BME4005 Assignment 5

Ran Hu 116010078

July 23, 2018

Problem: To study the relation between residue fluctuations (B-factor) and its evolutionary conservation (conservation scores or substitution rate).

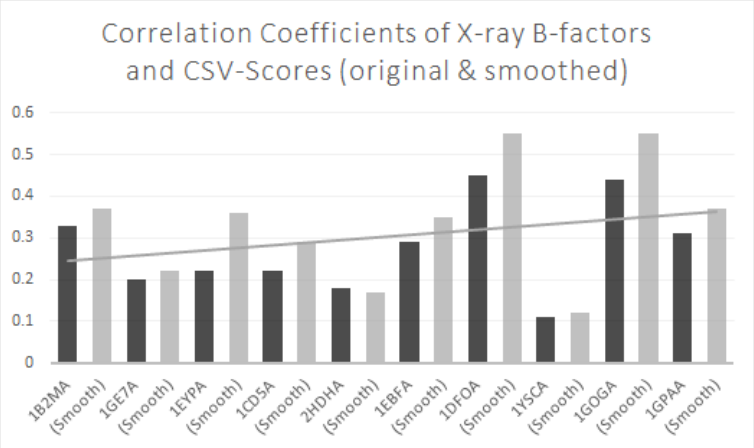
The 10 proteins I selected are: 1B2MA, 1GE7A, 1EYPA, 1CD5A, 2HDHA, 1EBFA, 1DFOA, 1YSCA, 1GOGA, 2GPAA. Their index in the list are: 3, 53, 104, 179, 219, 338, 401, 403, 518, 542.

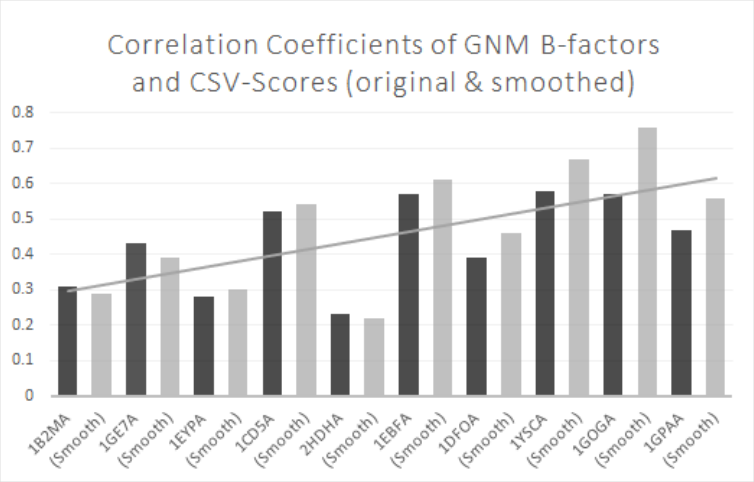
I calculated the Pearson's correlation coefficients using the cut-off radius ranging from 7 to 21 (integers). Then I use 5 points to "smooth" conservation scores and compare the smoothed conservation scores with GNM and X-ray B-factors. The results are as follows:

	Index	Diameter								GNM								X-ray	GNM_max	cut-off_max
Cut-off			7	8	9	10	11	12	13	14	15	16	17	18	19	20	21			
1B2MA	3	35.59	0.18	0.28	0.29	0.3	0.3	0.31	0.26	0.27	0.25	0.26	0.24	0.27	0.31	0.3	0.31	0.33	0.31	12
(Smooth)			0.05	0.13	0.15	0.18	0.15	0.2	0.17	0.19	0.18	0.19	0.19	0.23	0.27	0.29	0.28	0.37	0.29	20
1GE7A	53	45.62	0.33	0.35	0.37	0.4	0.42	0.41	0.43	0.4	0.42	0.41	0.41	0.42	0.42	0.42	0.4	0.2	0.43	13
(Smooth)			0.29	0.31	0.3	0.3	0.32	0.32	0.35	0.35	0.34	0.34	0.37	0.38	0.38	0.39	0.38	0.22	0.39	20
1EYPA	104	46.53	0.27	0.28	0.24	0.26	0.26	0.22	0.23	0.21	0.19	0.21	0.21	0.2	0.2	0.22	0.22	0.22	0.28	8
(Smooth)			0.3	0.28	0.26	0.24	0.22	0.2	0.2	0.2	0.19	0.23	0.22	0.22	0.23	0.24	0.26	0.36	0.3	7
1CD5A	179	52.94	0.38	0.43	0.37	0.46	0.47	0.5	0.52	0.5	0.49	0.5	0.5	0.49	0.49	0.48	0.47	0.22	0.52	13
(Smooth)			0.41	0.42	0.35	0.43	0.44	0.46	0.47	0.48	0.5	0.51	0.52	0.52	0.54	0.53	0.53	0.29	0.54	19
2HDHA	219	60.3	0.02	0.07	0.09	0.1	0.14	0.16	0.16	0.16	0.17	0.18	0.18	0.21	0.21	0.22	0.23	0.18	0.23	21
(Smooth)			-0.06	0	-0.01	0.01	0.06	0.1	0.12	0.13	0.15	0.17	0.17	0.2	0.21	0.22	0.22	0.17	0.22	21
1EBFA	338	67.31	0.55	0.57	0.55	0.54	0.54	0.53	0.54	0.52	0.52	0.53	0.53	0.54	0.55	0.56	0.57	0.29	0.57	8
(Smooth)			0.61	0.6	0.58	0.52	0.52	0.51	0.51	0.5	0.53	0.55	0.56	0.57	0.59	0.6	0.61	0.35	0.61	21
1DFOA	401	64.88	0.28	0.33	0.33	0.32	0.32	0.34	0.36	0.37	0.37	0.38	0.39	0.38	0.38	0.37	0.37	0.45	0.39	17
(Smooth)			0.35	0.4	0.4	0.39	0.38	0.4	0.42	0.44	0.44	0.45	0.46	0.45	0.45	0.45	0.45	0.55	0.46	17
1YSCA	403	60.12	0.49	0.57	0.54	0.54	0.51	0.55	0.56	0.56	0.57	0.58	0.56	0.57	0.56	0.57	0.57	0.11	0.58	16
(Smooth)			0.51	0.55	0.54	0.53	0.52	0.58	0.6	0.62	0.64	0.65	0.64	0.66	0.66	0.67	0.67	0.12	0.67	21
1GOGA	518	79.1	0.55	0.56	0.57	0.56	0.55	0.56	0.56	0.57	0.57	0.56	0.57	0.57	0.57	0.57	0.57	0.44	0.57	20
(Smooth)			0.63	0.64	0.7	0.68	0.67	0.7	0.71	0.72	0.72	0.72	0.74	0.74	0.74	0.75	0.76	0.55	0.76	21
1GPAA	542	89.84	0.47	0.28	0.29	0.38	0.39	0.4	0.38	0.37	0.37	0.37	0.41	0.41	0.42	0.42	0.44	0.31	0.47	7
(Smooth)			0.56	0.33	0.33	0.42	0.41	0.43	0.42	0.42	0.42	0.43	0.49	0.5	0.51	0.52	0.54	0.37	0.56	7
Average			0.352	0.372	0.364	0.386	0.39	0.398	0.4	0.393	0.392	0.398	0.4	0.406	0.411	0.413	0.415	0.275	0.415	21
(Smooth)			0.365	0.366	0.36	0.37	0.369	0.39	0.397	0.405	0.411	0.424	0.436	0.447	0.458	0.466	0.47	0.335	0.47	21

The average correlation coefficient of X-ray B-factors and smoothed conservation scores is 0.335. The average correlation coefficient of GNM B-factors and smoothed conservation scores is 0.47. The best cut-off radius among 7 to 21 is 21.

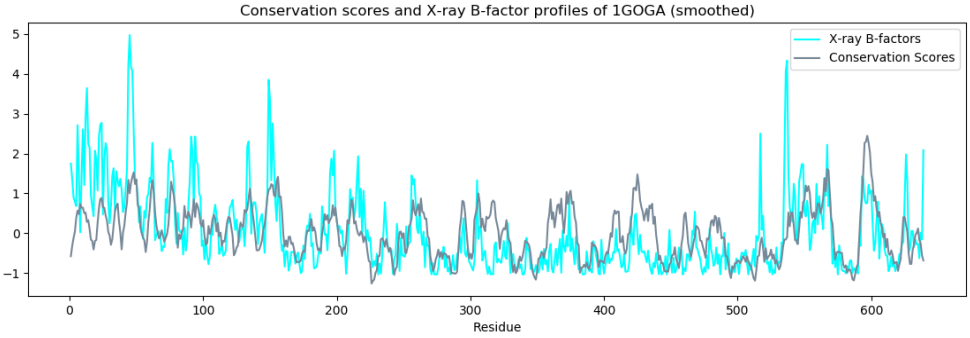
The results of all the 10 proteins are as follows. As is shown in the histogram, after smoothing the conservation scores, the results are generally better. The larger the protein is, the more obvious the difference is. And the trend shows that the correlation is usually larger for large proteins than small proteins. It may be because that smaller proteins are more flexible and unpredictable.

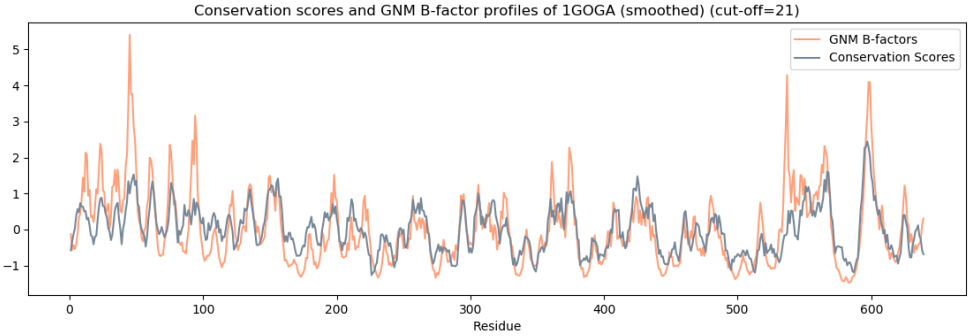




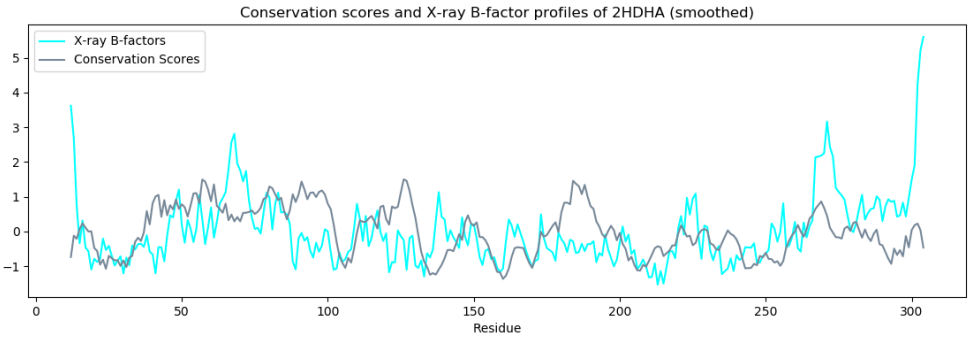
After drawing the graphs of each protein's correlation coefficient of conservation scores and GNM B-factors, as well as the correlation coefficient of conservation scores and X-ray B-factors, I find they are very consistent. This proves that we do can get evolutionary conservation information merely from the structural information.

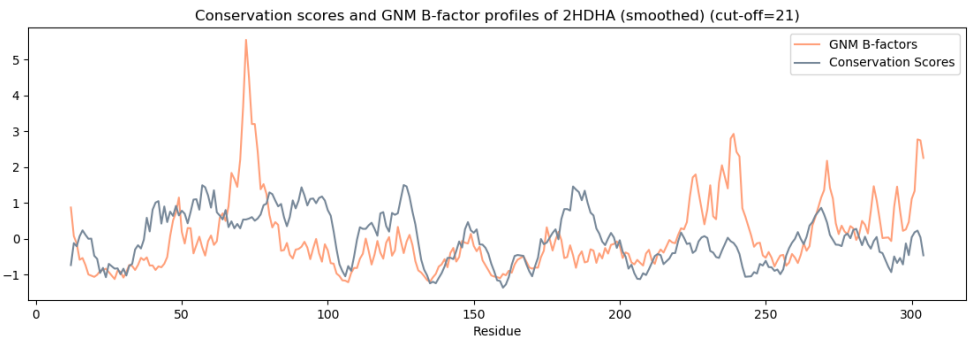
Here is the best correlation I get: 1GOGA (X-ray: 0.55, GNM: 0.76)



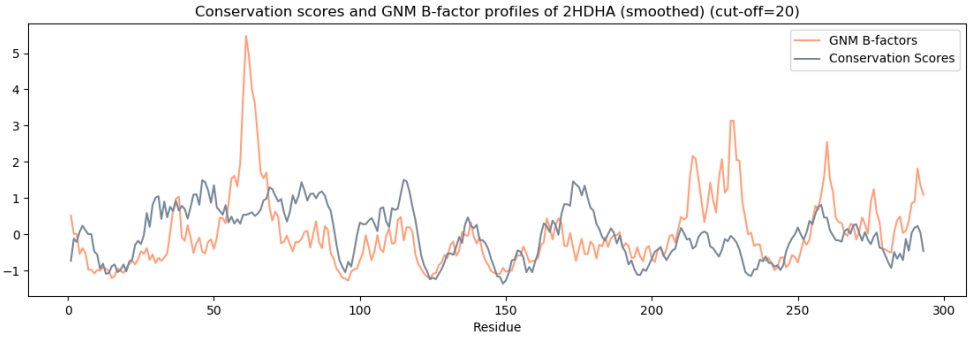


However, some proteins fail to have satisfying results. For example: 2HDHA (X-ray: 0.17, GNM: 0.22)





I found that there are many missing residues in 2HDHA's sequence. But I cannot get the complete sequece of it. I used $(PS)^2$ to get the new pdb file to compute GNM B-factors. The result is better (GNM: 0.29).

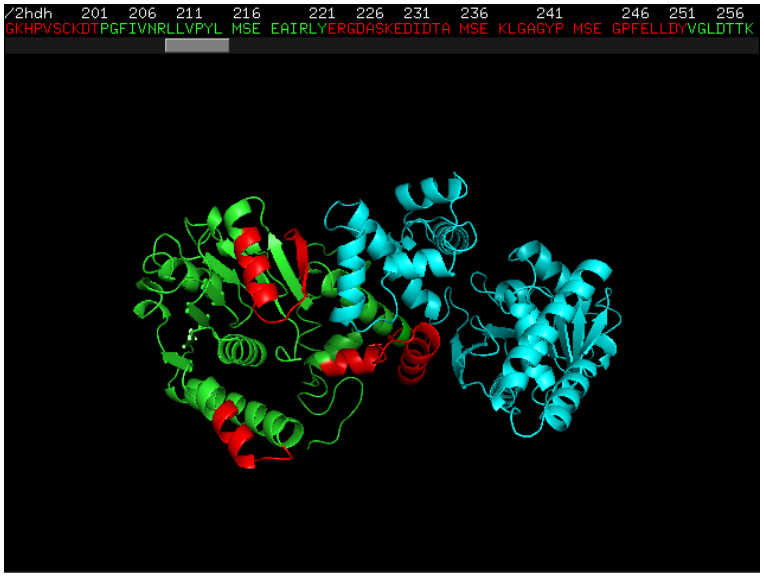


The new position information $(PS)^2$ provides is beneficial for us to get better results.

	Index	Diameter								GNM								X-ray	GNM_max	cut-off_max
Cut-off			7	8	9	10	11	12	13	14	15	16	17	18	19	20	21			
2HDHA	219	58.73	0.13	0.2	0.19	0.22	0.22	0.23	0.24	0.24	0.24	0.25	0.26	0.27	0.28	0.28	0.29	0.01	0.29	21
(Smooth)			0.06	0.15	0.12	0.15	0.16	0.19	0.21	0.21	0.23	0.24	0.25	0.27	0.28	0.29	0.29	0.03	0.29	20

2HDHA	219	60.3	0.02	0.07	0.09	0.1	0.14	0.16	0.16	0.16	0.17	0.18	0.18	0.21	0.21	0.22	0.23	0.18	0.23	21
(Smooth)			-0.06	0	-0.01	0.01	0.06	0.1	0.12	0.13	0.15	0.17	0.17	0.2	0.21	0.22	0.22	0.17	0.22	21

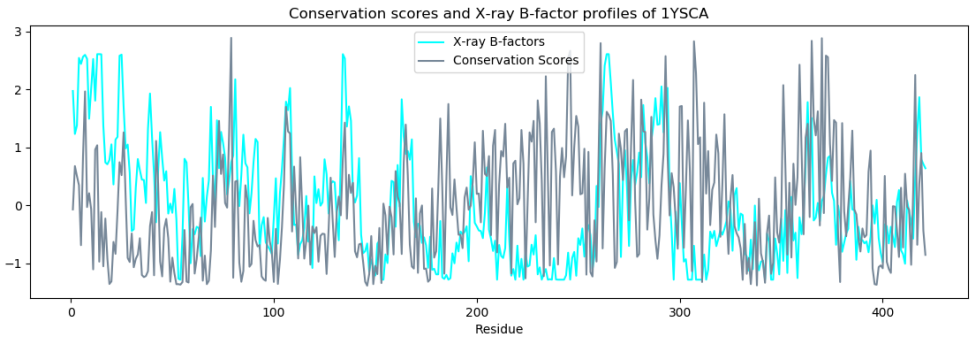
Furthermore, I selected the residues which do not have good agreement in the graphs (red), and found that these region are on the surface of the 2HDH, and some are on the contact surface between two chains.



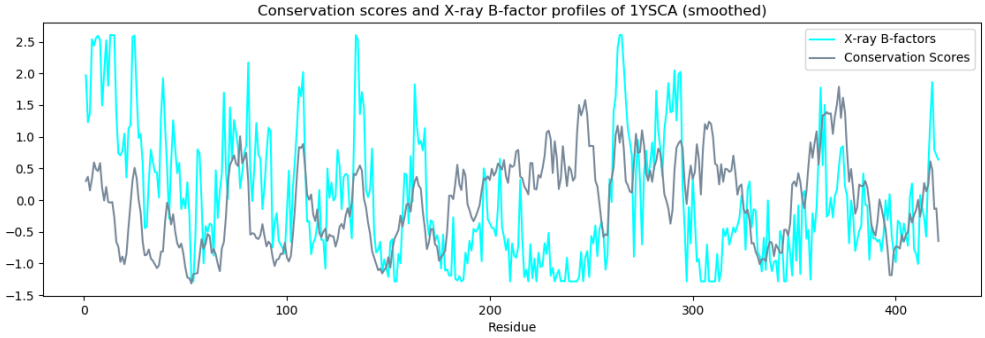
I guessed the other chain will influence the dynamic of residues. And I draw 1EBF, which also has two chains, but good correlation (X-ray: 0.35, GNM: 0.61, after smoothing the conservation scores). The red region are residues which do not have good agreement in the graphs, they are also on the surface or contact region, but do not influence the results much.

Smoothing can help moderate curves and gives better results. But for some proteins, the correlation is not good, and smoothing conservation scores does not help much. Take 1YSCA for example.

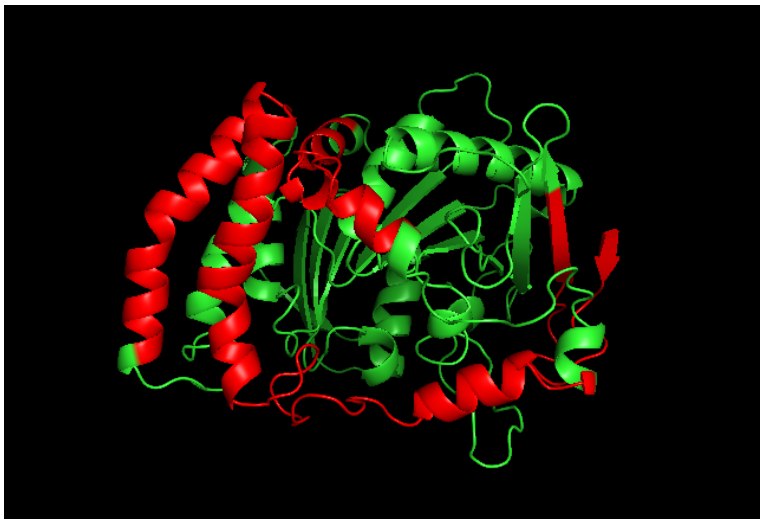
Orignial (X-ray: 0.11):



Smoothed (X-ray: 0.12):

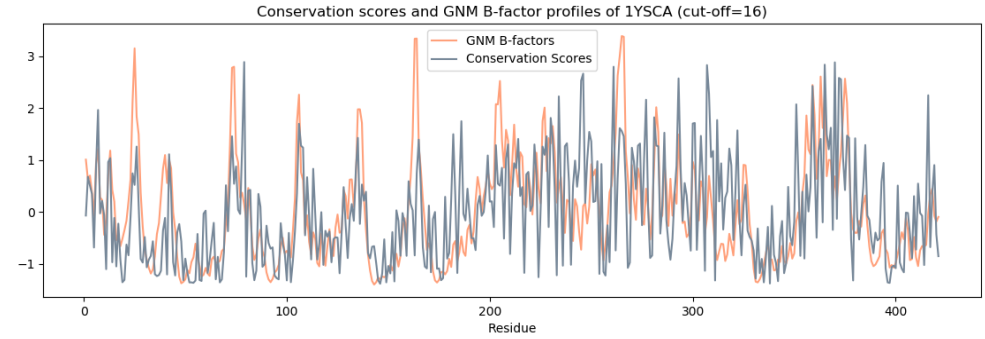


The area which has poor correspondence are on the surface.

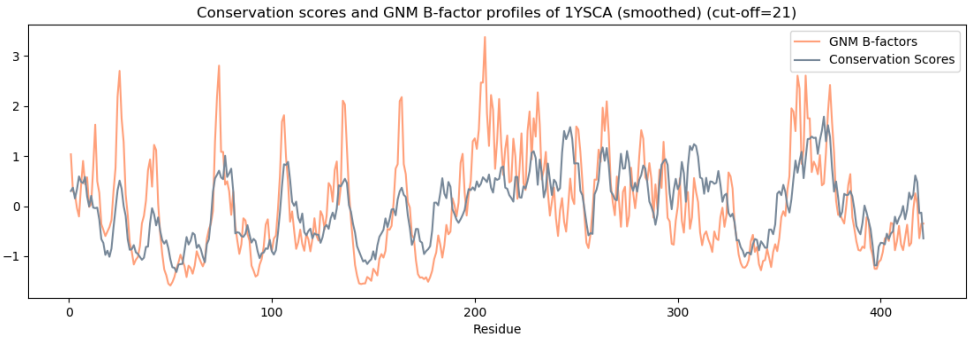


But 1YSCA has good GNM B-factors and conservation scores correspondance (original: 0.58, smoothed: 0.67):

Original (GNM: 0.58):

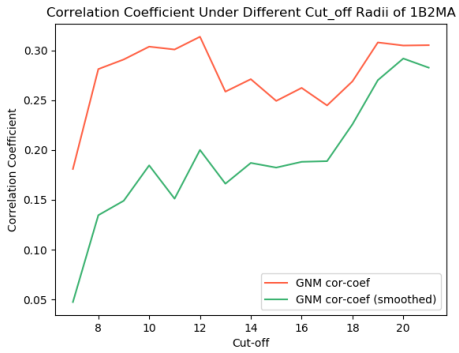


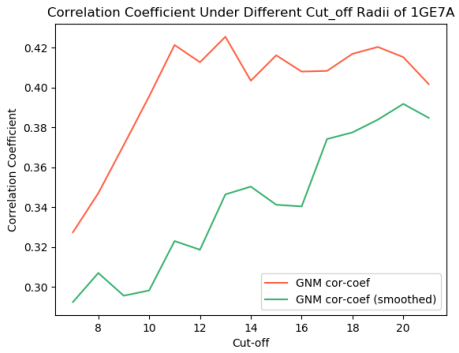
Smoothed (GNM: 0.67):

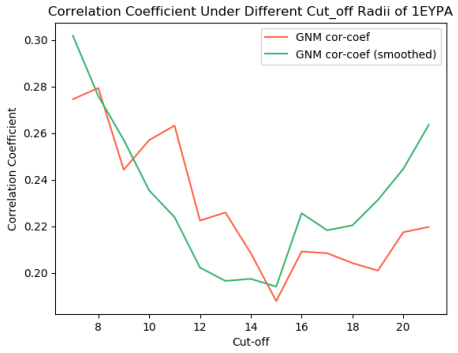


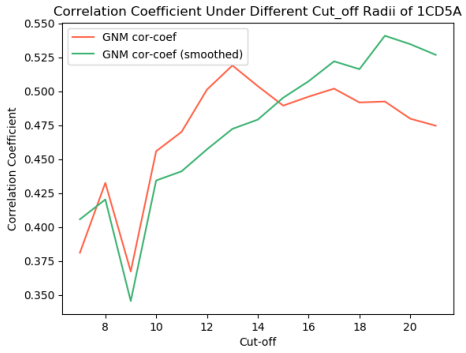
I think this example shows that GNM B-factor can better describe evolutionary conservation than X-ray B-factors, because there are many experimental errors.

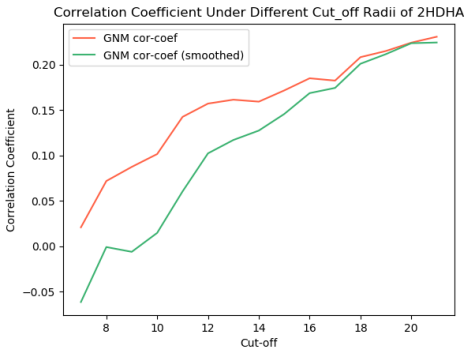
For different proteins, the influences of cut-off radii vary a lot.

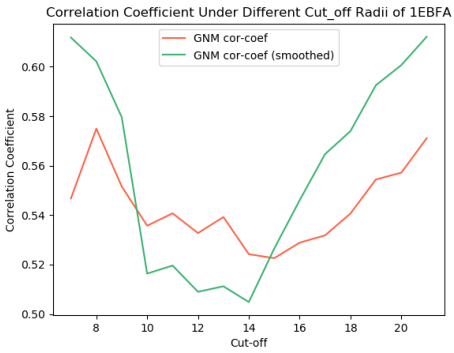


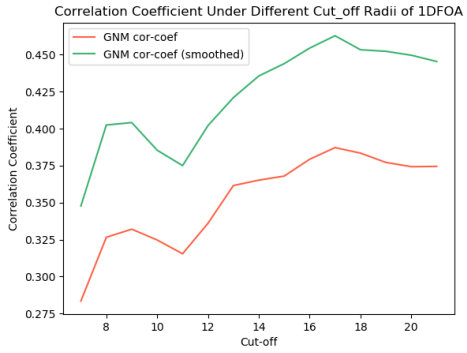


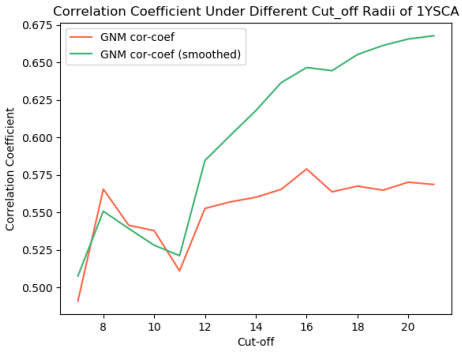


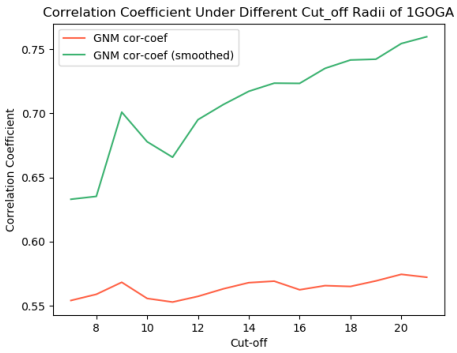


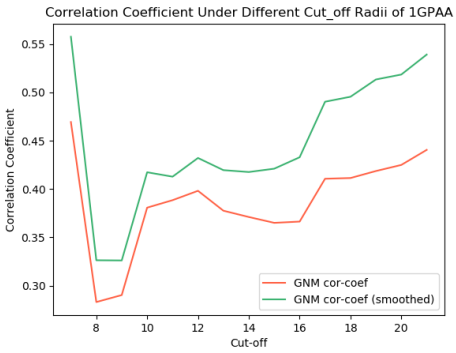












Although the trends are different, most proteins have better results as the radius becomes larger. I think it confirms with the evidence that the results of WCN is better than GNM in general, since WCN use all the pairwise reidues distances to build the model.