Favorable-Contact Model of Folding

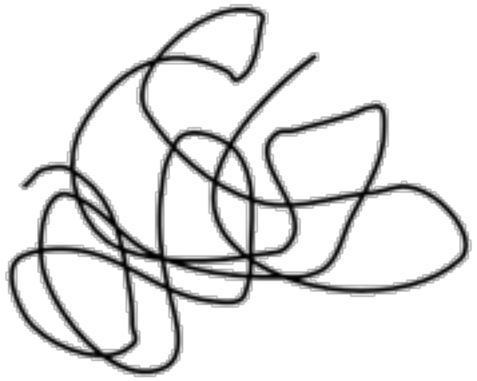
Mobolaji Williams — Shakhnovich Group Meeting— Mar. 21, 2017

Connecting Protein Folding and Protein Design

Design and Folding Problems

How do we model the way
structure determines sequence?

Study a model of possible sequence orders
for a given presumed structure



Structure

Protein Design

→

···A-R-H-G-L-H···

Sequence

←

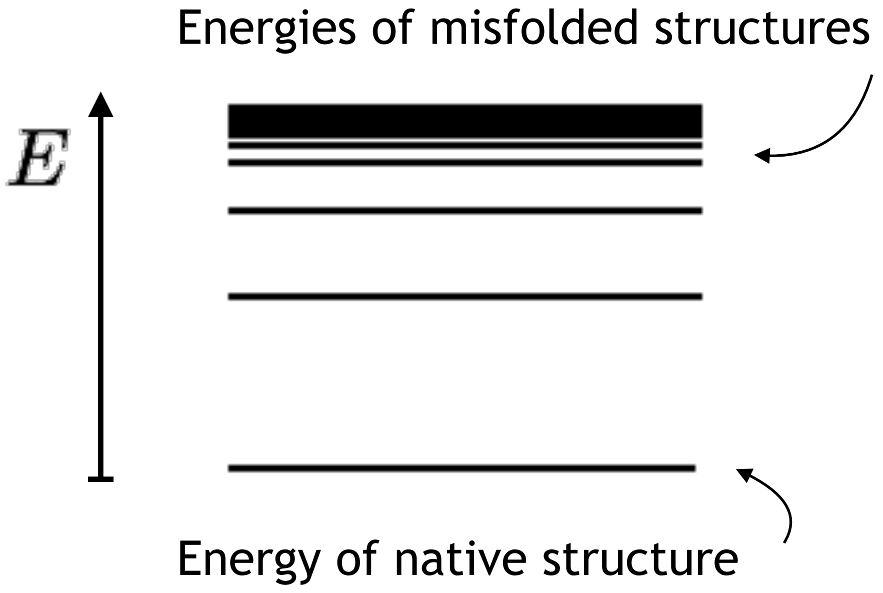
Protein Folding

How do we model the way
sequence determines structure?

Study a model of possible pairwise contacts
for a given sequence of contact regions

(Abstracted) Protein Design

Two-State Folding Model



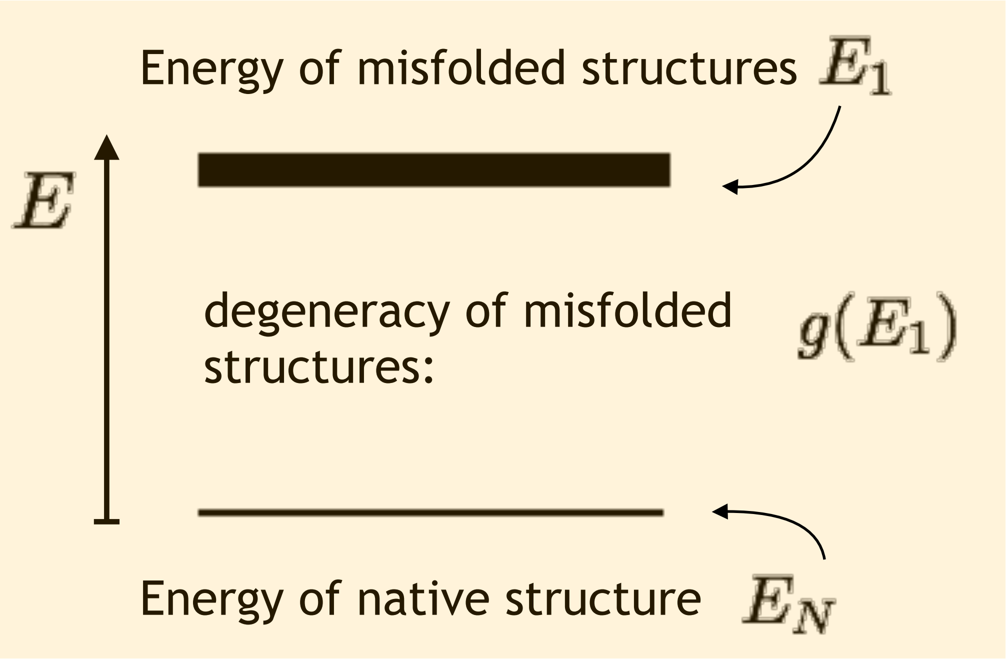
Energies of misfolded structures

E

Energy of native structure

(For structurally simple proteins)

Approximate as a two-state model



Energy of misfolded structures E_1

E

degeneracy of misfolded structures: $g(E_1)$

Energy of native structure E_N

Stability of Native Structure

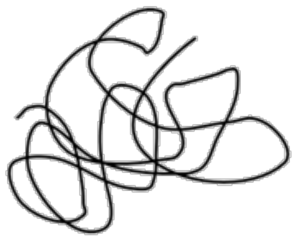
~ (Boltzmann Probability to be in Native Structure)

$$P_{\text{native}} = \frac{1}{1 + g(E_1)e^{-\beta(E_1 - E_N)}}$$

In order for sequence to properly fold the native structure must be highly stable (i.e., P_{native} must be as close to 1 as possible)

$E_1 - E_N \gg k_B T$: Energy gap between native and misfolded structures should be large

To find the sequence which maximizes energy gap, we **search over sequence space** for a given structure



Structure

$$\longrightarrow \quad \begin{array}{ll} \mathcal{S}_1 : & R - D - \cdots - L \\ \mathcal{S}_2 : & D - R - \cdots - L \\ \vdots & \qquad\quad \vdots \\ \mathcal{S}_{N_s} : & V - Y - \cdots - N \end{array}$$

Possible AA Sequences

Protein Folding and Design

Problem with formalism:

> *"For virtually any target structure, the lowest energy sequence will be a homopolymer, consisting of the amino acid with the largest self-attraction"**

*(Morrissey and Shakhnovich. "Design of proteins with selected thermal properties." (1996))

$$\longrightarrow \quad I - \cdots - I - \cdots I$$

Structure

Homopolymer?

This cannot be correct…

Suggested way forward:

Constrain amino acid composition

New Design Question:
Given **a fixed number of each type of amino-acid**
What sequence yields the lowest energy?

Implication of New Design Question

For a chain of length *N*, what is the size of our state space?

✗ 20^N (or 2^N under a polar/nonpolar framing)

✓ $\dfrac{N!}{n_1!n_2!\cdots n_{20}!}$ (or $\dfrac{N!}{n_{\text{polar}}!n_{\text{nonpolar}}!}$ under a polar/nonpolar framing)

We should search over the space of **permutations** of components

Partition Function of Permutations

Let’s define the system more precisely and introduce a Hamiltonian.

System Definition:

- subunits are labeled as ω_i with $i = 1, \dots, N$
- the ordering of subunits with the zero energy is $\vec{\omega} \equiv (\omega_1, \omega_2, \dots, \omega_N)$
- an arbitrary state is $\vec{\theta}$ where $\vec{\theta} \in \{\mathrm{perm}(\omega_1, \omega_2, \dots, \omega_N)\} \equiv Sym(\omega)$

Energy Definition:

- The state $\vec{\theta} = \vec{\omega}$ has zero energy and has subunits in the *correct* order. For all other states, there is an energy cost of λ_i for $\theta_i \neq \omega_i$.

$$\mathcal{H}_N(\{\theta_i\}) = \sum_{i=1}^{N} \lambda_i I_{\theta_i \neq \omega_i}$$

$$I_A = \begin{cases} 1 & A \text{ is true} \\ 0 & A \text{ is false} \end{cases}$$

Example: Three components

State	Energy
$(\omega_1, \omega_2, \omega_3)$	0
$(\omega_2, \omega_1, \omega_3)$	$\lambda_1 + \lambda_2$
$(\omega_3, \omega_2, \omega_1)$	$\lambda_1 + \lambda_3$
$(\omega_1, \omega_3, \omega_2)$	$\lambda_2 + \lambda_3$
$(\omega_2, \omega_3, \omega_1)$	$\lambda_1 + \lambda_2 + \lambda_3$
$(\omega_3, \omega_1, \omega_2)$	$\lambda_1 + \lambda_2 + \lambda_3$

Now let’s compute the partition function

$$Z_N(\{\beta\lambda_i\}) = \sum_{\vec{\theta} \in Sym(\omega)} \exp\left(-\beta \sum_{i=1}^{N} \lambda_i \, I_{\theta_i \neq \omega_i}\right)$$

$$\longrightarrow \boxed{Z_N(\{\beta\lambda_i\}) = \int_0^\infty ds\, e^{-s} \prod_{\ell=1}^{N} \left[1 + (s-1)e^{-\beta\lambda_\ell}\right]}$$

. . . so we can obtain a closed form expression . . .

. . .but what does it mean?

Model of Permutations

Energy Definition:

- The state $\vec{\theta} = \vec{\omega}$ has zero energy and has subunits in the *correct* order. For all other states, there is an energy cost of λ_i for $\theta_i \neq \omega_i$.

$$\mathcal{H}_N(\{\theta_i\}) = \sum_{i=1}^{N} \lambda_i I_{\theta_i \neq \omega_i}$$

Partition Function:

- $Z_N(\{\beta\lambda_i\}) = \int_0^\infty ds\, e^{-s} \prod_{\ell=1}^{N} \left[1 + (s-1)e^{-\beta\lambda_\ell}\right]$

What physics is contained in this partition function?

Progress comes from a simplification: $\lambda_i = \lambda_0 \quad \text{for all } i$ "The same energy penalty for each subunit"

The partition function then simplifies to

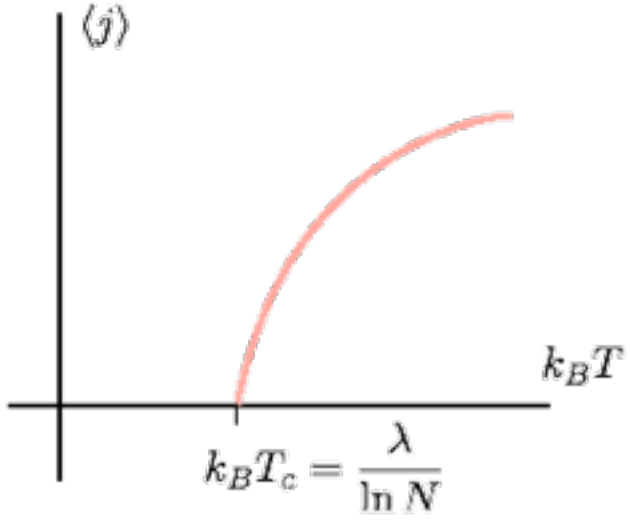
$$Z_N(\beta\lambda_0) = \int_0^\infty ds\, e^{-s} \left(1 + (s-1)e^{-\beta\lambda_0}\right)^N$$

There is a **transition temperature**

And the **average number of incorrect components** is

$$\langle j \rangle = \left\langle \sum_{i=1}^{N} I_{\theta_i \neq \omega_i} \right\rangle = \frac{\partial}{\partial(\beta\lambda_0)} \ln Z_N(\beta\lambda_0)$$

$$\longrightarrow \quad \boxed{\langle j \rangle \simeq \begin{cases} 0 & \text{for } T < T_c \\ N - e^{\beta\lambda_0} & \text{for } T > T_c \end{cases}}$$



→ Above a certain transition temperature, **there is a spectrum of sequences** (different from "correct sequence") **which yield the free energy minimum**

(Abstracted) Protein Folding

Motivation: N-mer Problems Lattice Proteins

We can study protein-folding through **lattice protein** models

Properties of Lattice Protein Models

- **Geometry** is made irrelevant (all proteins have the same shape)
- **Stability** is determined via contact densities and interaction energy between lattice sites
- **Primary Structure:** Sequence of monomers defined by their hydrophobicity
- **Secondary Structure:** None
- **Tertiary Structure:** Configuration of chain in the space of the cube

27-monomer lattice protein

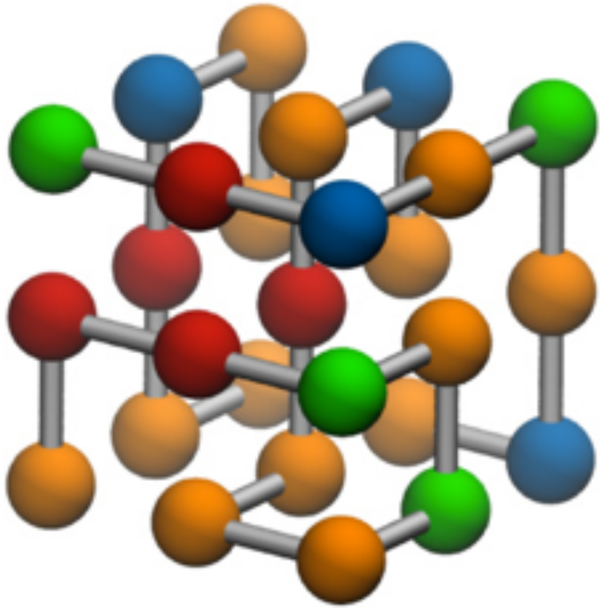


Figure from Whitford, Sanbonmatsu, and Onuchic.(2012)

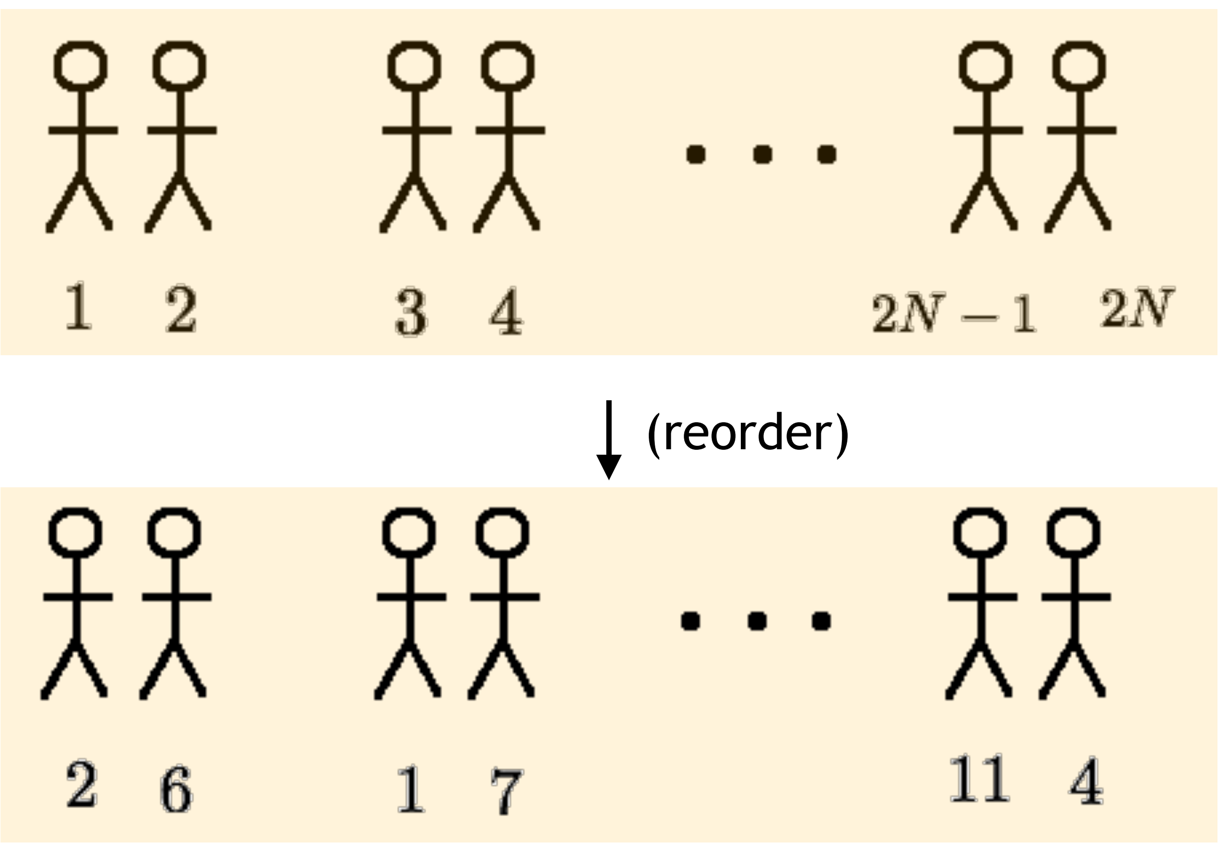
Configuration of chain in the 3X3X3 space represents a “folded” structure

Color represents hydrophobicity of monomer

Can we consider other statistical models of similar simplicity?

Dancing Partners Problem

- ➤ *N* partners (i.e., *2N* total people) arrive at a dancing party



- ➤ Each dance partner pair is re-ordered such that each person may or may not be with their original dance partner.

- ➤ In how many ways can this happen? (What is the size of the state space?)

$$|S_N| = \frac{(2N)!}{2^N N!}$$

Other Questions
- How many ways can re-order the dancing partners so that *k* pairs are not found in the original pairing?

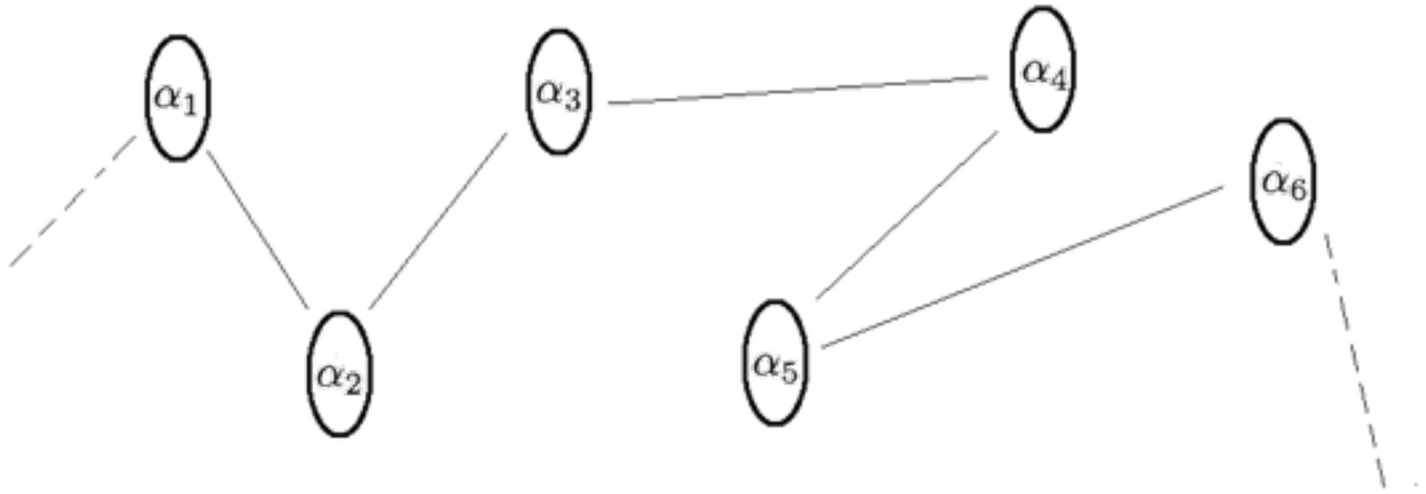
$$A_k = (-1)^k \sum_{\ell=0}^{k} (-1)^\ell \binom{k}{\ell} \prod_{i=1}^{\ell} (2i-1)$$

where $$\sum_{k=0}^{N} A_k = \frac{(2N)!}{2^N N!}$$

Main Point
There is an existing mathematical formalism to answer questions about rearranging an initial collection of pairs

Favorable Contact Model: Microstates

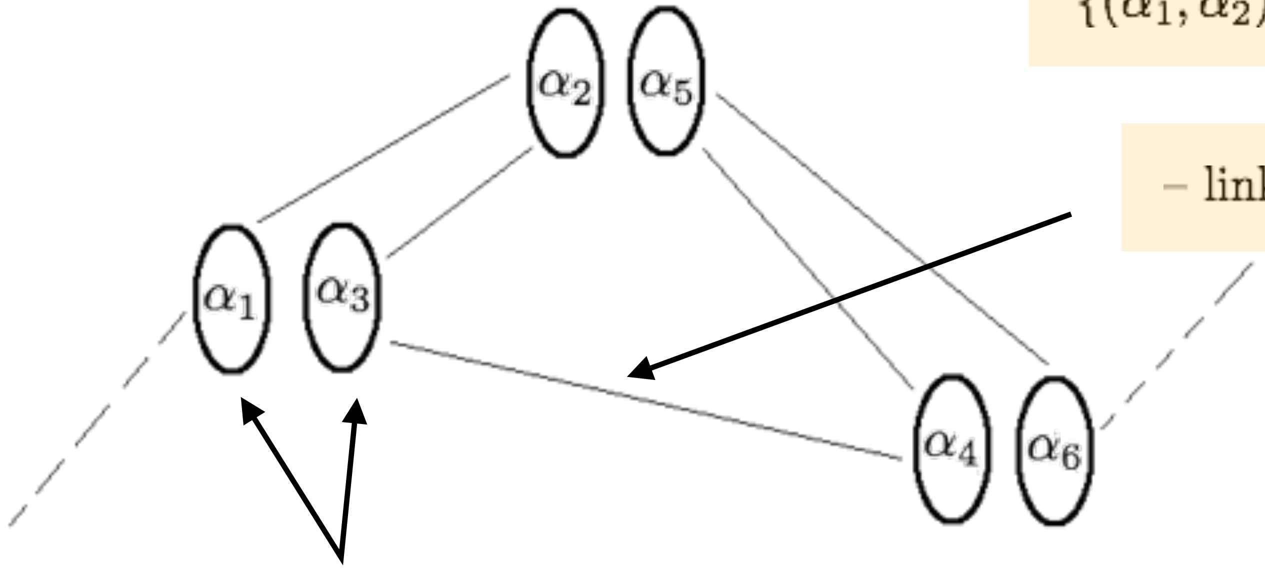
“Unfolded” configuration of 6 regions



Example of a microstate, i.e., a folded configuration of 6 regions

– regions are labeled as α_i with $i = 1, \dots, 2N$

– there is a single zero-energy (i.e., native) collection of pairwise contacts given by—for $2N$ regions—$\{(\alpha_1, \alpha_2), (\alpha_3, \alpha_4), \dots, (\alpha_{2N-1}, \alpha_{2N})\}$.



– links between regions are infinitely extendible

– only pairwise contacts between regions are possible

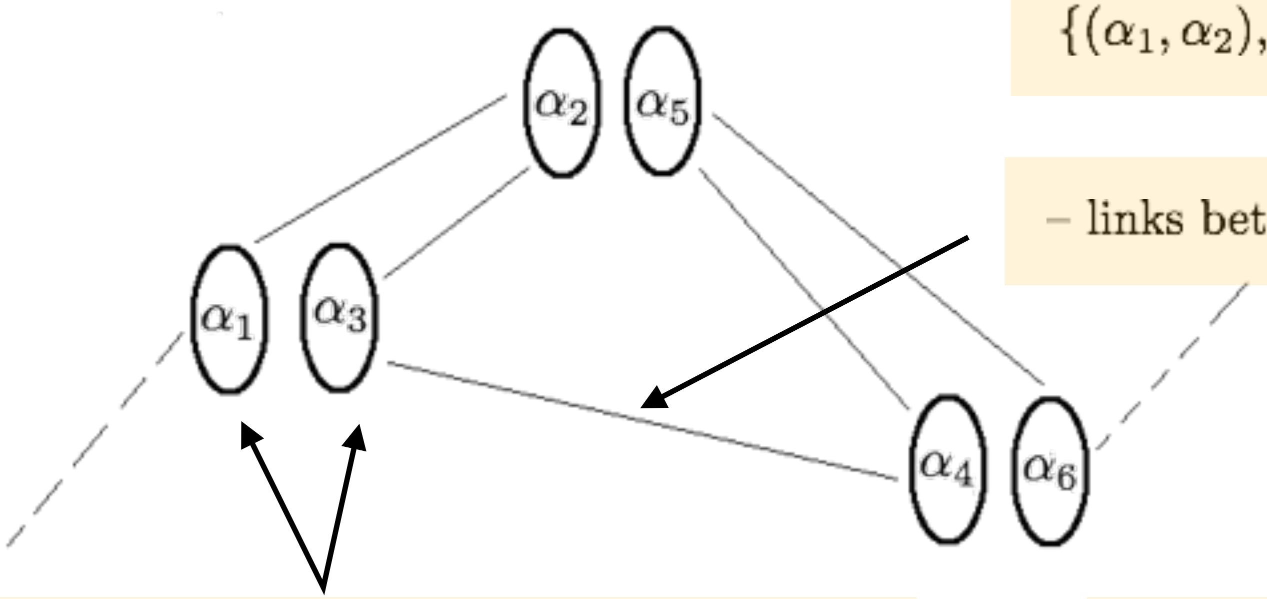
– microstates of system are given by collections of pairwise contacts (e.g., this microstate is $\{(\alpha_1, \alpha_3), (\alpha_2, \alpha_5), (\alpha_4, \alpha_6)\}$)

Favorable Contact Model: Energy

Example of a folded configuration of 6 regions

– regions are labeled as α_i with $i = 1, \ldots, 2N$

– there is a single zero-energy (i.e., native) collection of pairwise contacts given by—for $2N$ regions—$\{(\alpha_1, \alpha_2), (\alpha_3, \alpha_4), \ldots, (\alpha_{2N-1}, \alpha_{2N})\}$.



– links between regions are infinitely extendible

– only pairwise contacts between regions are possible

– microstates of system are given by collections of pairwise contacts (e.g., this microstate is $\{(\alpha_1, \alpha_3), (\alpha_2, \alpha_5), (\alpha_4, \alpha_6)\}$)

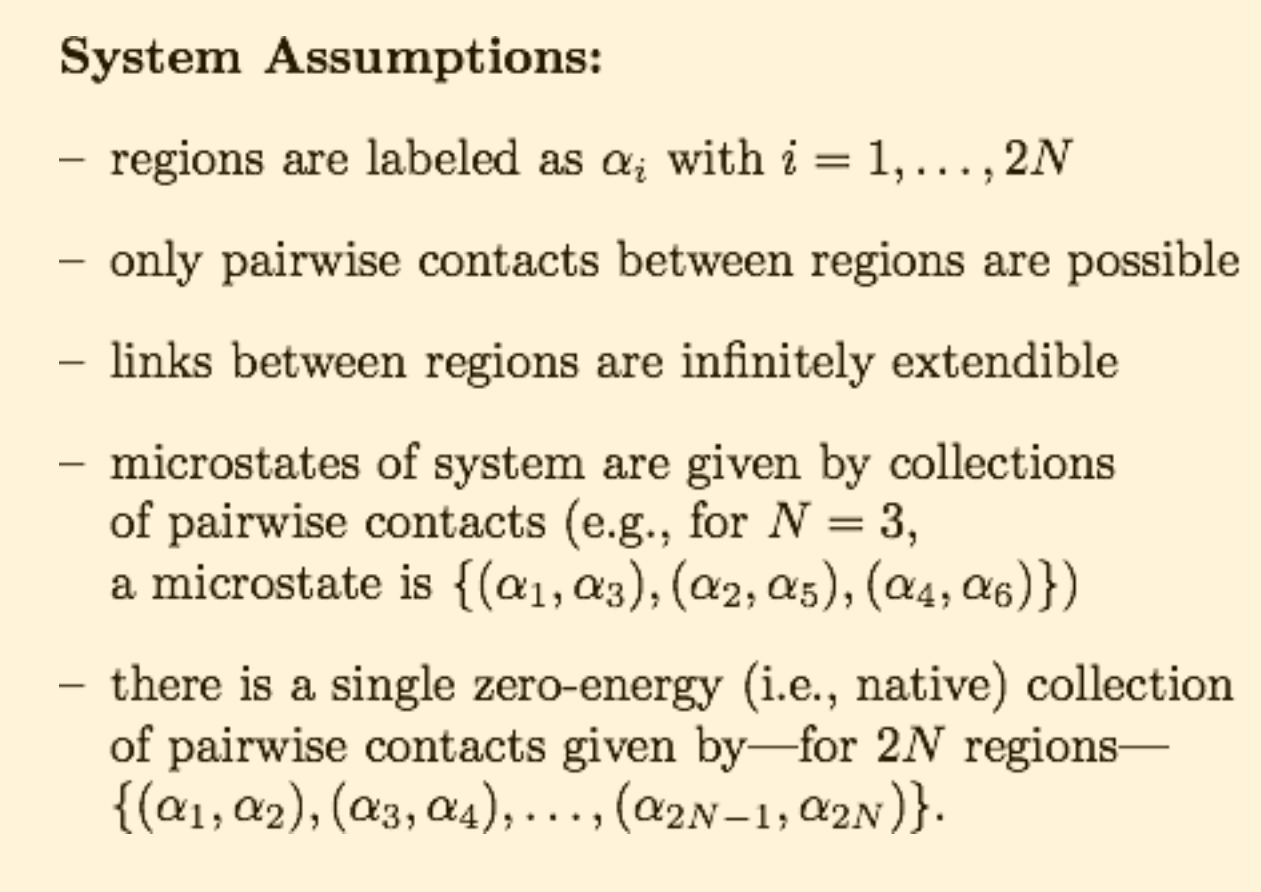
Energy Definition:

– The microstate $\{(\alpha_1, \alpha_2), (\alpha_3, \alpha_4), \ldots, (\alpha_{2N-1}, \alpha_{2N})\}$ has zero energy. For an arbitrary microstate, there is an energy cost of γ_i if the contact pair i is not an element of the set of pairs $\{(\alpha_1, \alpha_2), (\alpha_3, \alpha_4), \ldots, (\alpha_{2N-1}, \alpha_{2N})\}$.

Example: 6 components

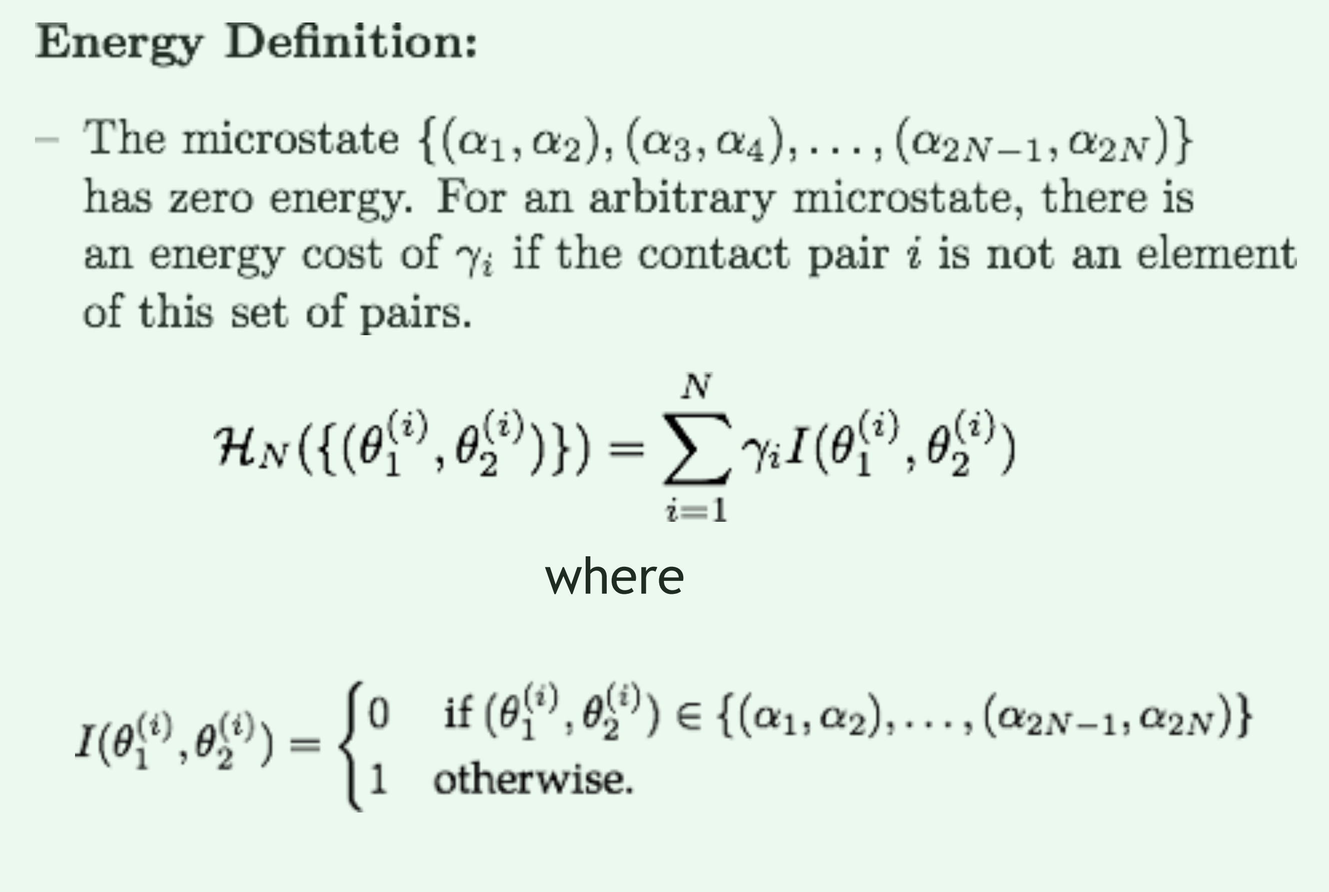
State	Energy
$(\alpha_1, \alpha_2), (\alpha_3, \alpha_4), (\alpha_5, \alpha_6)$	0
$(\alpha_1, \alpha_2), (\alpha_5, \alpha_4), (\alpha_3, \alpha_6)$	$\gamma_2 + \gamma_3$
$(\alpha_3, \alpha_6), (\alpha_1, \alpha_2), (\alpha_4, \alpha_5)$	$\gamma_1 + \gamma_3$
$(\alpha_1, \alpha_5), (\alpha_2, \alpha_4), (\alpha_3, \alpha_6)$	$\gamma_1 + \gamma_2 + \gamma_3$
$(\alpha_2, \alpha_5), (\alpha_3, \alpha_1), (\alpha_4, \alpha_6)$	$\gamma_1 + \gamma_2 + \gamma_3$

Favorable Contact Model: Statistical Mechanics



System Assumptions:

- regions are labeled as α_i with $i = 1, \ldots, 2N$
- only pairwise contacts between regions are possible
- links between regions are infinitely extendible
- microstates of system are given by collections of pairwise contacts (e.g., for $N = 3$, a microstate is $\{(\alpha_1, \alpha_3), (\alpha_2, \alpha_5), (\alpha_4, \alpha_6)\}$)
- there is a single zero-energy (i.e., native) collection of pairwise contacts given by—for $2N$ regions—$\{(\alpha_1, \alpha_2), (\alpha_3, \alpha_4), \ldots, (\alpha_{2N-1}, \alpha_{2N})\}$.



Energy Definition:

- The microstate $\{(\alpha_1, \alpha_2), (\alpha_3, \alpha_4), \ldots, (\alpha_{2N-1}, \alpha_{2N})\}$ has zero energy. For an arbitrary microstate, there is an energy cost of γ_i if the contact pair i is not an element of this set of pairs.

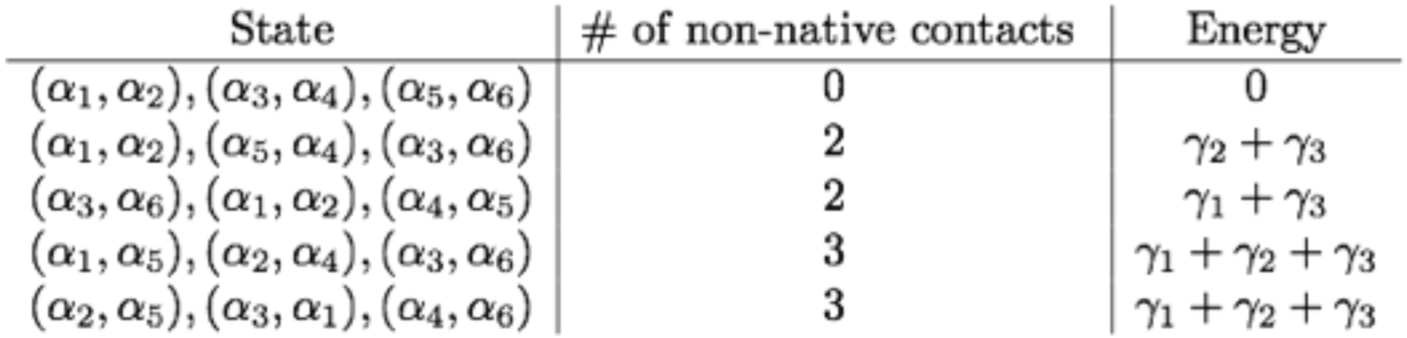
$$\mathcal{H}_N(\{(\theta_1^{(i)}, \theta_2^{(i)})\}) = \sum_{i=1}^{N} \gamma_i I(\theta_1^{(i)}, \theta_2^{(i)})$$

where

$$I(\theta_1^{(i)}, \theta_2^{(i)}) = \begin{cases} 0 & \text{if } (\theta_1^{(i)}, \theta_2^{(i)}) \in \{(\alpha_1, \alpha_2), \ldots, (\alpha_{2N-1}, \alpha_{2N})\} \\ 1 & \text{otherwise.} \end{cases}$$

Example: 6 components

$\left\langle I(\theta_1^{(i)}, \theta_2^{(i)}) \right\rangle$: average number of non-native contacts



State	# of non-native contacts	Energy
$(\alpha_1, \alpha_2), (\alpha_3, \alpha_4), (\alpha_5, \alpha_6)$	0	0
$(\alpha_1, \alpha_2), (\alpha_5, \alpha_4), (\alpha_3, \alpha_6)$	2	$\gamma_2 + \gamma_3$
$(\alpha_3, \alpha_6), (\alpha_1, \alpha_2), (\alpha_4, \alpha_5)$	2	$\gamma_1 + \gamma_3$
$(\alpha_1, \alpha_5), (\alpha_2, \alpha_4), (\alpha_3, \alpha_6)$	3	$\gamma_1 + \gamma_2 + \gamma_3$
$(\alpha_2, \alpha_5), (\alpha_3, \alpha_1), (\alpha_4, \alpha_6)$	3	$\gamma_1 + \gamma_2 + \gamma_3$

What are the equilibrium statistical mechanical properties of this system at an arbitrary *T*?

Favorable Contact Model: Statistical Mechanics

Energy Definition:

– The microstate $\{(\alpha_1, \alpha_2), (\alpha_3, \alpha_4), \ldots, (\alpha_{2N-1}, \alpha_{2N})\}$ has zero energy. For an arbitrary microstate, there is an energy cost of γ_i if the contact pair i is not an element of this set of pairs.

$$\mathcal{H}_N(\{(\theta_1^{(i)}, \theta_2^{(i)})\}) = \sum_{i=1}^{N} \gamma_i I(\theta_1^{(i)}, \theta_2^{(i)})$$

where

$$I(\theta_1^{(i)}, \theta_2^{(i)}) = \begin{cases} 0 & \text{if } (\theta_1^{(i)}, \theta_2^{(i)}) \in \{(\alpha_1, \alpha_2), \ldots, (\alpha_{2N-1}, \alpha_{2N})\} \\ 1 & \text{otherwise.} \end{cases}$$

What are the equilibrium statistical mechanical properties of this system at an arbitrary *T*?

↓

What is the partition function for this system?

$$Z_N(\{\beta\gamma_i\}) = \sum_{\{(\theta_1^{(i)}, \theta_2^{(i)})\}} \exp\left(-\beta \sum_{j=1}^{N} \gamma_j I(\theta_1^{(j)}, \theta_2^{(j)})\right)$$

(Some mathematical work, i.e., dancing partners)

$$\boxed{Z_N(\{\beta\gamma_i\}) = \frac{1}{\sqrt{\pi}} \int_0^\infty dt\, t^{-1/2} e^{-t} \prod_{k=1}^{N} \left(1 + (2t-1)e^{-\beta\gamma_k}\right)}$$

An unphysical but analytically soluble folding model

Allows us to compute:

— two-region and four-region correlations

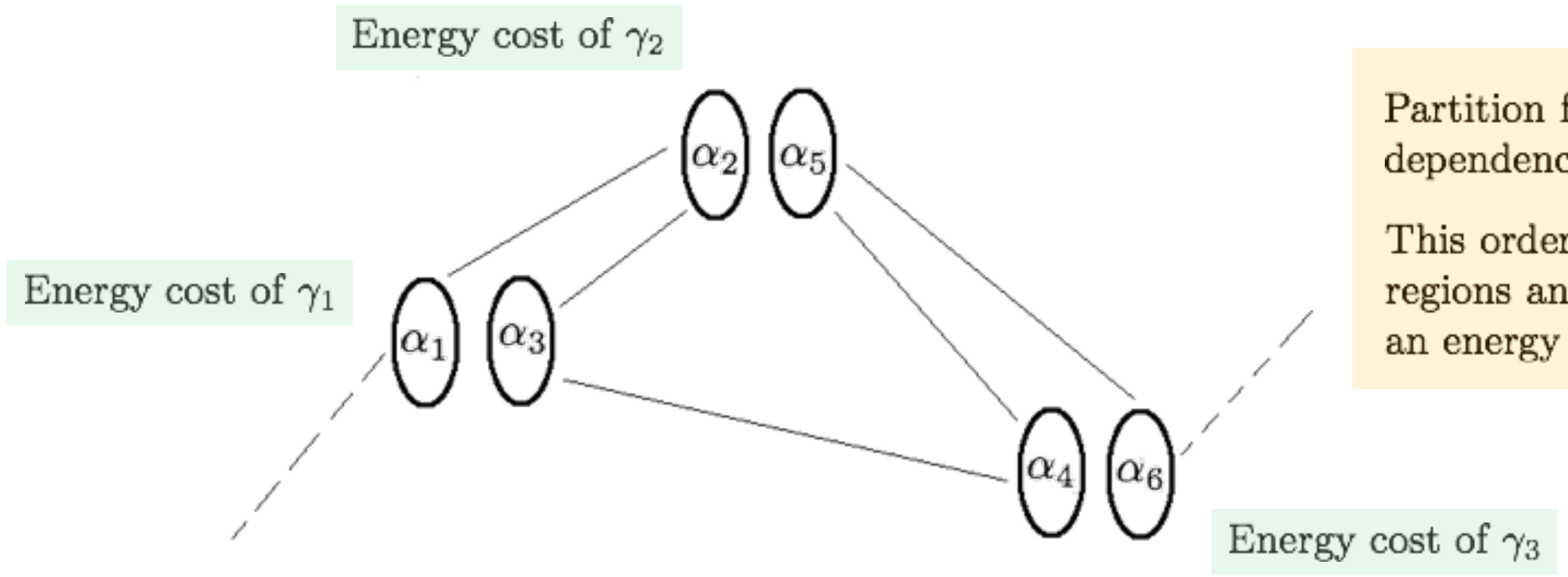
— free energy as a function of deviation from native state

— native-to-unfolded transition temperatures

* Very similar in form to the result from the statistical mechanics of the symmetric group: $Z_N(\{\beta\lambda_i\}) = \int_0^\infty ds\, e^{-s} \prod_{\ell=1}^{N} \left[1 + (s-1)e^{-\beta\lambda_\ell}\right]$

FCM: Unphysical Assumption

Favorable Contact Partition Function: $Z_N(\{\beta\gamma_i\}) = \frac{1}{\sqrt{\pi}} \int_0^\infty dt\, t^{-1/2} e^{-t} \prod_{k=1}^{N} \left(1 + (2t-1)e^{-\beta\gamma_k}\right)$



Energy cost of γ_2

Energy cost of γ_1

Energy cost of γ_3

Partition function has an explicit dependence on γ_k

This orders the contacts between regions and associates this order with an energy cost.

...but the energy cost should depend on **the properties** of the adjacent regions and *not* on the arbitrary ordering of the contacts.

For solubility, our model does not consider the **specific properties** of interactions between regions…

…but we can consider **distributions** of these interaction properties.

→ Each γ_k is drawn from distribution $\rho_0(\gamma)$

FCM: Parameterizing Ignorance

Favorable Contact Partition Function:

$$Z_N(\{\beta\gamma_i\}) = \frac{1}{\sqrt{\pi}} \int_0^\infty dt\, t^{-1/2} e^{-t} \prod_{k=1}^{N} \left(1 + (2t-1)e^{-\beta\gamma_k}\right)$$

Instead of giving each contact k an energy cost of γ_k

we say the energy cost for contact k is drawn from a distribution $\rho_0(\gamma)$

Example: Gaussian Distribution of Energy Costs

Each γ_k is drawn from a normal distribution with mean γ_0 and variance σ_γ^2.

$$\gamma_k \sim \mathcal{N}(\gamma_0, \sigma_\gamma^2)$$

Mean energy cost; proxy for the **stability** of the native configuration

If γ_0 ↑, then stability increases
If γ_0 ↓, then stability decreases

Variance of energy cost; proxy for the **heterogeneity** of the chain of regions

If σ_γ ↑, then chain is more heterogeneous
If σ_γ ↓, then chain is more homogeneous

Question
How do we determine the **average properties** of a system defined by such a distribution of interaction energies?

→

Answer
Compute the quenched free energy!

Generalized Favorable Contact Model

Favorable Contact Free Energy (with energy cost distributions) $\longrightarrow$ **Quenched Free Energy**

$$\langle \ln Z_N(\{\beta\gamma_i\})\rangle = \int_{-\infty}^{\infty} \prod_{j=1}^{N} d\gamma_j\, \rho_0(\gamma_j)\, \ln \frac{1}{\sqrt{\pi}} \int_0^{\infty} dt\, t^{-1/2} e^{-t} \prod_{k=1}^{N} \left(1+(2t-1)e^{-\beta\gamma_k}\right)$$

"Generalized Favorable Contact Model"

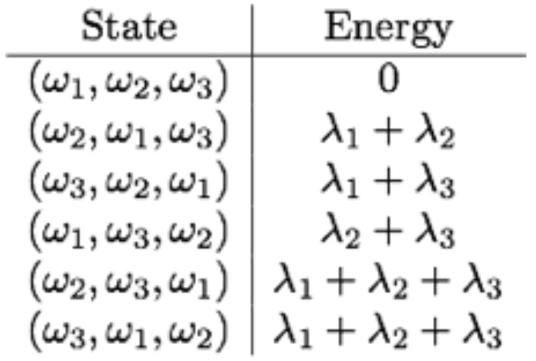
OK. It's a complicated mathematical expression. So what?

This favorable contact free energy is similar in form to that for the **permutation glass**

$$-\beta F_{\text{perm. glass}} = \int_{-\infty}^{\infty} \prod_{j=1}^{N} d\lambda_j\, \rho_0(\lambda_j)\, \ln \int_0^{\infty} ds\, e^{-s} \prod_{k=1}^{N} \left(1+(s-1)e^{-\beta\lambda_k}\right)$$

Permutation Glass (definition)
physical system where the state space consists of permutations of a list and the Hamiltonian depends on random parameters. (similar to spin glass)

e.g.



State	Energy
$(\omega_1, \omega_2, \omega_3)$	0
$(\omega_2, \omega_1, \omega_3)$	$\lambda_1 + \lambda_2$
$(\omega_3, \omega_2, \omega_1)$	$\lambda_1 + \lambda_3$
$(\omega_1, \omega_3, \omega_2)$	$\lambda_2 + \lambda_3$
$(\omega_2, \omega_3, \omega_1)$	$\lambda_1 + \lambda_2 + \lambda_3$
$(\omega_3, \omega_1, \omega_2)$	$\lambda_1 + \lambda_2 + \lambda_3$

where each λ_i is drawn from a probability distribution

Physical Results of Permutation Glass

For a chain length *N*, a non-degenerate free energy minimum exists only if

$$\frac{\langle \lambda \rangle}{\sigma_\lambda} \gtrsim \sqrt{2 \ln N}$$

mean/variance of energy cost for deviating from non-degenerate microstate

Can we establish similar constraints for the generalized favorable contact model?

Generalized FCM: Questions

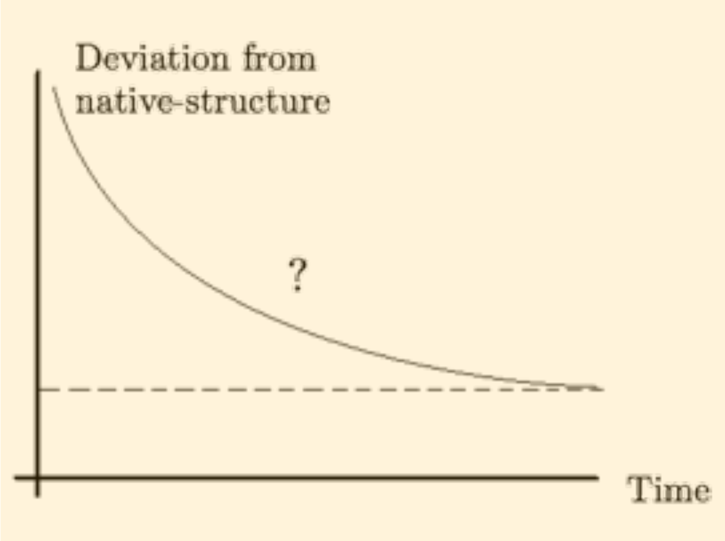
Questions to ask about General Favorable Contact Model

- **Existence of Native State:** How do the stability and heterogeneity properties of this abstracted polypeptide affect the existence of a native state?

- **Relaxation (“Folding”) Time:** How much time does it take the system to reach thermal equilibrium? Such a time can be taken to be how long it takes the abstracted polypeptide to “fold.”

- **Non-native contacts:** How does the average number of non-native contacts vary with temperature and parameters?

- **Regional Contact Correlations:** What is the four-region correlation? (i.e., the correlation between two different two-region contact regions)

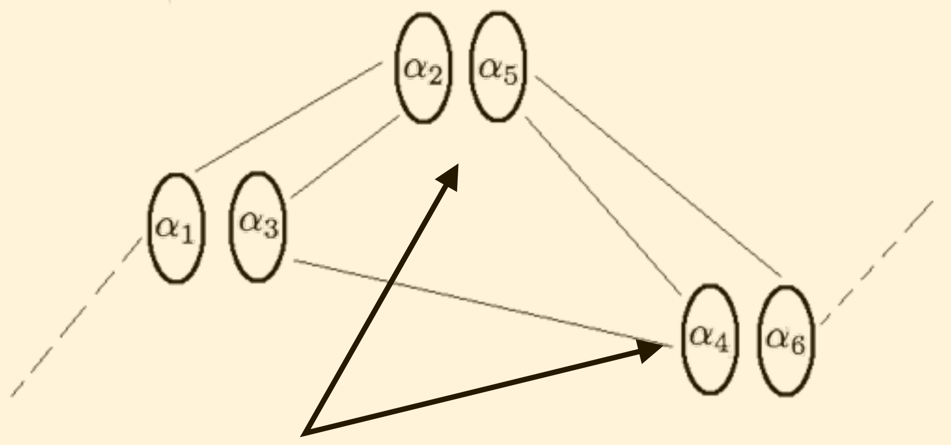


$\left\langle I(\theta_1^{(i)}, \theta_2^{(i)}) \right\rangle$: average number of non-native contacts

$\langle\gamma\rangle$: proxy for stability of native state

σ_γ : proxy for heterogeneity of chain

$\Longrightarrow$ How does this affect the existence of native state?



correlation between two “two-region-contacts”?

Generalized FCM: Unphysical Assumptions

Unphysical assumptions of model

- **Even number of regions:** Assumed the chain consisted of exactly an even number of regions.

- **No specific interaction matrix:** In lieu of an interaction matrix, which could account for how specific regions interact with each other, we assumed the interaction energies could be modeled by quenched disorder.

- **Infinitely extendable chain:** To allow for all possible interacting combinations, we allowed our abstracted polypeptide to be infinite extendable.

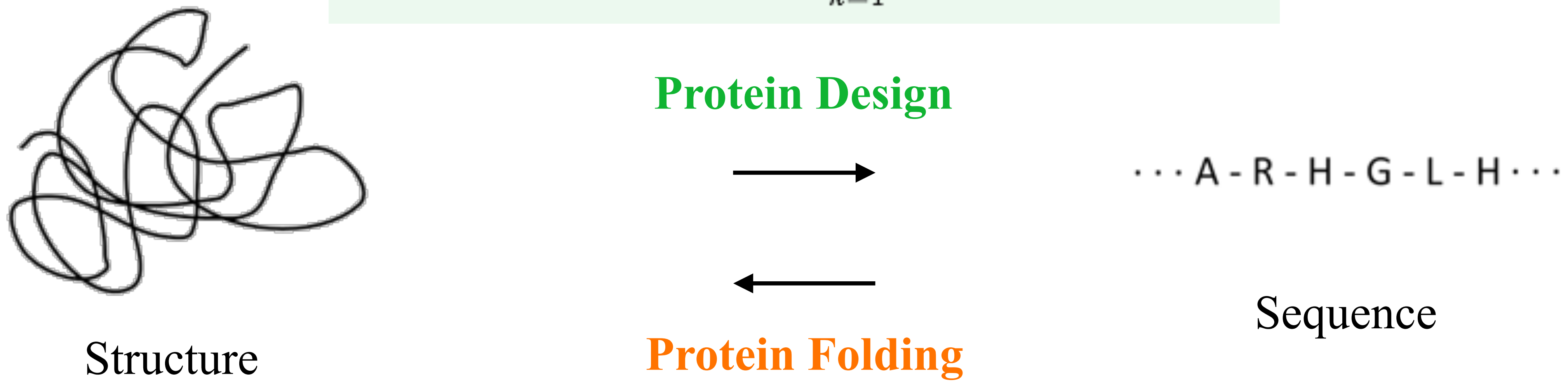
- **Only includes interacting pairs:** Although two regions of a polypeptide chain can sometimes define interaction, they do not exclusively do so.

Connecting Protein Folding and Protein Design

(Abstracted) Design and Folding Problems

Study a model of possible sequence orders
for a given presumed structure

$$Z_N^{\text{perm.}}(\{\beta\lambda_i\}) = \int_0^\infty ds\, e^{-s} \prod_{k=1}^{N} \left(1 + (s-1)e^{-\beta\lambda_k}\right)$$



Protein Design

$$\longrightarrow$$

$$\longleftarrow$$

Protein Folding

$\cdots$A-R-H-G-L-H$\cdots$

Structure

Sequence

Study a model of possible pairwise contacts
for a given sequence of contact regions

$$Z_N^{\text{FCM}}(\{\beta\gamma_i\}) = \frac{1}{\sqrt{\pi}} \int_0^\infty dt\, t^{-1/2} e^{-t} \prod_{k=1}^{N} \left(1 + (2t-1)e^{-\beta\gamma_k}\right)$$

END