

# **Decoding the Structural Keywords in Protein Structure Universe**

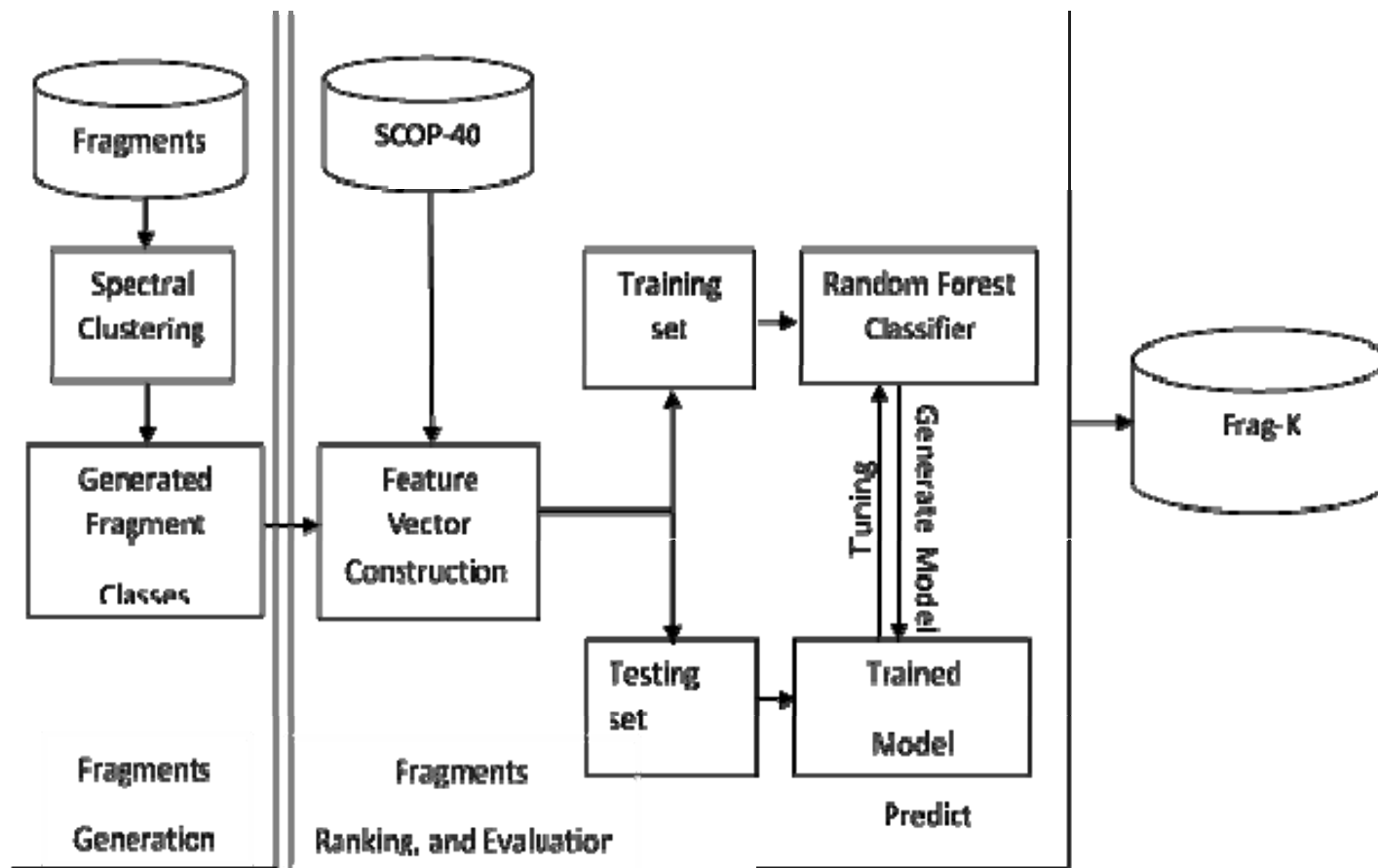
**Wessam Elhefnawy, Min Li, Jianxin Wang, and Yaohang Li**

Elhefnawy W, Li M, Wang JX et al. Decoding the structural keywords in protein structure universe. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 34(1): 3–15 Jan. 2019. DOI 10.1007/s11390-019-1895-y

# Frag-K

- a protein structural dictionary
- composed of a set of backbone fragments ranging from 4 to 20 residues
- effectively distinguish between major protein folds
- ~400 4- to 20-residue Frag-K fragments is capable of classifying major SCOP folds with high accuracy

# Generation of Frag-K Library



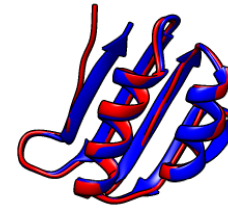
# Frag-K outperforms Fragbag with better precision, recall, and F-measure in classifying protein folds in EDD dataset

	Fragbag			Frag-K					
	L4 to L12			L4 to L12			L4 to L20		
SCOP Fold Classes	Precision	Recall	F	Precision	Recall	F	Precision	Recall	F
a.39	1.00	0.60	0.75	1.00	0.67	0.80	1.00	0.87	0.93
c.23	0.90	0.69	0.78	0.78	0.90	0.84	0.85	0.96	0.90
c.47	0.79	0.65	0.71	0.95	0.91	0.93	0.95	0.89	0.92
c.69	0.89	0.74	0.81	1.00	0.96	0.98	1.00	1.00	1.00
a.1	0.80	0.80	0.80	0.90	0.90	0.90	1.00	0.90	0.95
a.3	1.00	0.14	0.25	0.71	0.71	0.71	0.88	1.00	0.93
c.2	0.75	0.78	0.76	0.98	0.95	0.96	1.00	0.98	0.99
c.3	1.00	0.32	0.48	1.00	0.82	0.90	1.00	0.95	0.98
c.1	0.66	0.94	0.78	0.85	0.96	0.90	0.98	0.96	0.97
b.1	0.64	0.80	0.71	0.90	0.94	0.92	0.93	0.98	0.95
b.6	0.25	0.12	0.17	0.38	0.38	0.38	1.00	0.63	0.77
b.40	0.84	0.36	0.50	0.78	0.80	0.79	0.93	0.84	0.88
b.42	0.40	0.18	0.25	1.00	0.82	0.90	1.00	1.00	1.00
c.93	1.00	0.56	0.71	0.92	0.67	0.77	1.00	1.00	1.00
b.47	1.00	0.89	0.94	0.80	0.89	0.84	1.00	0.89	0.94
b.60	1.00	0.62	0.77	1.00	0.75	0.86	1.00	0.75	0.86
c.37	0.86	0.52	0.65	1.00	0.80	0.89	0.95	0.90	0.92
b.29	0.83	0.36	0.50	1.00	0.71	0.83	1.00	0.93	0.96
g.3	0.39	0.64	0.48	0.96	0.92	0.94	0.91	0.84	0.87
a.26	0.33	0.12	0.18	0.63	0.63	0.63	0.75	0.75	0.75
a.24	0.73	0.53	0.62	0.89	0.53	0.67	0.83	0.67	0.74
c.55	0.67	0.19	0.30	0.39	0.67	0.49	0.61	0.90	0.73
d.15	0.56	0.70	0.62	0.88	0.81	0.85	0.93	0.93	0.93
a.4	0.56	0.93	0.70	0.82	0.97	0.89	0.89	0.99	0.93
d.58	0.61	0.79	0.69	0.76	0.76	0.76	0.84	0.84	0.84
b.121	0.75	0.55	0.63	0.80	0.73	0.76	1.00	0.91	0.95
b.34	0.62	0.49	0.55	0.90	0.70	0.79	0.86	0.81	0.83
<b>Avg/total</b>	<b>0.73</b>	<b>0.56</b>	<b>0.60</b>	<b>0.85</b>	<b>0.79</b>	<b>0.81</b>	<b>0.93</b>	<b>0.89</b>	<b>0.90</b>

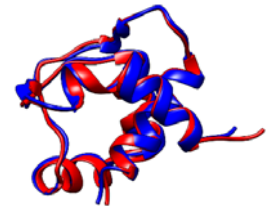
# Top-200 most effective Fold-Keys fragments for fold classification



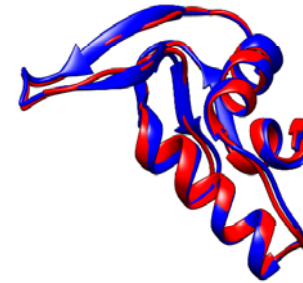
# Frag-K to assemble protein structures with global fit



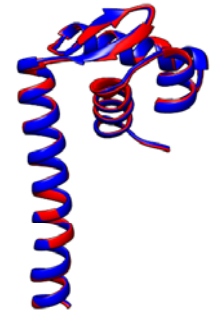
d4maka\_d.58: Ferredoxin-like  
*alpha+beta sandwich with antiparallel beta-sheet;*  
*(beta-alpha-beta), 0.51Å.*



d4j20a\_a.3: Cytochrome  
*core: 3 helices; folded leaf, opened, 0.64Å.*



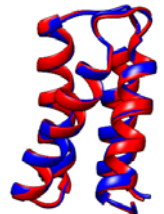
d1dp7p\_a.4: DNA/RNA-binding 3-helical bundle  
*core: 3-helices; bundle, closed or partly opened, right-*  
*handed twist; up-and down, 1.41Å.*



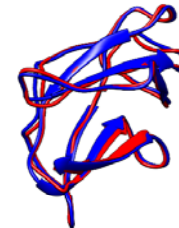
d1r7ja\_a.4: DNA/RNA-binding 3-helical bundle  
*core: 3-helices; bundle, closed or partly opened,*  
*right-handed twist; up-and down, 1.27Å*



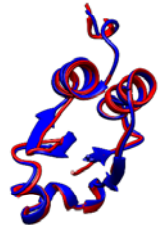
d2ve8a\_a.4: DNA/RNA-binding 3-helical bundle  
*core: 3-helices; bundle, closed or partly opened, right-*  
*handed twist; up-and down, 0.40Å.*



d1l1a1\_a.24: Four-helical up-and-down bundle  
*core: 4 helices; bundle, closed or partly opened,*  
*handed twist; up-and-down, 0.39Å.*



d3uzqb\_b.1: Immunoglobulin-like beta-sandwich  
*sandwich; 7 strands in 2 sheets; Greek-key, 1.78Å*



d3eina1\_c.47: Thioredoxin fold  
*core: 3 layers, a*  
*mixed beta-sheet of 4 strands, order 4312; strand*  
*antiparallel to the rest, 0.65Å.*