

Tertiary alphabet for the observable protein structural universe

Presented By: Eugene Choi

Mackenzie, C. O., Zhou, J., & Grigoryan, G. (2016). Tertiary alphabet for the observable protein structural universe. *Proceedings of the National Academy of Sciences*

Outline

- **Background**
- **Methods**
- **Results**
- **Results Analysis/Implications**
- **Discussion**

Background: Protein Structures

Primary:

- Sequence of amino acids

Secondary:

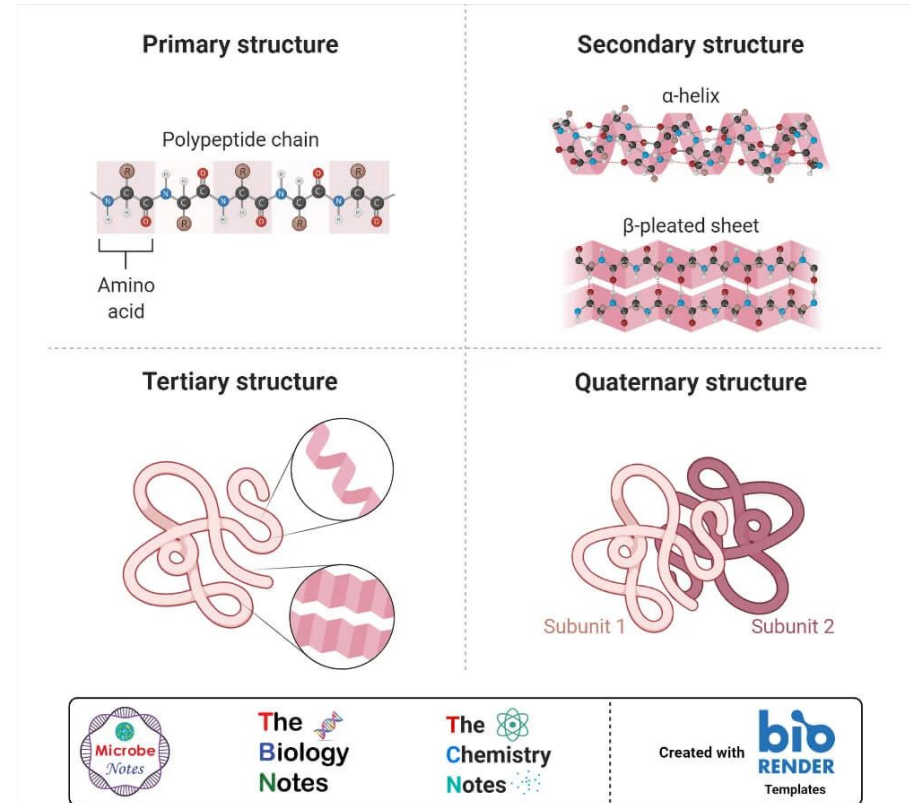
- Localized conformation of the chain

Tertiary:

- Overall 3D structure resulting from folding and covalent cross-linking of a protein

Quaternary:

- Association of several protein chains or subunits into a closely packed arrangement



Background: Vocab

Degeneracy

- Different amino acid sequences can fold into the same or similar 3D structures

Motif

- Recurring pattern of secondary or tertiary structure that is found in multiple proteins

Designability

- The varying degrees of feasibility in engineering a given protein structure using naturally occurring amino acids
- Higher designability -> easier to engineer -> more frequently occurring
- Lower designability -> harder to engineer -> less frequently occurring

Background: Task/Motivation

Question:

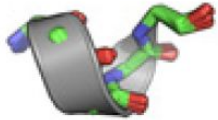
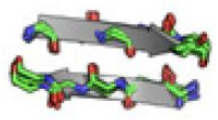
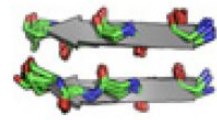

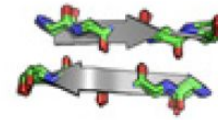
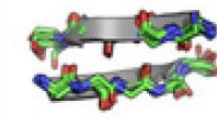
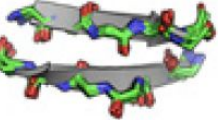

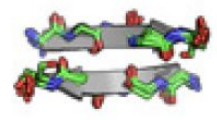
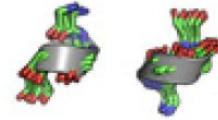

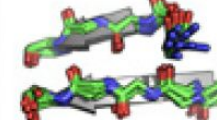

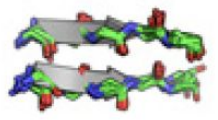


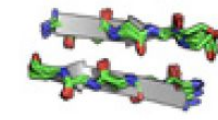
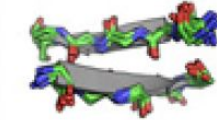
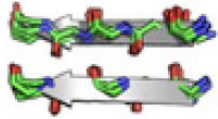
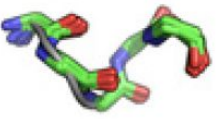

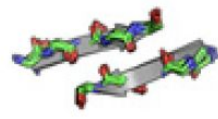

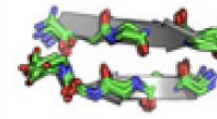
- Can the universe of allowed local 3D structural environment of proteins be modeled by standard reusable building blocks (TERMS)?

Motivation:

- Provides insight between sequence to structure
- “... there is another fundamental reason to expect degeneracy - namely the differential designability of protein structure”

TERMs

A

 <p>1; 686,323 (13%/13%)</p>	 <p>2; 68,544 (6%/6%)</p>	 <p>3; 31,472 (3%/3%)</p>	 <p>4; 51,233 (3%/1%)</p>	 <p>5; 70,713 (5%/1%)</p>	 <p>6; 27,692 (4%/1%)</p>
 <p>7; 7,474 (1%/0.7%)</p>	 <p>8; 19,434 (2%/0.6%)</p>	 <p>9; 26,462 (3%/0.5%)</p>	 <p>10; 13,009 (1%/0.4%)</p>	 <p>11; 384,921 (12%/0.4%)</p>	 <p>12; 4,726 (0.9%/0.4%)</p>
 <p>13; 6,586 (0.5%/0.4%)</p>	 <p>14; 23,295 (3%/0.3%)</p>	 <p>15; 16,468 (2%/0.3%)</p>	 <p>16; 9,440 (1%/0.3%)</p>	 <p>17; 56,575 (4%/0.3%)</p>	 <p>18; 6,429 (1%/0.3%)</p>
 <p>19; 16,981 (2%/0.3%)</p>	 <p>20; 9,030 (0.5%/0.2%)</p>	 <p>21; 23,099 (2%/0.2%)</p>	 <p>22; 38,035 (4%/0.2%)</p>	 <p>23; 20,517 (1%/0.2%)</p>	 <p>24; 18,140 (3%/0.2%)</p>

Methods

Motif Creation

1. **Generate candidate motif for every residue**
 - a. Each candidate motif is initially composed of residue i and $i - 2$ to $i + 2$ residues
2. **For every pair of residue within the amino acid sequence “contact degree” calculated**
 - a. Find all possible rotamers that don't clash with backbone
 - b. Contact degree calculated as weighted fraction of rotamer combinations i and j that have closely approaching non hydrogen atoms (most likely to form contacts)
 - c. Rotamers: describe different ways side chain can be oriented around it's central bond

Motif Creation (continued)

$$\tau(i, j) = \frac{\sum_{a=1}^{20} \sum_{b=1}^{20} \sum_{r_i \in R_i(a)} \sum_{r_j \in R_j(b)} C_{ij}(r_i, r_j) Pr(a) Pr(b) p(r_i) p(r_j)}{\sum_{a=1}^{20} \sum_{b=1}^{20} \sum_{r_i \in R_i(a)} \sum_{r_j \in R_j(b)} Pr(a) Pr(b) p(r_i) p(r_j)}, \quad [2]$$

where $R_i(a)$ is the set of nonclashing rotamers of amino acid a at position i , $C_{ij}(i, j)$ is a logical variable indicating whether rotamers r_i and r_j at positions i and j , respectively, have nonhydrogen atom pairs within 3 Å, $Pr(a)$ is the frequency of amino acid a in the structural database, and $p(r_i)$ is the probability of rotamer r_i from the rotamer library. Contact degree varies from 0 to 1, with higher values corresponding to position pairs more likely to influence each other's amino acids identities. If $\tau(i, j)$ was above 0.05, residues i and j were said to form a PC (55).

Structural search and matching with MASTER

1. Regular RMSD biased towards smaller motifs
2. σ max resolution parameter typically set at 1 Å
3. Rest is normalization component

$$c(t) = \sigma_{max} \sqrt{\left(1 - \frac{2}{N(N-1)} \sum_k \sum_{i=1}^{n_k-1} \sum_{j=i+1}^{n_k} e^{(i-j)/L} \right)}. \quad [1]$$

Minimum Set Cover

1. **NP - complete problem: Finding minimum number of TERMS that covers structural universe**
2. **Greedy Solution:**
 - a. **Choose the TERM element if not chosen that covers most structural units**
 - b. **Iterate until 99% of all elements has been covered**

Discovering TERMS

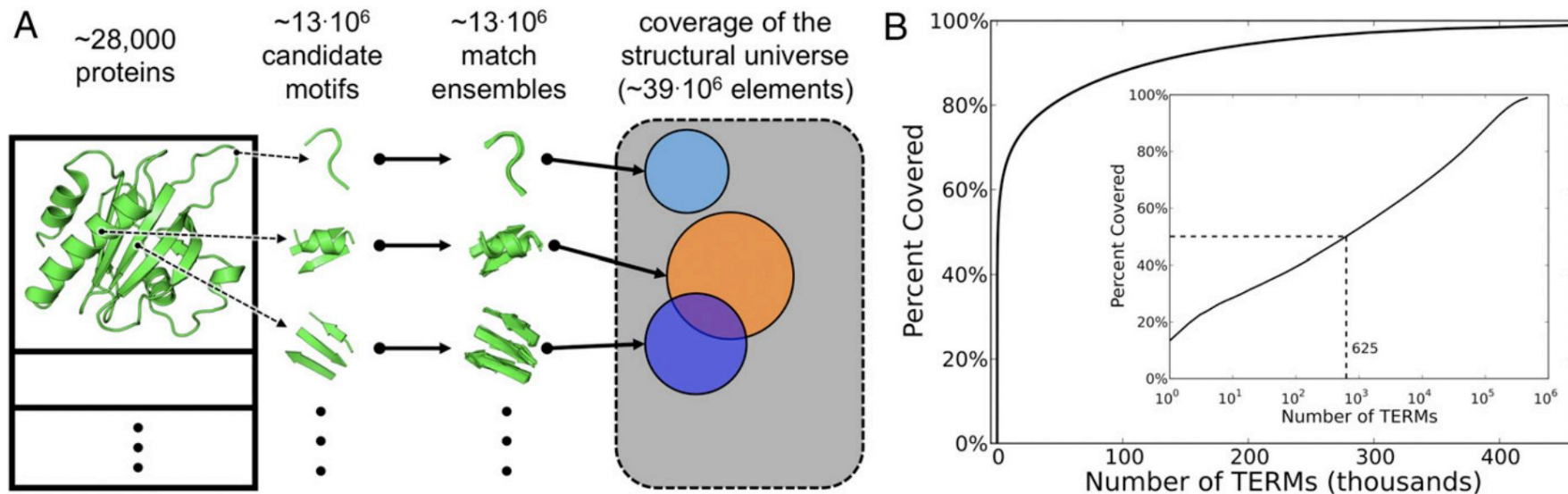


Fig. 1. Discovering TERMS that optimally describe the protein structural universe. **(A)** A candidate motif is defined around each residue in the database, structural matches (from within the database) to each motif are identified using MASTER (58), and these matches are used in defining the coverage of every motif. Next, the set cover problem is solved to find the minimal set of motifs that jointly cover the structural universe. **(B)** Coverage of the universe as a function of the number of TERMS, in the order discovered by the greedy algorithm (inset uses logarithmic scale along the x axis).

Results

A small number of TERMS Describe most of the structural universe

- Substantial Degeneracy Found
 - 635 TERMS describe over 50% of structural universe
 - 458,000 TERMS describes 99% of all structures in the PDB database

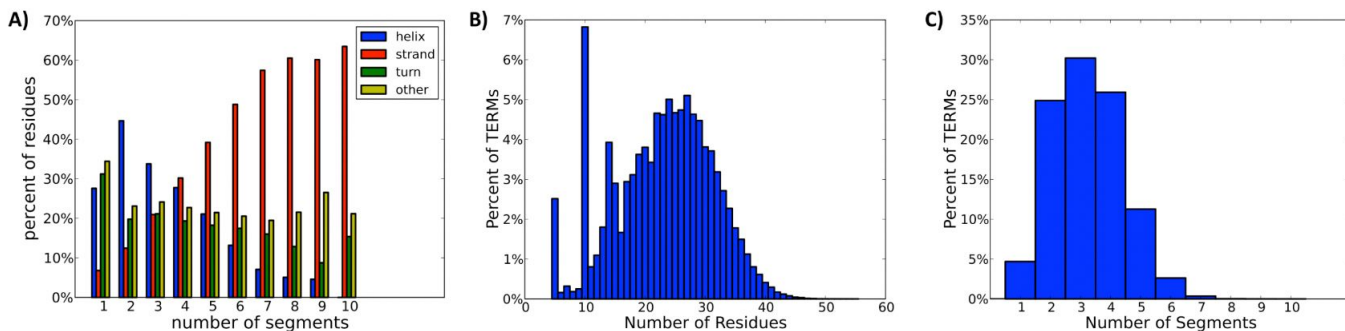
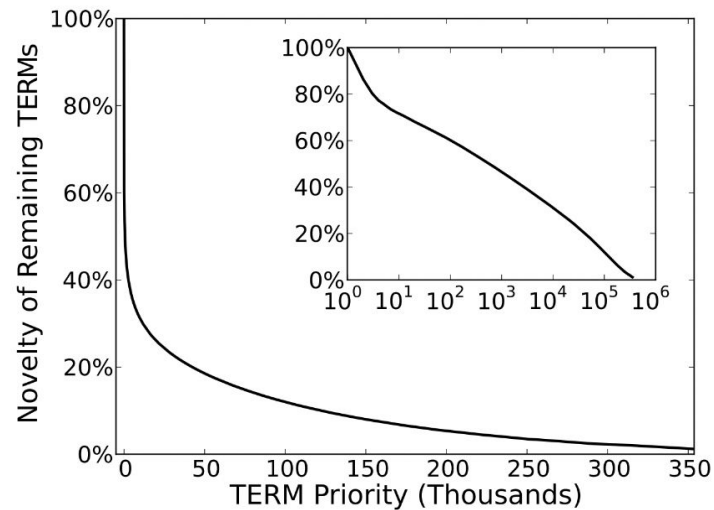


Figure S3 Statistics of universal TERMS. **A)** Secondary structure composition as a function of the number of segments. The Y-axis represents the percent of centroid residues, among TERMS with the given number of segments, that have a specific secondary structure classification (see SI Methods). **B)** The distribution of TERM sizes (i.e., the number of residues). **C)** The distribution of the number of TERM segment.

How many structures are novel?

- **Priority:**
 - Rank order TERMS based on their ability to capture the structural diversity of proteins
- **Novelty Measure:**
 - (# of structures covered by exclusively low priority) / (# of structure covered by low priority)



The PDB is Close to Saturated in TERMS

- Will new proteins still be representable by universal TERMS?
 - Take sequences with less than 35% sequence identity (~1,000)
 - High priority TERMS actually universal
 - New rare motifs likely to continue arising

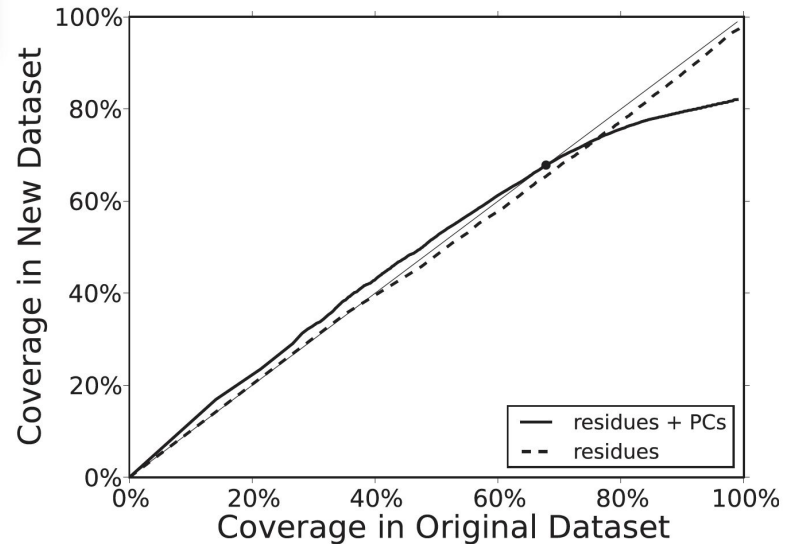


Fig. 3. Coverage in a test dataset of 1,095 proteins highly divergent from those used to create universal TERMS. The thick solid and dashed lines represent coverage of all universe elements and just residues, respectively. The thin line designates $x = y$. Up to 68% (indicated by a bold dot) the two sets are covered roughly identically.

Implications

TERMs Sequence Statistic enables design

- **Structure -> Sequence**
- **Experiment (use TERMs to reconstruct sequence and compare with native sequence)**
 - **Calculate positional self and pairwise pseudo energy**
 - **Self Pseudo energy: likelihood of amino acid at certain position**
 - **Interaction pseudo-energy: how often two specific amino acids interact at specific positions**
 - **Integer Linear programming to find lowest pseudo energy**

TERMs Sequence Statistic enables design

- X-Ray (Rigid Structures)
- NMR (Dynamic Structures)
 - TERM more advantageous at dynamic representations
 - More “loose” identifies sequence with broader structural patterns
 - Rosetta more advantageous with more precise measurements from X-Ray

Table 1. Sequence recovery results

Dataset*	Method	SID, % [†]	B/E, % [‡]	Cons., % [§]	Top 3, % [¶]	Cov., % [#]
X-ray-1 (66), 64	TERMs	29.3	30/27	49.5	51.8	96
	Rosetta	35.5	39/29	50.6		
	Combined	36.1	39/32	51.1		
X-ray-2 (67), 11	TERMs	26.7	30/22	50.8	49.9	98
	Rosetta	31.9	38/23	49.1		
	Combined	32.6	38/26	48.6		
NMR-1 (26), 5	TERMs	25.4	25/26	60.0	45.3	93
	Rosetta	28.3	32/22	54.6		
	Combined	32.0	33/29	55.3		
NMR-2 (67), 11	TERMs	20.6	23/17	44.8	54.6	90
	Rosetta	22.9	25/19	41.6		
	Combined	24.5	27/20	42.3		

TERMs Sequence Statistic enables design

- Sequences designed using TERMS only
24% identical to corresponding ones from Rosetta (orthogonal)
- Combined Method:
 - Use TERMS to limit amino acid choice to about 10 possibilities

Table 1. Sequence recovery results

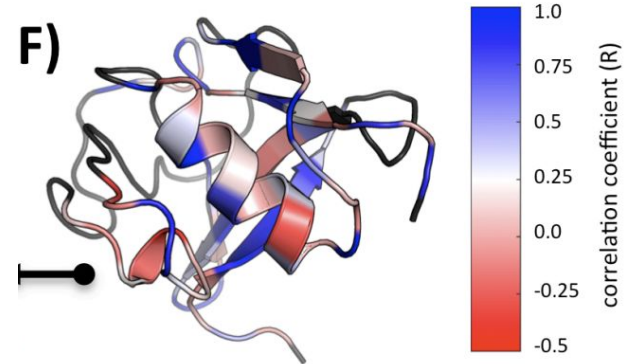
Dataset*	Method	SID, % [†]	B/E, % [‡]	Cons., % [§]	Top 3, % [¶]	Cov., % [#]
X-ray-1 (66), 64	TERMs	29.3	30/27	49.5	51.8	96
	Rosetta	35.5	39/29	50.6		
	Combined	36.1	39/32	51.1		
X-ray-2 (67), 11	TERMs	26.7	30/22	50.8	49.9	98
	Rosetta	31.9	38/23	49.1		
	Combined	32.6	38/26	48.6		
NMR-1 (26), 5	TERMs	25.4	25/26	60.0	45.3	93
	Rosetta	28.3	32/22	54.6		
	Combined	32.0	33/29	55.3		
NMR-2 (67), 11	TERMs	20.6	23/17	44.8	54.6	90
	Rosetta	22.9	25/19	41.6		
	Combined	24.5	27/20	42.3		

TERMs Explain Evolutionary Variation

- Can TERMs explain evolutionary variation?
 - Used pseudo-energies to predict sequence variations for various proteins
 - Small variants using Monte Carlo simulations
 - Using BLAST they found actual evolutionary variations (MSAs)
 - Compared frequency of amino acids in predicted and actual evolutionary sequences

TERMs Explain Evolutionary Variation

- **Correlation: 0.51** between predicted and actual evolutionary variants
- **TERM's top amino acid prediction** matched evolution **35% time**
- **These fractions are higher than native sequence recovery rate**
 - **TERMs are able to capture broader evolutionary trends better than exact structures**

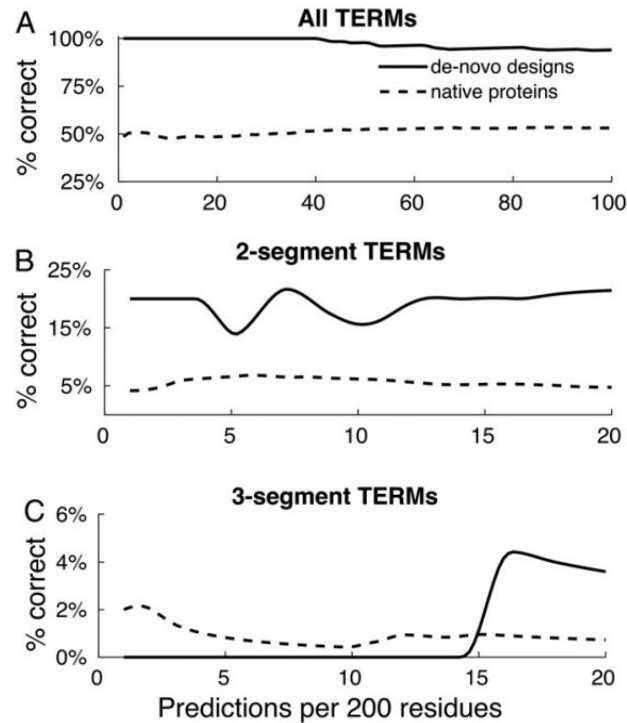


TERMs Map Sequence to Structure

- **Sequence -> Structure**
- **Challenges:**
 - **TERM only encompasses partial structure large influence from environment**
 - **Number of possible alignments of a TERM grows exponentially with increase in disjoint segments**
- **Weak Coupling Framework**
 - **Statistical approach**
 - **Score structure based on 4,000 highest priority TERMS between sequence to TERM**

TERMs Sequence to Structure

- Better performance on de novo structure
 - Have more consistent structures
- Performs poorly on multisegment TERMs



Discussion

Discussion/Summary

- **Objective: Develop a systematic decomposition of protein structure space to understand structure-sequence relationships.**
- **TERMs can provide:**
 - **Structure -> Sequence insight**
 - **Evolutionary insight**
 - **Sequence -> Structure**
- **Why do TERMs recur: Likely due to both biophysical principles and evolutionary constraints.**

Questions

**How are TERMs structures used in current times with Alphafold2?
Do TERMs still provide any valuable insight into protein structure?**

Discuss: 2 min

Thank you!