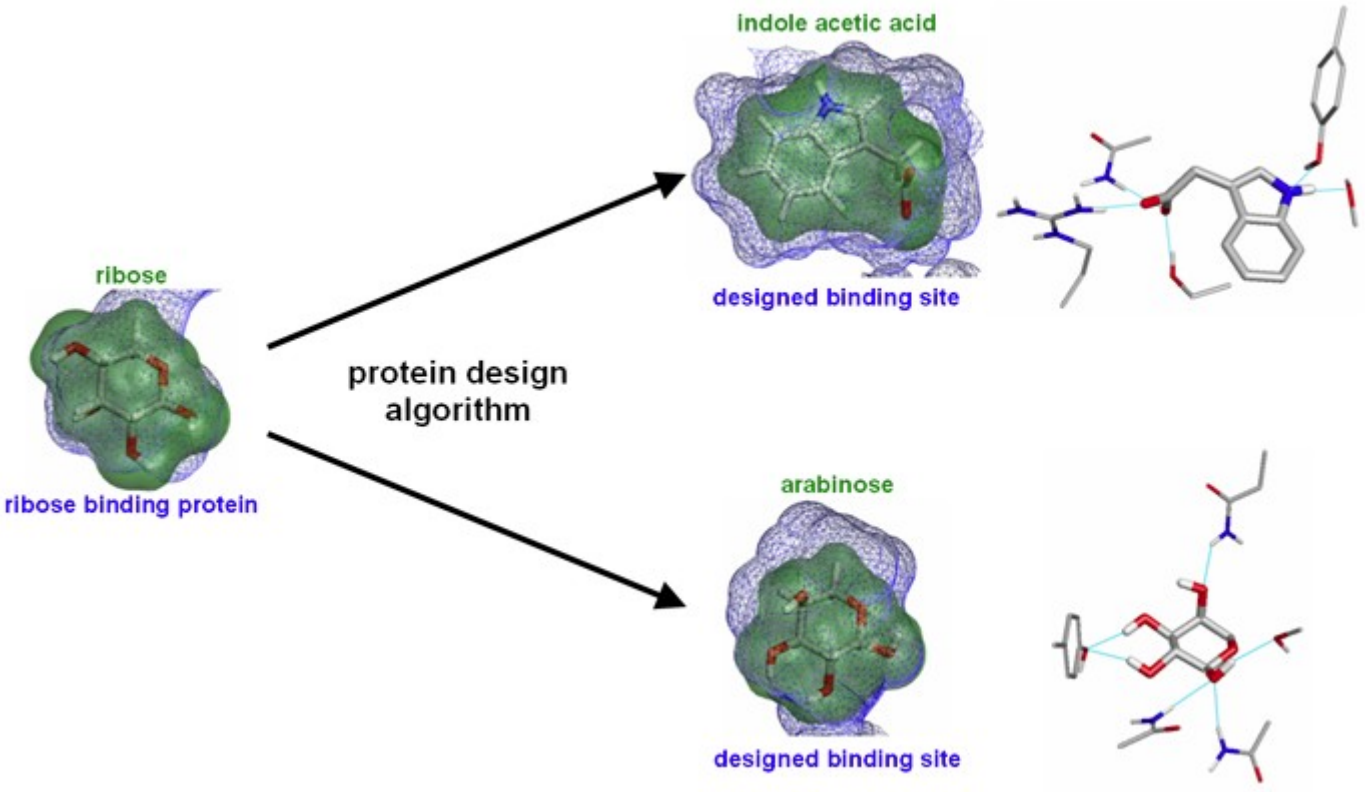
- *Subhodeep Moitra*

Protein Engineering Cycle



Source : itooktheredpill.dyndns.org

Protein Design



Source : http://www.stanford.edu/~boas/science/protein_design/4.png

Theory

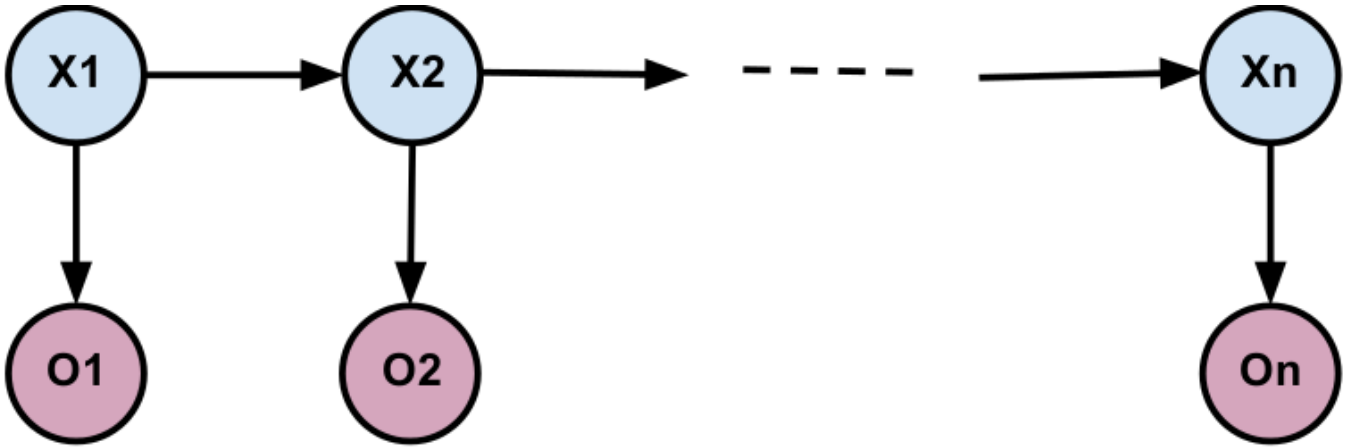


Figure 1: A canonical HMM

The graphical representation of the HMM factorizes according to the following probability distribution :

$$P(X_1, \ldots, X_n | \mathbf{O_1}, \ldots, \mathbf{O_n}) = P(X_1) \prod_{i=1}^{n} P(\mathbf{O_i} | X_i) \prod_{i=1}^{n-1} P(X_{i+1} | X_i) \qquad (1)$$

Calculation of the Optimal Probability value:

$$V(i,k) = \begin{cases} P(\mathbf{O_1}|X_1{=}k)P(X_1{=}k) & \text{if } i = 1 \\ P(\mathbf{O_i}|X_i{=}k)\max_j\{P(X_i{=}k|X_{i-1}{=}j)V(i-1,j)\} & \text{if } 2 \leq i \leq n \end{cases}$$

Retrieval of the optimal sequence corresponding to the optimal probability value:

$$X_n^\star = \arg\max_k V(n,k)$$

$$X_i^\star = Ptr(X_{i+1}^\star, i+1) \qquad \text{if } 1 \leq i \leq n-1$$

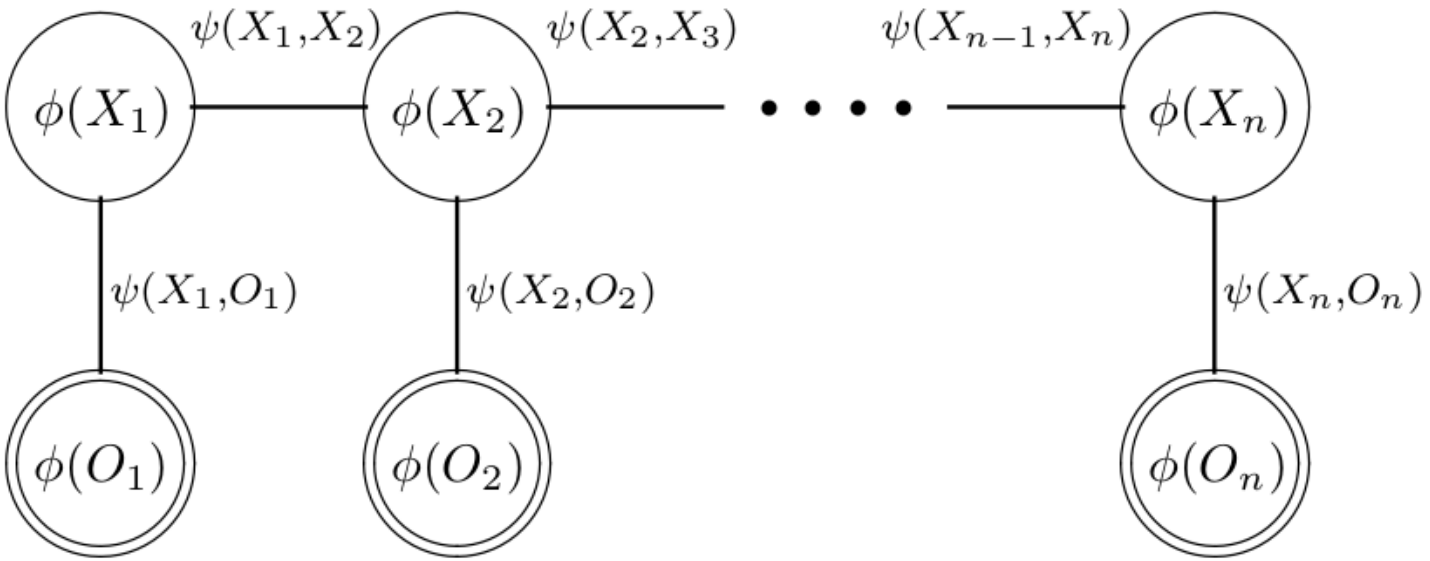


Figure 2: A chain structured Markov Random Field. Probability of a sequence assignment is defined as a function of the node and edge potentials

The probability of the a sequence assignment is defined as follows :

$$P(X_1, X_2, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{n} \phi(X_i)\phi(\mathbf{O_i})\psi(\mathbf{O_i}, X_i) \prod_{i=1}^{n-1} \psi(X_i, X_{i+1})$$

$$\text{where} \quad Z = \sum_{X_1, X_2, \ldots, X_n} \left[\prod_{i=1}^{n} \phi(X_i)\phi(\mathbf{O_i})\psi(\mathbf{O_i}, X_i) \prod_{i=1}^{n-1} \psi(X_i, X_{i+1}) \right]$$

The main steps in the max-product algorithm are :

Calculation of the Optimal Probability value:

$$V(i,k) = \begin{cases} \phi(\mathbf{O_1})\psi(\mathbf{O_1}, X_1{=}k)\phi(X_1{=}k) & \text{if } i = 1 \\ \phi(\mathbf{O_i})\phi(X_i{=}k)\psi(\mathbf{O_i}, X_i{=}k)\max_j\{\psi(X_i{=}k, X_{i-1}{=}j)V(i-1,j)\} & \text{if } 2 \leq i \leq n \end{cases}$$

The main steps in the max-product algorithm are :

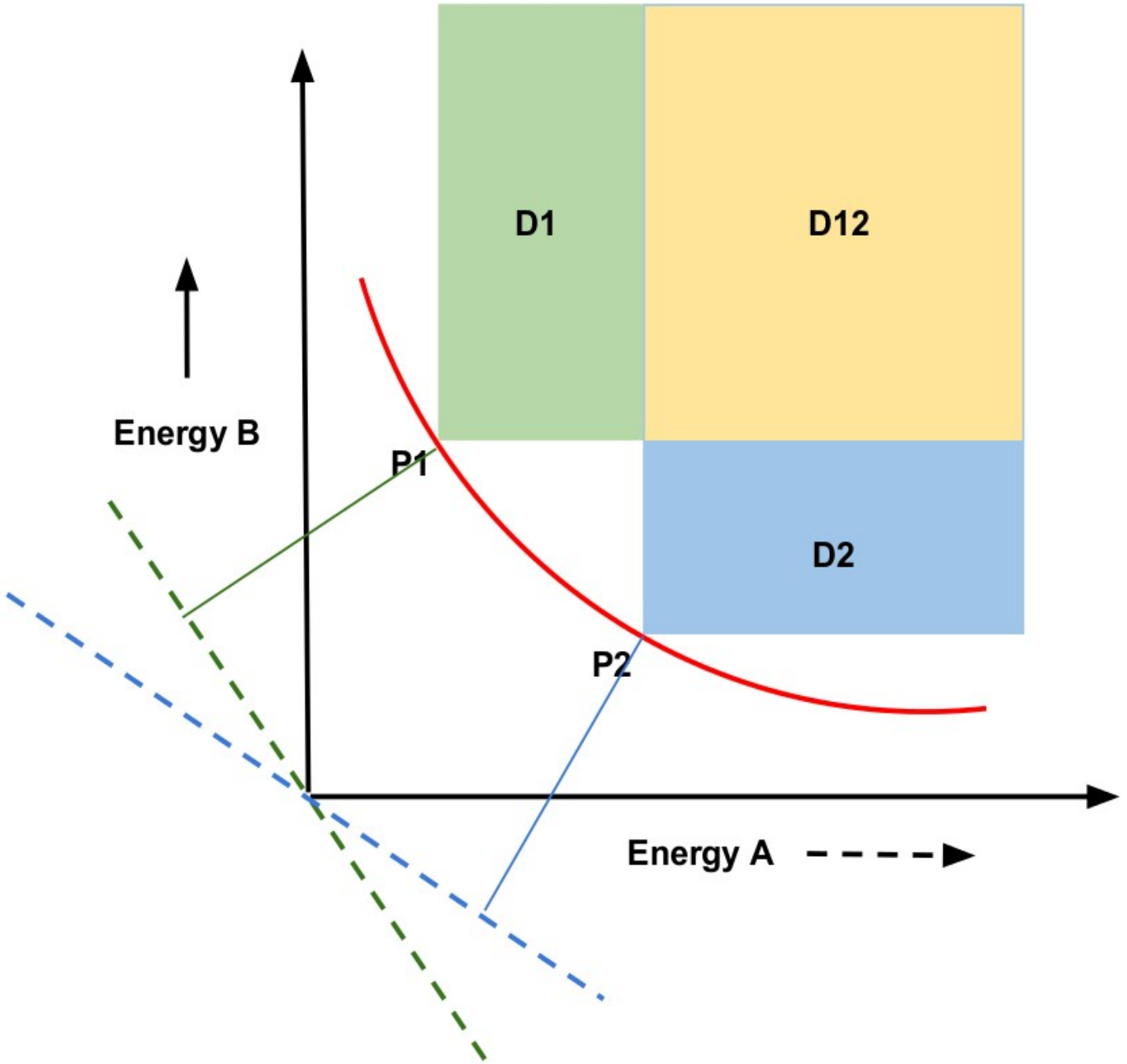
Calculation of the Optimal Probability value:

$$V(i,k) = \begin{cases} \phi(\mathbf{O_1})\psi(\mathbf{O_1}, X_1{=}k)\phi(X_1{=}k) & \text{if } i = 1 \\ \phi(\mathbf{O_i})\phi(X_i{=}k)\psi(\mathbf{O_i}, X_i{=}k)\max_j\{\psi(X_i{=}k, X_{i-1}{=}j)V(i-1,j)\} & \text{if } 2 \leq i \leq n \end{cases}$$

Retrieval of the optimal sequence corresponding to the optimal probability value:

$$X_n^\star = \arg\max_k V(n,k)$$

$$X_i^\star = Ptr(X_{i+1}^\star, i+1) \qquad \text{if } 1 \leq i \leq n-1$$


D1
D12
D2
Energy B
P1
P2
Energy A

For multi-state protein design we can combine the energy functions as follows.

$$E_{AB}(X) = \theta_A E_A(X) + \theta_B E_B(X) \quad \text{where} \quad \theta_A, \theta_B \geq 0 \text{ and } \theta_A + \theta_B = 1$$

Here $E_A(X)$ is any energy function mapping sequence to energy values for state A. The lower the energy the better the fit for sequence X to state A. $E_A(X)$ can also be written as a CMRF.

$$E_A(X) = -\log \left[\prod_{i=1}^{n} \phi(X_i)\phi(\mathbf{O_i^A})\psi(\mathbf{O_i^A}, X_i) \prod_{i=1}^{n-1} \psi(X_i, X_{i+1}) \right]$$

$$
\begin{aligned}
X_A &= \arg\min_X E_A(X) \\
&= \arg\min_X -\log\left[\prod_{i=1}^{n} \phi(X_i)\phi(\mathbf{O_i^A})\psi(\mathbf{O_i^A}, X_i) \prod_{i=1}^{n-1} \psi(X_i, X_{i+1})\right] \\
&= \arg\min_X -\left[\sum_{i=1}^{n} \log\phi(X_i) + \sum_{i=1}^{n} \log\phi(\mathbf{O_i^A}) + \sum_{i=1}^{n} \log\psi(\mathbf{O_i^A}, X_i) + \sum_{i=1}^{n-1} \log\psi(X_i, X_{i+1})\right] \\
&= \arg\min_X -\left[\sum_{i=1}^{n} \log\phi(X_i) + \sum_{i=1}^{n} \log\psi(\mathbf{O_i^A}, X_i) + \sum_{i=1}^{n-1} \log\psi(X_i, X_{i+1})\right]
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
X_B &= \arg\min_X E_B(X) \\
&= \arg\min_X -\left[\sum_{i=1}^{n} \log\phi(X_i) + \sum_{i=1}^{n} \log\psi(\mathbf{O_i^B}, X_i) + \sum_{i=1}^{n-1} \log\psi(X_i, X_{i+1})\right]
\end{aligned}
$$

Taking the convex combination of the energy functions we get

$$X_{AB} = \arg\min_{X} \theta_A E_A(X) + \theta_B E_B(X)$$
$$\text{s.t} \quad \theta_A + \theta_B = 1$$
$$\theta_A, \theta_B \geq 0$$

Taking the convex combination of the energy functions we get

$$X_{AB} = \arg\min_X \theta_A E_A(X) + \theta_B E_B(X)$$

$$\text{s.t} \quad \theta_A + \theta_B = 1$$

$$\theta_A, \theta_B \geq 0$$

$$X_{AB} = \arg\min_X -\theta_A \left[\sum_{i=1}^{n} \log \phi(X_i) + \sum_{i=1}^{n} \log \psi(\mathbf{O_i^A}, X_i) + \sum_{i=1}^{n-1} \log \psi(X_i, X_{i+1}) \right]$$

$$- \theta_B \left[\sum_{i=1}^{n} \log \phi(X_i) + \sum_{i=1}^{n} \log \psi(\mathbf{O_i^B}, X_i) + \sum_{i=1}^{n-1} \log \psi(X_i, X_{i+1}) \right]$$

Taking the convex combination of the energy functions we get

$$X_{AB} = \arg\min_X \theta_A E_A(X) + \theta_B E_B(X)$$

$$\text{s.t} \quad \theta_A + \theta_B = 1$$

$$\theta_A, \theta_B \geq 0$$

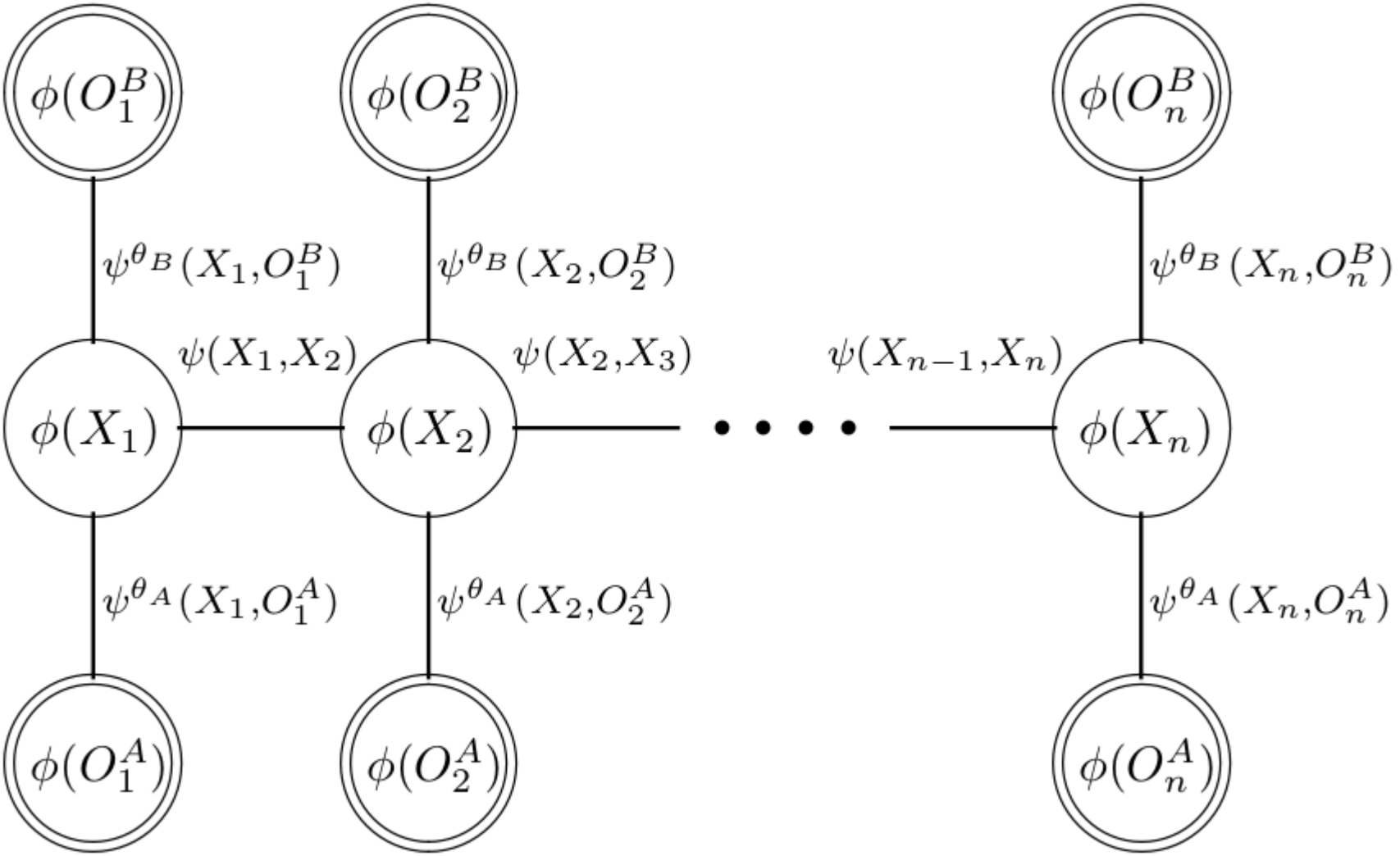
$$X_{AB} = \arg\min_X -\theta_A \left[\sum_{i=1}^{n} \log \phi(X_i) + \sum_{i=1}^{n} \log \psi(\mathbf{O_i^A}, X_i) + \sum_{i=1}^{n-1} \log \psi(X_i, X_{i+1}) \right]$$

$$- \theta_B \left[\sum_{i=1}^{n} \log \phi(X_i) + \sum_{i=1}^{n} \log \psi(\mathbf{O_i^B}, X_i) + \sum_{i=1}^{n-1} \log \psi(X_i, X_{i+1}) \right]$$

$$= \arg\min_X - \left(\sum_{i=1}^{n} \log \phi(X_i) + \sum_{i=1}^{n-1} \log \psi(X_i, X_{i+1}) \right)$$

$$- \left(\sum_{i=1}^{n} (\log \psi(\mathbf{O_i^A}, X_i)^{\theta_A} + \log \psi(\mathbf{O_i^B}, X_i)^{\theta_B}) \right)$$

We get – Tree MRF(TMRFs)



The main steps in the max-product algorithm are :

Calculation of the Optimal Probability value:

$$V(i,k) = \begin{cases} \psi(\mathbf{O_1^A}, X_1{=}k)^{\theta_A}\psi(\mathbf{O_1^B}, X_1{=}k)^{\theta_B}\phi(X_1{=}k) & \text{if } i = 1 \\ \phi(X_i{=}k)\psi(\mathbf{O_i^A}, X_i{=}k)^{\theta_A}\psi(\mathbf{O_1^B}, X_1{=}k)^{\theta_B}\cdots & \\ \max_j\{\psi(X_i{=}k, X_{i-1}{=}j)V(i-1,j)\} & \text{if } 2 \leq i \leq n \end{cases}$$

Retrieval of the optimal sequence corresponding to the optimal probability value:

$$X_n^\star = \arg\max_k V(n,k)$$

$$X_i^\star = Ptr(X_{i+1}^\star, i+1) \qquad \text{if } 1 \leq i \leq n-1$$

Algorithm 2 Pareto-Optimal MultiState Protein Design

procedure PARETO-FRONTIER(ϕ, ψ, O^A, O^B) ▷ Find Pareto frontier
  Initialize queue $Q \leftarrow \emptyset$
  $X_A \leftarrow$ CMRF-DECODE(ϕ, ψ, O^A) ▷ $\arg\min_X E_A(X)$
  $X_B \leftarrow$ CMRF-DECODE(ϕ, ψ, O^B) ▷ $\arg\min_X E_B(X)$
  Enqueue (X_A, X_B)
  $C_H \leftarrow \{X_A, X_B\}$
  repeat
    Dequeue from Q a pair (X_1, X_2) and assert$(E_A(X_1) < E_A(X_2))$
    $m \leftarrow \frac{E_B(X_1) - E_B(X_2)}{E_A(X_1) - E_A(X_2)}$
    $\theta_A \leftarrow \frac{-m}{1-m}$
    $\theta_B \leftarrow \frac{1}{1-m}$
    $X_{AB} \leftarrow$TMRF-DECODE($\phi, \psi, O^A, O^B, \theta_A, \theta_B$)
      ▷ $\arg\min_X \theta_A E_A(X) + \theta_B E_B(X)$
    if $X_{AB} \neq X_A$ and $X_{AB} \neq X_B$ **then**
      $C_H \leftarrow C_H \cup \{X_{AB}\}$
      Enqueue (X_A, X_{AB}) and (X_{AB}, X_B)
    end if
  until Q is empty
  return C_H
end procedure

procedure CMRF-DECODE(ϕ, ψ, O^A) ▷ Max decode CMRF

  Allocate $V[N][K]$ ▷ N is length of seq,K is numAA

  Allocate $X_A[N]$

  for $k \leftarrow 1, K$ **do**

    $V[1][k] \leftarrow \psi(\mathbf{O_1^A}, X_1{=}k)\phi(X_1{=}k)$

  end for

  for $i \leftarrow 1, N$ **do**

    for $k \leftarrow 1, K$ **do**

      $V[i][k] \leftarrow \phi(X_i{=}k)\psi(\mathbf{O_i^A}, X_i{=}k) \max_j \{\psi(X_i{=}k, X_{i-1}{=}j)V(i-1, j)\}$

      $Ptr[i][k] \leftarrow \arg\max_j \{\psi(X_i{=}k, X_{i-1}{=}j)V(i-1, j)\}$

    end for

  end for

▷ Now retrieve the pointers

  $X_A[N] \leftarrow \arg\max_k V[N][k]$

  for $i \leftarrow N-1, 1$ **do**

    $X_A[i] \leftarrow Ptr[i+1][X_A[i+1]]$

  end for

  return X_A

end procedure

Summary

Contributions

A fast and exact protein design method

Flexibility in choosing features

Enumerate the Pareto Frontier

Limitations

Sequential dependency between variables

Thank You