Protein Folding - seeing is deceiving

Bioinformatics Institute Agency for Science Technology and Research Singapore 1 October 2021

> George Rose Johns Hopkins University grose@jhu.edu

> > 1

Protein Folding - seeing is deceiving Covido ergo Zoom

- with apologies to Descartes

Protein Folding – seeing is deceiving Recent paper Protein Science (2021) 30: 1606–1616 and further thoughts from

Biochemistry, "Protein folding from a physical-chemical perspective: entropy as organizer", invited Perspective, in preparation.

Protein Folding Problem

<u>Ribonuclease A</u>

LYS GLU THR ALA ALA ALA LYS PHE GLU ARG GLN HIS MET ASP SER SER THR SER ALA ALA SER SER ASN TYR CYS ASN GLN MET LYS SER ARG ASN LEU THR LYS ASP ARG CYS PRO VAL ASN THR PHE VAL HIS SER LEU ALA ASP VAL GLN ALA VAL CYS GLN LYS ASN VAL ALA CYS LYS ASN GLY GLN THR CYS TYR GLN SER TYR SER THR MET SER ILE THR ASP CYS ARG GLU THR GLY LYS TYR PRO ASN CYS ALA TYR LYS GLN ALA ASN LYS HIS ILE ILE VAL GLU GLY ASN PRO TYR VAL PRO VAL HIS ASP ALA SER VAL



Definition of the problem for this talk: <u>Predict conformation from sequence</u>.

Protein Folding Problem

The stunning success of deep-learning artificial intelligence (AI) approaches has transformed the field.

Jumper, J., Evans, R., Pritzel, A., et al. (2021) Highly accurate protein structure prediction with AlphaFold, Nature 596, 583–589.

Tunyasuvunakool, K., Adler, J., Wu, Z., et al. (2021) Highly accurate protein structure prediction for the human proteome, Nature 596, 590–596.

Baek, M., DiMaio, F., Anishchenko, I., et al. (2021) Accurate prediction of protein structures and interactions using a three-<u>track neural</u> network, Science 373, 871–876.

Progress in Science

observation \rightarrow pattern recognition \rightarrow theory/models

Where are we in this progression? observation: 50th year of the protein data bank (PDB) pattern recognition: deep learning AI

theory/models: ??

Protein Folding Problem

observation \rightarrow pattern recognition \rightarrow theory/models

In imperfect analogy, protein structure prediction using AI is akin to Mendeleev's compilation of the periodic table of the elements prior to its eventual derivation from quantum mechanics (e.g., molecular orbital theory).

Outline of this seminar

The current paradigm Excluding interactions Entropy as organizer

Framing the problem

Friedrich Nietzsche: what we see depends on the perspective from which we look.

Tony Schwartz (my dentist): if you're not looking for it, you don't see it.

In both politics and science, what we see depends on the perspective from which we look.

The current paradigm



Intuitively, what determines this conformation?

The current paradigm



Conditioning expectations about protein folding: many favorable interactions

The Anfinsen Hypothesis

What am I thinking?*



All backbones are the same. Side chains discriminate. The most energetically favorable constellation of interactions between and among the side chains corresponds to the native conformation.

$U(nfolded) \rightleftharpoons N(ative)$

*But, of course it doesn't matter: this is thermodynamics.

The current paradigm



The Anfinsen Hypothesis $U \rightleftharpoons N$

Native state = minimum free-energy conformer All backbones are the same - side chains discriminate

Haber & Anfinsen (1961) J Biol Chem 236:422-424

The current paradiam (sort of) $U \rightleftharpoons N$

featureless random coil

Native state = minimum free-energy conformer
Unfolded state is a featureless landscape
All backbones are the same – side chains discriminate
Organizing interactions are visible in N

Current Paradigm - what the mechanism?

The "what you see is what you get" view.



Force field minimization Protein = $\frac{A}{r^{12}} - \frac{B}{r^6} - \frac{\sum q_i q_j}{\epsilon r_{ij}}$ - H-bonds-torsions-dipoles ...

Also: knowledge-based potentials, contact energies, Go models, lattice models ...

All are attractive (i.e. stabilizing) interactions

Current Paradigm - what's the mechanism?

The "what you see is what you get" view.

Force field minimization Protein = $\frac{A}{r^{12}} - \frac{B}{r^6} - \frac{\sum q_i q_j}{\epsilon r_{ij}}$ - H-bonds-torsions-dipoles ...

Also: knowledge-based potentials, contact energies, Go models, lattice models ...

but seeing is deceiving

Seeing is deceiving



Seeing is deceiving

Tim Noble and Sue Webster

Now for something completely different



Re-framing the question

Disfavored interactions

The what you see is what you get view, i.e. Organizing interactions are visible in N

Proposing instead that substantial organization results from elimination of unfavorable interactions – excluding interactions.

But first, what's an excluding interaction?

What's an Excluding interaction

Driving forces are attractive interactions. Excluding forces are disfavored interactions. They exclude high-energy interactions, reducing entropy loss on folding.

Knowledge-based potentials, contact energies, Go models, lattice models ... all based on attractive interactions.

Two primary excluding forces: (i) sterics (ii) hydrogen-bond satisfaction

Excluding interactions are not visible in the X-ray structure

Excluding interactions

Condition expectations with a toy model

Exhaustive menu of Ala tetramers with a 5-state model



54 = 625 possible distinct conformers

这好我们的我 整 致 致 致 推 推 推 她 推 她 推 她 推 她 你 你 你 你 你 你 我 我 我 The set set at any shirt of the way set we say the set of the man and the set at at the set 把人 你不 你是 那年 要求 要求 要求 医子 医子 医子 医子 医子 经外 经外 医子 化化 医子 化化 医子 化化 and some there are and the and the set whe will not and the will the son the set are the the the Job hud by Job and Jos TE are by ma phy man and my phy and by by high high his and 「 こう こう ちょ ちょ ちょ ちゃ ちゃ ちゃ かん かん かん かん かん かん かん ちゃ ちゃ ちゃ ちゃ ちゃ ちゃ ちゃ when when when the the the the with when when when when when when the when the when the when the when when when 大子 大兵 大人 如何 如何 如何 如何 如何 如何 不可 医化 如何 医化 如何 医子 化合 化合 化合 化合 化合 人口 人口 the set with the rest which have not show that have the top way the top with here will have the top and the あっち ちょうちょう ちょう ちょうちょう ちょうちょうちょうちょう ちょうちょう with the the set at the set at the set and the set and the set and the set at the the set and man mat the 分子 文章 王林 圣林 林林 林林 林林 齐子 齐山 是子 子子 子子 子子 法律 王林 法律 林子 法律 不幸 大学 不幸 大学 不幸

It the phy the The set at some the set of the se the state of the s that show the part that when the show that the show that the show in any and the the the the the server and the life in the tell the the tell the start we start the start the start the start the tell the tell the when when when the the the the with with with when when the the when the the when the the the the the ないっちゃ ちな なか かか かか かか かき ひろ ひん かん ひん ひん ひん なか たち かん ひん ひん ひん ひん ひん ひん ひん by the by the the the the the the has not the the has not the the the the the the the the with show more than and that and that and that and that and that and the star she that the star and and and and the an the way we want the fet and we want was and me we want and and the want the source and the source and

that share the state with a state of the sta man and and and the state of the the Me was the star of the star of the Me Me Me May May May and the star of th to the source source to the after a first and the source of the source of the source to the source tothe source to the source to when the way and the the the the the when the when the when the the start the the the the the ********************************* A the set of the set o

城	* I Tak		XA	the	127	Hory		Harry	上本					Z,	¥				Att	the hits
	Ft.			th	r.	14									441		Lat		then	met
	savy	two	1 mg	-17					the	my	オント	way	maring	(tant	the	when	rout
			57	Xonth											~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	tot	٤		243	
				Ht.	site.				-45	the	华	-the	-Jerry					**	やちょ	
文文					2mg		mary .							the		my the	a set			set
tur	xwy		2mg	V.	1	t				Thread	2-1		Howest	·			stathe	start for	Harry-	furthe
No.		starts		15	-13	Jt.		-FL	start f	where	MA	ward					where	with	in	ALAX
六年	in	The	谷中	-53		A		M	the	Young		House	JHY J				And.	try	-777	Art
wit .	丸		wit	white	YER			solyth		the		the	,the	with			147	th.	++1	
林	松			Mar	They		my		my						Them	Yt	star a		the	story.
	maria	through		14		-12		-5.5	mut	****	sight	mar and a second	Arr Harr				munt	mut	my	maryor
the start	with	wit	mat	,At	ngh.	,FX		,IL	my	with 1	444	stight	why	•			the	the the	and the host	software
				414						the		John -	,Fx			林东	L 473			
野	ない			net?	The second	447			my.					int the	the th	tor	y stra	1th	the	the state
thread a		typest		****		14			my t	441	442	ship h	here				story a	type	sough	2 Aug
		myt				#th		#th	the	AM	attat	Are	Jones				where	the	. foto for	whether

The organizing power of excluding interactions

Conformers with steric clashes or backbone polar groups that lack hydrogen bond partners make a negligible contribution to the overall thermodynamic population and are not visible in the native structure.

Excluding interactions are not visible in the X-ray structure

Not captured in knowledge-based potentials, contact energies, Go models, lattice models

What's an Excluding interaction

Two primary excluding forces: (i) sterics (ii) hydrogen-bond satisfaction

An example of each type of excluding force

Excluding interactions – sterics

The Flory isolated-pair hypothesis: The simplifying assumption that each φ, ψ pair is sterically independent of all but its adjacent chain neighbors.

Excluding interactions – sterics

An α-helix cannot be followed by a contiguous β-strand

Systematic local steric restrictions extend beyond adjacent chain neighbors.



Pappu, Srinivasan & GDR(2000) PNAS 97:12565–12570. Fitzkee & GDR (2004) Protein Science 13: 633–639. Fitzkee & GDR (2005) "J. Mol. Biol. 353: 873–887.

Excluding interactions – H-bond satisfaction

Protein Folding: current paradigm $U \rightleftharpoons N$

All backbones are the same – side chains discriminate

The most energetically favorable constellation of interactions between and among the side chains corresponds to the native conformation.

Force field minimization

Protein = $\frac{A}{r^{12}} - \frac{B}{r^6} - \frac{\sum q_i q_j}{\epsilon r_{ij}} - H$ -bonds-torsions-dipoles ...

All backbones are the same — side chains discriminate —

When a protein folds, many backbone polar groups are removed from solvent access. These groups must find intra-molecular hydrogen-bond partners. Why?

GDR et al, (2006) "A backbone-based theory of protein folding." Proc Nat. Acad. Sci. 103: 16623–16633.

It's the 21st century and we're still unsure of how much a hydrogen bond is worth. But we do have a good idea that the energetic cost of a completely unsatisfied H-bond = ~+5 kcal/mol.

$$\frac{P_u}{P_u} = e^{\frac{-\Delta E_{hb}}{RT}} = 0.02\%$$

 P_u - probability of an unsatisfied hydrogen bond ΔE_{hb} - energy of a hydrogen bond (~ -5 kcal/mol) R - gas constant T - temperature

Fleming & GDR (2005) "Do all backbone polar groups in proteins form hydrogen bonds?" Protein Sci 14:1911–1917.
Panasik, Fleming & GDR (2005) "Hydrogen-bonded turns in proteins: The case for a recount" Protein Science 14: 2910–2914

Backbone-dominated model of folding A backbone hydrogen bond may add little to the stability of the native state, but a completely unsatisfied backbone hydrogen bond would be dramatically destabilizing (+5 kcal/mol), shifting the U≓N folding equilibrium far to the left,

rivaling the entire $\Delta G_{conformation}$ for a typical protein \approx [-5, -15] kcal/mol.

Fleming & GDR (2005) "Do all backbone polar groups in proteins form hydrogen bonds?" Protein Sci 14:1911–1917.
Panasik, Fleming & GDR (2005) "Hydrogen-bonded turns in proteins: The case for a recount" Protein Science 14: 2910–2914

Backbone-dominated model of folding $\Delta G_{\text{conformation}}$ for a protein \approx [-5, -15] kcal/mol. If a backbone polar group is satisfied by water when unfolded but left unsatisfied when folded, the $U \rightleftharpoons N$ would be shifted far to the left. There are only two extensible hydrogen-bond-satisfying conformers: α -helix and β -strands. Necessarily, all proteins are built on scaffolds of these two hydrogenbonded elements.

GDR et al, (2006) "A backbone-based theory of protein folding." Proc Nat. Acad. Sci. 103: 16623–16633.

How many distinct scaffolds are possible?

Using lysozyme (129 residues) as a template, a typical domain might have ~ 10 elements of α -helix and/or β -sheet = 2¹⁰ possibilities X complexity from interconnecting loops.

Interconnecting loops are typically short and therefore constraining.



... Only a few thousand scaffolds are possible

All backbones are the same — side chains discriminate

For a protein domain (e.g. lysozyme or ribonuclease), only a few thousand backbone scaffolds are possible (not some incomprehensibly large number).

 Chothia (1992) Proteins. One thousand families for the molecular biologist, Nature 357, 543–544

• Przytycka, Aurora & GDR (1999). "A protein taxonomy based on secondary structure." <u>Nat Struct Biol</u> 6(7): 672–682.

Of thermodynamic necessity, proteins are built on scaffolds of α -helix and/or β -sheet, and only a few thousand backbone scaffolds are possible.

All backbones are the same — side chains discriminate

For a protein domain (e.g. lysozyme or ribonuclease), only a few thousand backbone scaffolds are possible (not some incomprehensibly large number).

side chains discriminate among these alternatives

Side chains discriminate among these alternatives

All backbones are the same — side chains discriminate

But even here, steric excluding interactions impose substantial restrictions.

A Ramachandran-type map for side chains Side chain conformational bias local organization – sterics only Blocked mono-peptides: CH3-CO-AA-NH2





GDR (2019) Ramachandran maps for side chains in globular proteins. Proteins 87:357–364.

A Ramachandran-type map for side chains Side chain conformational bias local organization – sterics only Blocked mono-peptides: CH3-CO-AA-NH2

For each clash-free backbone conformation, generate the full range of side chains conformations: $\chi 1 = [-180, 180]$, $\chi 2 = [-180, 180]$; exclude those with steric clashes. For each allowed side chain conformation, increment the backbone population count.

GDR (2019) Ramachandran maps for side chains in globular proteins. Proteins 87:357–364.

Excluding interactions – sterics







<u>A long-standing question</u>: Why don't individual molecules get stuck in meta-stable traps en route from U to N?

Answer: of thermodynamic necessity, proteins are built on scaffolds of α -helix and/or β -sheet, and only a few thousand backbone scaffolds are possible.

Entropy is the primary organizer in protein folding

Of thermodynamic necessity, globular protein are built on hydrogen bond-satisfied scaffolds of α-helices and/or β-strands. Scaffold folding is highly cooperative, not residue by residue. If not, dangling unsatisfied backbone polar groups would shift the U≓N equilibrium far to the left.

Entropy is the primary organizer in protein folding

Of thermodynamic necessity, globular protein are built on hydrogen bond-satisfied scaffolds of α-helices and/or β-strands. Scaffold folding is highly cooperative, not residue by residue. If not, dangling unsatisfied backbone polar groups wo uld shift the U≕N equilibrium far to the left.

Under folding conditions, the selection of scaffold segments is limited to the two possible alternatives – α -helix or β -strand – implying a substantial degree of prior organization. If so, scaffold assembly of these preorganized components comes at a dramatically reduced cost in conformational entropy.

Entropy is the primary organizer in protein folding

How to count: tiling ϕ, ψ -space into mesostates



Tile space into a 60° x 60° grid of mesostates. 60% of the 14 allowed mesostates is accessible = 8.4 grid equivalents.

If residue by residue ΔS_{folding} = - R ln(8.4/1) cal/mol/degree/residue ΔG_{folding} = 1.28 kcal/mol/residue at 27°C If segment by segment ΔS_{segment} = - R ln(2/1) cal/mol/degree/segment ΔG_{segment} = .42 kcal/mol/segment at 27°C

13 kcal/mol for a 10-residue segment if residue by residue 0.42 kcal/mol for a 10-residue segment if segment by segment Hydrogen-bonding as a thermodynamic pivot
<u>The big 3</u>:
Conformational entropy always favors U
The hydrophobic effect always favors N
Hydrogen-bonding favors U under unfolding conditions but favors N under folding conditions.

The thermodynamic-pivot hypothesis hinges on whether water is a poor solvent for the protein backbone.

In conclusion

Conformers with unsatisfied backbone hydrogen bonds will be culled, thereby rarefying the folding population and reducing the entropy cost of folding.

The ideas proposed here suggest the existence of a quintessential simplicity that underlies the apparent complexity of protein folding.

Thank you!