Sampling rare events: folding pathways

Alberto Perez

Mar 10th 2022 — OpenEye CUP XXI
Everything that living things can do can be understood in terms of the jiggling and wiggling of atoms

- Richard Feynman
We can understand MD as jumps between states at certain timescales

Sometimes we are interested in the native state but we don’t even know what it looks like.

Modeled system

How do we know we have found the native state?
How long do we need to sample?
Statistical significance?
What is the physical process?

Unknown Native State
Our group works in understanding molecular interactions

Protein Folding/Structure determination

Peptide-protein binding

DNA-protein binding

Peptide self-assembly
Conventional MD is inefficient at exploring the energy landscape.
We leverage biophysical data that has been insufficient for structural determination.
We incorporate data into simulations through Bayesian inference.

\[
p(x|D) = \frac{p(D|x)p(x)}{p(D)} \sim p(D|x)p(x)
\]

\[
p(x) \sim \exp[-\beta E_{\text{force}}(x)]
\]
We use generalized ensemble methods for enhanced sampling

- High Temperature / Weak Restraints
- Low Temperature / Strong Restraints
Phase transitions lead to local exchanges and limit sampling efficiency.
What types of data?

• Sparsely labeled NMR

• Cryo-EM (CryoFold)

• Φ-value analysis

• Chemical Shift Perturbation NMR

• Paramagnetic relaxation enhancement (PRE)

• General knowledge

• …
What types of data?

- Sparsely labeled NMR
- Cryo-EM (CryoFold)
- Φ-value analysis
- Chemical Shift Perturbation NMR
- Paramagnetic relaxation enhancement (PRE)
- General knowledge
- ...
Protein Folding problem

• (1) What structure encoded by a sequence
• (2) How do proteins fold that fast (pathways)
• (3) Can we design new proteins
Blind competition events are a great way to validate methodologies

Protein G and L are two proteins with same topology and different folding pathways

G: MTYKLILNGKTLKGETTTEAVDAATAEKVFQKQYANDNGVDGEWTYDDATKTFTVTE

L: MEEVTIKANLIFANGSTQTAEFKGFTEKATSEAYAYADTLKKDNGEWTVDVADKGYTLNIKFAG

G: ms folder

L: s folder
Protein G and L are two proteins with same topology and different folding pathways

G
MTYKLILNGKTLKGETTTEAEDAAATAAEKVFKQYANDNGVGDGEWTYDDATKTFVTE
G
DTYKL[VIVLNGTTFTTY]TTEAVDAATAEKVFKQYANDNGVGDGEWTYADATKTFVTE

L
MEEVTIKANLIIFANGSTQTAETFKGTKEKATSEAYAYADTLKDNGEWTVDVADKGYTLENIKFAG
L
EEVTIKANLIIFANGSTQTAETFKGTKEKALSEVLAYADTLKDNGEWTIDKRVTNGVILLIKFAG

Gmut: µs folder
G: ms folder
Lmut: ms-s folder
L: s folder
Φ-value analysis informs us of residues likely to be making native interactions in the transition state ensemble.

\[
\phi = \frac{\left( \Delta G_W^{TS \rightarrow D} - \Delta G_M^{TS \rightarrow D} \right)}{\left( \Delta G_W^{N \rightarrow D} - \Delta G_M^{N \rightarrow D} \right)} = \frac{\Delta \Delta G^{TS \rightarrow D}}{\Delta \Delta G^{N \rightarrow D}}
\]
Φ-value analysis informs us of residues likely to be making native interactions in the transition state

\[
\phi = \frac{\left( \Delta G^{TS\rightarrow D}_W - \Delta G^{TS\rightarrow D}_M \right)}{\left( \Delta G^{N\rightarrow D}_W - \Delta G^{N\rightarrow D}_M \right)} = \frac{\Delta \Delta G^{TS\rightarrow D}}{\Delta \Delta G^{N\rightarrow D}}
\]

Mutant residue ordered in TS

Mutant residue disordered in TS
What’s the minimal amount of information that will focus sampling and identify folding pathways?
Protein G folds through the second hairpin, through an intermediate.
Protein L and its mutant fold through the first hairpin, with no intermediates
MELD captures TSE and intermediates. Is this systematic?
For protein G and $G_{\text{mut}}$ we have few folding events.
Markov State Models
Using Adaptive sampling capture the folding kinetics of Protein G mutant
Peptide Binding

Chemical Shift Perturbation Data
Intrinsically disordered peptides fold upon binding

SRLTWVQRSQNPLKIRLTLTREP
PEIKLKTIKTIQNGRELFESSLCGLNLEVQASE
NLQSSIVKKPLPLTQPG
KWTLERLKRKYRN
KVAFLKILGGF
RSVVKIKLGRK

ET domain of BRD3

ET is an interaction hub involved in gene regulation and virus entry.

---

Peptide Binding

Chemical Shift Perturbation Data
Intrinsically disordered peptides fold upon binding

SRLTWVQRSTQNLKIRLRTAE
PEIKKT KKTIQNGRELFFESSLCGLLNEVQASE
NLQSSIVFKKPLPLTQPG
KWTLERLKRKYRN
KVAPLKLGGF
RSVKVKIKLGRK

ET domain of BRD3
ET is an interaction hub involved in gene regulation and virus entry.
Peptides bind ET with a wide range of binding affinities

Kd ~ 10 nM → ~2-3 months
Kd ~ 10 µM → ~2-3 years

NMR
NOESY (protein-peptide)
Chemical Shift Perturbation (CSP) (Protein)
Chemical shift perturbation provides indirect data about where the peptide might bind

Free ET
Bound ET

Allosteric changes?
Direct contacts?
Our ensembles sample multiple bound/misbound states and identify the native bound structure.
Competitive binding simulations help us determine binding affinity.

\[ \Delta G_{BA} = -k_b T \ln \left( \frac{p_B}{p_A} \right) \]
Computed binding free energies are in agreement with ITC data

$\Delta \Delta G_{\text{bind}} \sim 2.4 \text{ kcal/mol.}$

$\Delta \Delta G_{\text{bind}} \sim 2.6 \text{ kcal/mol.}$

Predicted

Experiment

We used the lessons learnt from folding routes to design a novel peptide inhibitor.
Structural Biology needs physical modeling

- Learn about why/how
- Lead to design principles (smooth landscapes)
- Transferable to new materials
- Simpler and slower folding pathways for protein L over G
- Bridges experiments and atomistic structures