Computational Prediction of Protein-DNA Interactions

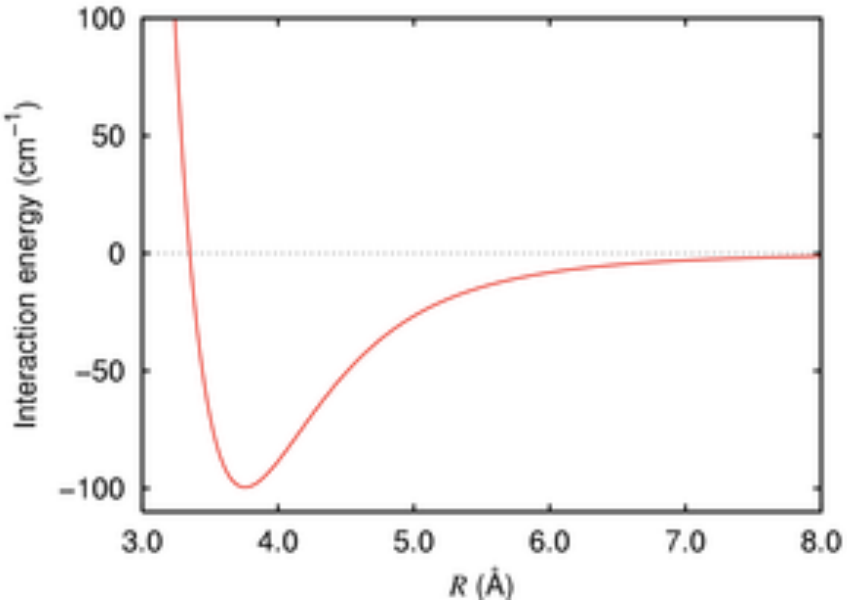
Xide Xia
Advisor: Dr. Mohammed AlQuraishi

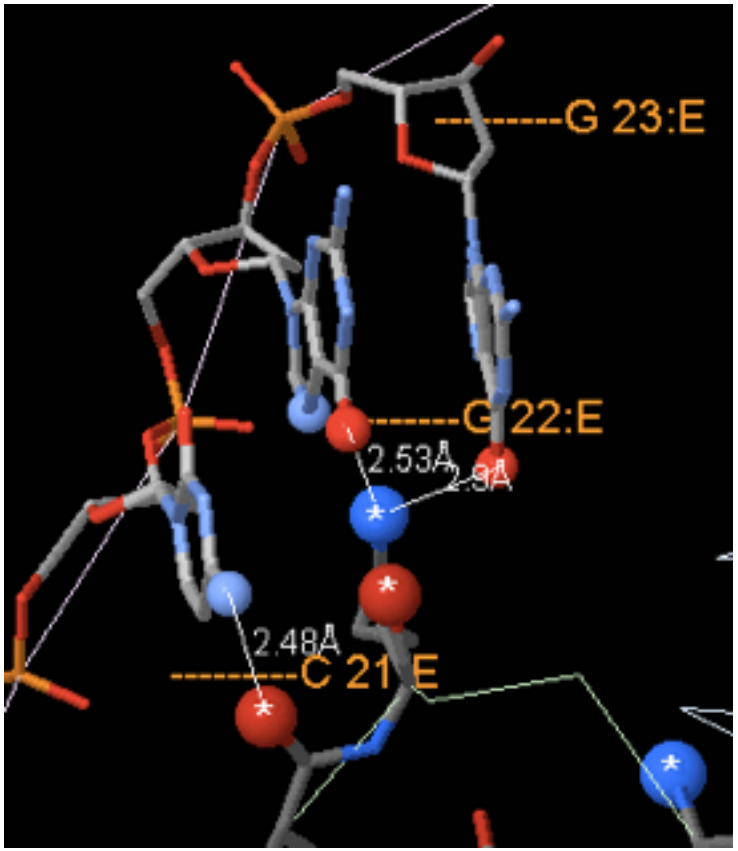
Position Weight Matrix (PWM)

	1	2	3	4	5	6	7	8	9	10
A	0.0625	0.375	0.5625	0.2500	0.8125	0.1875	0.3125	0.125	0.3750	0.4375
C	0.3125	0.500	0.1875	0.1875	0.0625	0.5625	0.1875	0.375	0.4375	0.1875
G	0.5000	0.125	0.2500	0.0625	0.0000	0.1250	0.4375	0.125	0.0625	0.1250
T	0.1250	0.000	0.0000	0.5000	0.1250	0.1250	0.0625	0.375	0.1250	0.2500

PWMs are often represented graphically as sequence logos.

Phase 1. Select a Model





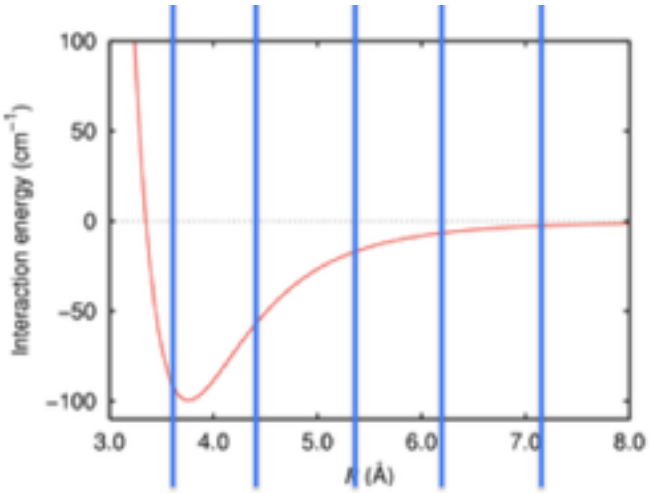
Atom pairs: {Protein atom, DNA atom} 27*37 = 999

Protein

{Ala | Cys | Ile | Leu | Met | Val, C, β} → 1
{Ile, C, γ1 | γ2 | δ1} → 2
{Leu, C, γ | δ1 | δ2} → 2
{Met, C, γ | ε} → 2
{Val, C, γ1 | γ2} → 2
{Met, S, δ} → 3
{Cys, S, γ} → 4
{His | Phe | Trp | Tyr, C, β} → 5
{His | Phe | Trp | Tyr, C, γ | δ1 | δ2 | ε1 | ε2 | ε3
{Trp, N, ε1} → 7
{Tyr, O, η} → 8
{His, N, δ1 | ε2} → 9
{Asn | Gln | Thr | Ser, C, β} → 10
{Asn, O, δ1} → 11
{Gln, O, ε1} → 11
{Thr, O, γ1} → 11
{Ser, O, γ} → 11
{Gln | Thr, C, γ | γ2} → 12
{Asn, C, γ} → 13
{Gln, C, δ} → 13
{Asn, N, δ2} → 14
{Gln, N, ε2} → 14
{Arg | Lys, C, β} → 15
{Arg, C, γ | δ} → 16
{Lys, C, γ | δ | ε} → 16
{Arg, N, η1 | η2} → 17
{Lys, N, ζ} → 17
{Arg, C, ζ} → 18
{Arg, N, ε} → 19
{Glu, C, β | γ} → 20
{Asp, C, β} → 20
{Glu, C, δ} → 21
{Asp, C, γ} → 21
{Glu, O, ε1 | ε2} → 22
{Asp, O, δ1 | δ2} → 22
{Pro, C, β | γ | δ} → 23
{_, N, } → 24
{_, C, α} → 25
{_, C, } → 26
{_, O, } → 27

DNA

{_, O, P1 | P2} → 1
{_, P, } → 2
{_, O, 5'} → 3
{_, C, 5'} → 4
{_, C, 4'} → 5
{_, C, 3'} → 6
{_, C, 2'} → 7
{_, C, 1'} → 8
{_, O, 4'} → 9
{_, O, 3'} → 10
{DA | DG, N, 9} | {DT | DC, N, 1}
{DA | DG, C, 8} → 12
{DA | DG, N, 7} → 13
{DA | DG, C, 5} → 14
{DA | DG, C, 4} → 15
{DA | DG, N, 3} → 16
{DA, C, 2} → 17
{DA, N, 1} → 18
{DA, C, 6} → 19
{DA, N, 6} → 20
{DG, C, 2} → 21
{DG, N, 2} → 22
{DG, C, 6} → 23
{DG, O, 6} → 24
{DT | DC, C, 6} → 25
{DC, C, 5} → 26
{DC, C, 4} → 27
{DC, N, 3} → 28
{DT | DC, C, 2} → 29
{DT | DC, O, 2} → 30
{DT, C, 5} → 31
{DT, C, 7} → 32
{DT, C, 4} → 33
{DT, O, 4} → 34
{DT, N, 3} → 35
{DC, N, 4} → 36
{DG, N, 1} → 37



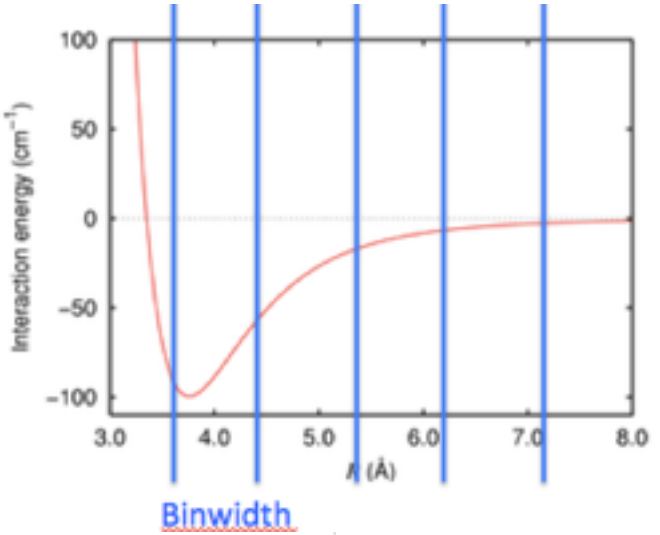
Binwidth

$$E_{A,B} = \sum_{a \in A, b \in B} \alpha_{a,b}\, Distbin(a,b)$$

β_A

A			T			G			C		
1	...	N	1	...	N	1	...	N	1	...	N
0 3 ... 1 5											

4* N* 999



Binwidth

A											
1* Binwidth			...	n* Binwidth				...	N* Binwidth		
0	3	...	1	5		4	10	...	5	9	

N* 999

$$E_{A,B} = \sum_{a \in A, b \in B} \alpha_{a,b}\, Distbin(a,b)$$

{atom1 ID, atom2 ID}	{1,1}	{1,2}	{1,3}	...	{i, j}	...	{27,37}
Count of times of {atom1, atom2} appears within dist0 = n* Binwidth	C_1	C_2	C_3	...	C_n	...	C_{999}

999

3DNA: Base mutation ➔ Input Feature Vecto $\boldsymbol{\beta_i = \{\beta_A, \beta_T, \beta_G, \beta_C\}}$

Output vector ***P***: ***PWM***

Kullback-Leibler divergence (KLD)

$$D_{KL}(P||Q) = \sum_i P(i)\mathrm{Ln}\left(\frac{P(i)}{Q(i)}\right)$$

PWM (P)

Prediction (Q)

KLD = 1.3

PWM (P)

Prediction (Q)

KLD = 4.4

Custom Model (Similar to Multinomial Logistic Regression)

Define: X (Length = 999*N)

$$E_{A,B} = \sum_{a \in A, b \in B} \alpha_{a,b}\, Distbin(a,b) \quad \rightarrow \quad X \cdot \beta_A$$

$$\Pr(Y = A) = \frac{e^{\beta_A X}}{1 + \sum_{k=1}^{4} e^{\beta_k X}}, \quad \Pr(Y = T) = \frac{e^{\beta_T X}}{1 + \sum_{k=1}^{4} e^{\beta_k X}}$$

$$\Pr(Y = G) = \frac{e^{\beta_G X}}{1 + \sum_{k=1}^{4} e^{\beta_k X}}, \quad \Pr(Y = C) = \frac{e^{\beta_C X}}{1 + \sum_{k=1}^{4} e^{\beta_k X}}$$

Prediction Q

Goal: Minimizing $D_{KL}(P||Q) = \sum_i P(i) \mathrm{Ln}\left(\frac{P(i)}{Q(i)}\right)$

CVX (a Matlab-based modeling system for convex optimization)

```
cvx_begin
        variable X
        Q = exp(X * phi)
        penalty = KL_div(P, Q)

        minimize (penalty + |X|)
cvx_end
```

Data (size = 700)

Data	PWM	Protein Sequence	Protein Structure	DNA Structure
Phase 1	✓	✓	✓	✓

Result

	Nbins=2	Nbins=3	Nbins=4	Nbins=5
KLD	***2.5421***	***2.4352***	***2.5435***	***2.5641***

Binwidth = 1.3Å

Phase 2. Train Model on More Structure Data

Data	PWM	Protein Sequence	Protein Structure	DNA Structure
Phase 1	✓	✓	✓	✓
Phase 2	✓	✓	✓	✓

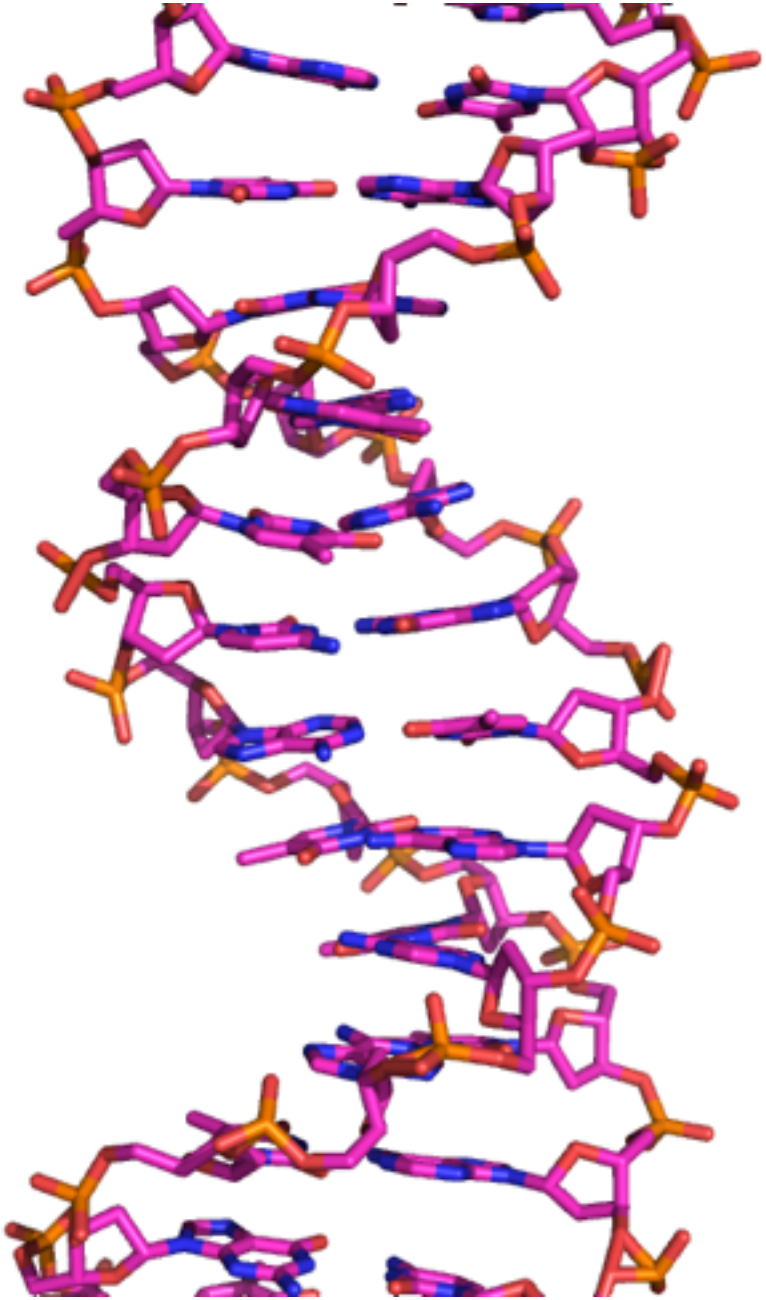
3d-footprint

```
>1a02_FJN:      STRUCTURE OF THE DNA BINDING DOMAINS OF NFAT, FOS AND JUN  BOUND TO DNA        organism=HOMO SAPIENS
IC=12.114 |tag=multimer
rrirrerNkmAAaksRnrrreltdtlqaetdqledeksalqteianllkekek/rkrmrNriAaskSRkrkleriarleekvktlkaqnselastanmlreqvaql/wplssqsgsyel
rievqpkphhRahYetEgsRgavkaptgghpvvqlhgymenkplglqifigtaderilkphafyqvhritgktvtttsyekivgntkvleiplepknnmratidcagilklrnadielr
kgetdigRkntrvrlvfrvhipessgrivslqtasnpiecsQRsahelpmverqdtdsclvyggqqmiltgqnftseskvvftekttdgqqiwemeatvdkdksqpnmlfveipeyrnk
hirtpvkvnfyvingkrkrsqpqhftyhpv  interface= F:8,11,12,16, J:6,9,13,14, N:23,26,29,32,139,173,174,
A |   0   0  96  68  24  59  24  18   0   0  24   6   0   0  96
C |   0   0   0   4  24  13  24  13   4   0  24  13   0  96   0
G |  96  96   0  18  24  11  24  11   0  96  24   4   0   0   0
T |   4   0   0   6  24  13  24  54  92   0  24  73  96   0   0
>1a02_N:p53-like_transcription_factors;E_set_domains;   STRUCTURE OF THE DNA BINDING DOMAINS OF NFAT, FOS AND JUN
BOUND TO DNA    organism=HOMO SAPIENS   IC=3.592 |tag=redundant
wplssqsgsyelrievqpkphhRahYetEgsRgavkaptgghpvvqlhgymenkplglqifigtaderilkphafyqvhritgktvtttsyekivgntkvleiplepknnmratidcag
ilklrnadielrkgetdigRkntrvrlvfrvhipessgrivslqtasnpiecsQRsahelpmverqdtdsclvyggqqmiltgqnftseskvvftekttdgqqiwemeatvdkdksqpn
mlfveipeyrnkhirtpvkvnfyvingkrkrsqpqhftyhpv      interface= N:23,26,29,32,139,173,174,
A |   3   0  94  41
C |   3   0   0  12
G |  86  96   1  28
T |   4   0   1  15
>1a0a_AB:       PHOSPHATE SYSTEM POSITIVE REGULATORY PROTEIN PHO4/DNA  COMPLEX  organism=SACCHAROMYCES CEREVISIAE
IC=7.140 |tag=multimer
mKResHkhaEqaRRnrlavalhelaslipaewkqqnvsaapskattveaacryirhlqqngst/mkResHKhaEqaRRnrlavalhelaslipaewkqqnvsaapskattveaacryir
hlqqngst        interface= A:2,3,6,10,13,14, B:3,6,7,10,13,14,
A |   0   0  57   0   0  13   0
C |  96  96  13  96   0  16   0
G |   0   0  13   0  96  13  96
T |   0   0  13   0   0  54   0
```

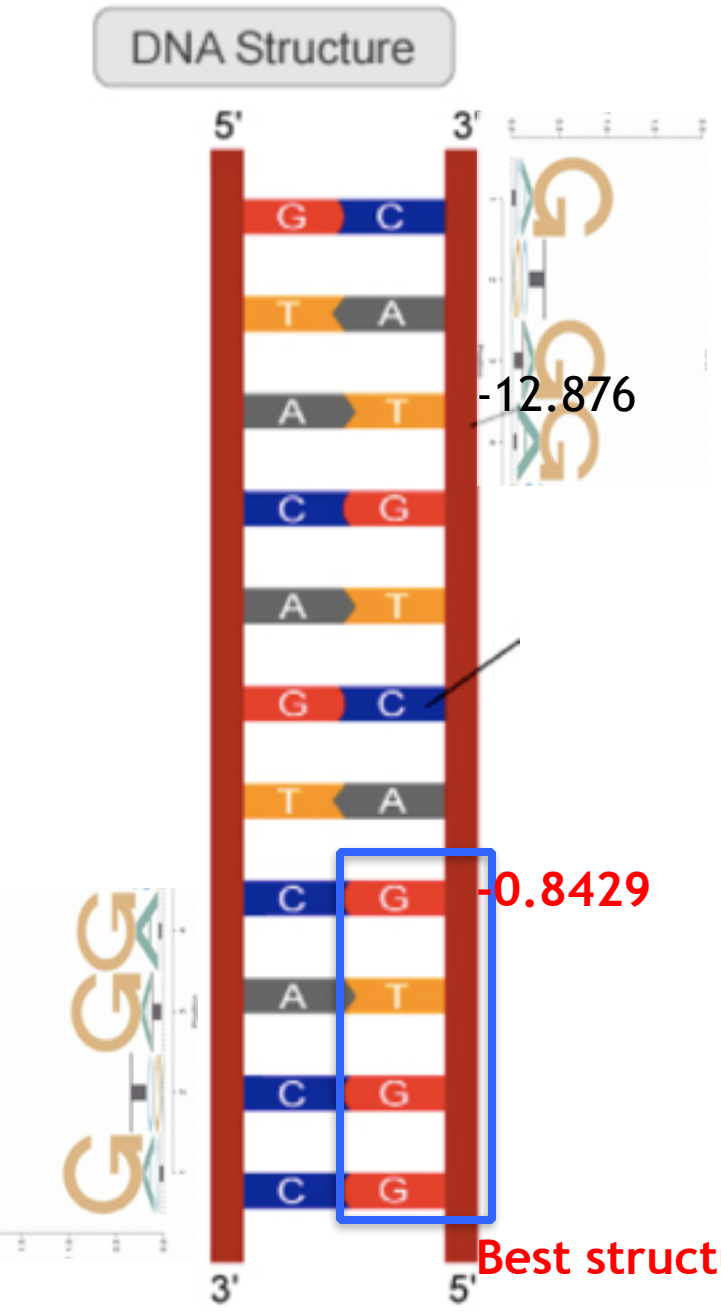
PDB name & Protein ID

PWM

(size: 1200 * 10 ≈ 12,000)



Fit the PWM along the DNA strand



Base\position	1	2	3	4
A	0.11	0.07	0.10	0.26
C	0.04	0.20	0.02	0.01
G	0.81	0.20	0.78	0.70
T	0.04	0.53	0.10	0.03

$$Current\ Score = \sum_{i=1}^{Length(PWM)} Log(P_i)$$

$$S_3 = \sum_{i=1}^{4} Log(P_i)$$

$$= Log(0.04) + Log(0.2) + Log(0.1) + Log(0.01)$$

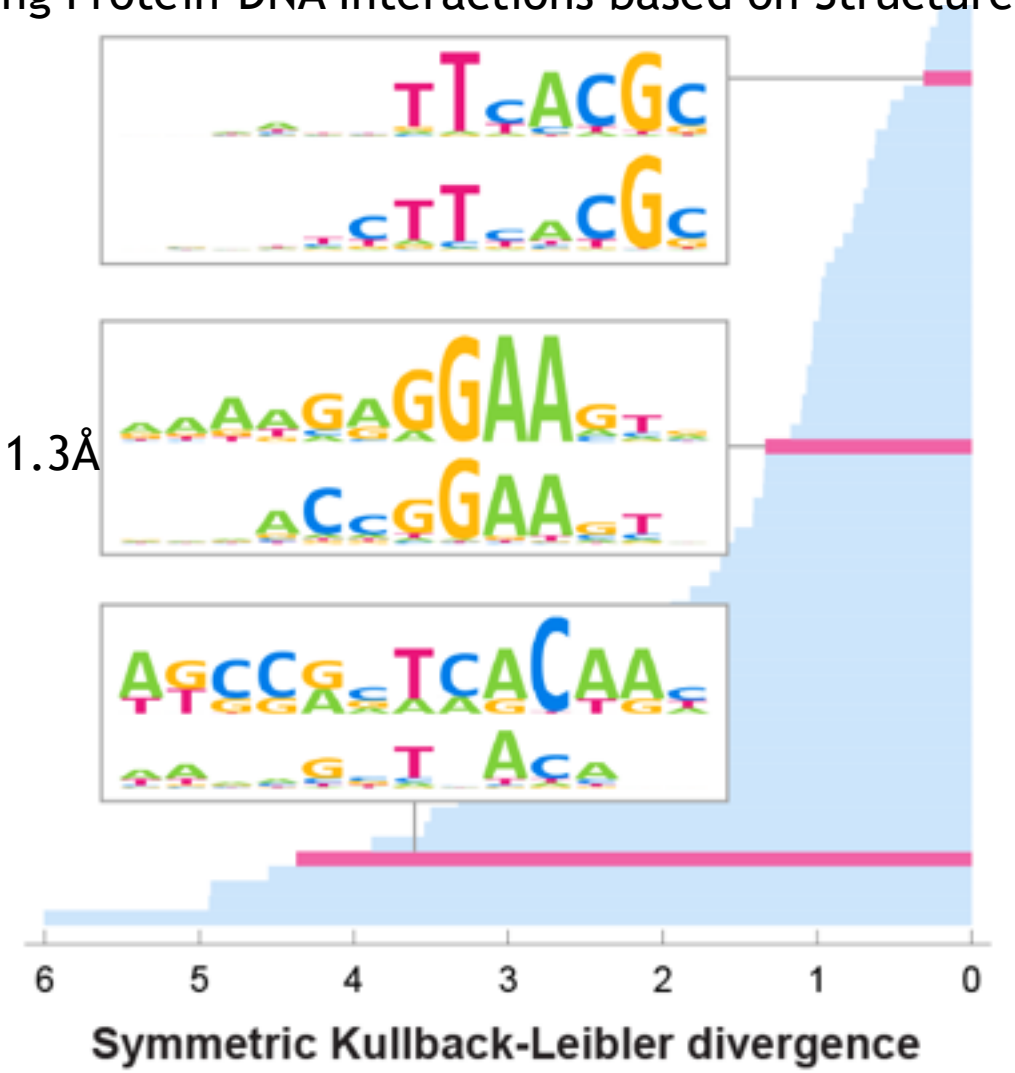
$$= -12.8755$$

New Result

Predicting Protein-DNA Interactions based on Structures

Lambda	(
Nbin=3	*1*
Nbin=4	*1*

Binwidth = 1.3Å



Old data
2.4352
2.5435

Phase 3. Train Model on Sequence

	PWM	Protein Sequence	Protein Structure	DNA Structure
Phase 1	✓	✓	✓	✓
Phase 2	✓	✓	✓	✓
Phase 3	✓	✓	✗	✗

There are many databases available such as CIS-BP, UniProbe, and JASPAR.

CIS-BP

Protein Sequence

```
('M0010_1.02', 'T000596_1.02', 'MS07_1.02', 'pTH7225', 'PBM', 'VSGETEPSASAT#TWGHKREREEFSLPPQPLITGSAVTKECESSMSLERPKKYRGVR
QRP#GK#AAEIRDPHKATRV#LGTFETAEAAARAYDAAALRFRGSKAKLNFPENVGTQTIQRNSHFLQNSWQPSLTYIDQCPTLLSYSRCMEQQQ', 'NDGCYGNH', 'DNCRGCVN'),
Pos  A    C    G    T
1    0.184462664307868  0.256581686929458  0.319720469565776  0.239235179196898
2    0.105898899930707  0.380880190302537  0.185394291902765  0.327826617863991
3    0.0455569631043099 0.000793710479183719  0.952379394258818  0.00126993215768892
4    0.0370386028966732 0.782082856690142  0.0541609782265034  0.126717562186682
5    0.0301800872826301 0.62464428573705  0.110739693670301  0.234435933310018
6    0.136135030144868  0.0930974935640104 0.701799156755609  0.068968319535512
7    0.17668837744013  0.191404234286779  0.36299447050365  0.26891291776944
8    0.199359808183106  0.485466980109179  0.0970388363926729  0.218134375315043

('M0011_1.02', 'T000600_1.02', 'MS29_1.02', 'pTH8109', 'PBM', 'MRRGRGSSAVAGPTVVAAINGSVKEIRFRGVRKRP#GRFAAEIRDP#KKARV#LGTFDS
AEEAARAYDSAARNLRGPKAKTNFPIDSSSPPPPNLRFNQIRNQNQNQVDPFMDHRLFTDHQQQFPIV', 'CGCCGSCR', 'YGSCGGCG'),
Pos  A    C    G    T
1    0.0801358912869704 0.759592326139089  0.0801358912869704  0.0801358912869704
2    0.0144203312392918 0.00014277555824672  0.985294117647059  0.00014277555824672
3    0.00014277555824672  0.885351227869789  0.114363221016562  0.00014277555824672
4    0.00014277555824672  0.985294117647059  0.0144203312392918  0.00014277555824672
5    0.0144203312392918 0.00014277555824672  0.985294117647059  0.00014277555824672
6    0.0572529982866933 0.642632781267847  0.185750999428898  0.114363221016562
7    0.0794733502538071 0.904346446700508  0.000158629441624365  0.0160215736040609
8    0.581087360694796  0.0234665427509294 0.279042750929368  0.116403345724907
```

PWM

(size: 5000 * 10 ≈ 50,000)

HHalign

No	Hit	Prob	E-value	P-value	Score	SS	Cols	Query HMM	Template HMM
1	gi|114796395|emb|CAK25951#9 pu	100.0	6.5E-64	6.5E-64	322.3	9.8	91	1-91	1-93 (93)

```
No 1
>gi|114796395|emb|CAK25951#9 putative ATP-dependent DNA ligase [Bacteriophage\nLKD16]
Probab=100.00  E-value=6.5e-64  Score=322.27  Aligned_cols=91  Identities=60%  Similarity=0.972  Sum_probs=90.2

Q ss_pred              CceeeeeEEEEEEecccCCcccceEEEEEEecCCCEEeeeccCHHHHHHHHHHHHHccCC--ccccccccEEEEEehhhc
Q gi|33300828|re    1 PEITVDGRIVGYVMGKTGKNVGRVVGYRVELEDGSTVAATGLSEEHIQLLTCAYLNAHID--EAMPNYGRIVEVSAMERS   78 (91)
Q Consensus         1 p~~~v~g~ivg~vmgktg~nvg~vvgyrv~ledgs~v~atgls~~~i~~lt~~~~~~~~~--~~m~~~grivevsamers   78 (91)
                        ||++|+|++||||||||+|||||||||||||+|||||||++|+|||||+++++++++  |+|+|+||||||+|||||
T Consensus         1 pslavegivvgfvmgktganvgkvvgyrvdledgtivsatgltrdriemltteaellggadhpgmadlgrvvevtamers   80 (93)
T gi|114796395|e    1 PSLAVEGIVVGFVMGKTGANVGKVVGYRVDLEDGTIVSATGLTRDRIEMLTTEAELLGGADHPGMADLGRVVEVTAMERS   80 (93)
T ss_pred              CceeeeeEEEEEEecccCCcccceEEEEEeccCCcEEeccccHHHHHHHHHHHHHhcCCCCCChhhcccEEEEEeeecc
Confidence          899999999999999999999999999999999999999999999999999999999998  9999999999999999


Q ss_pred              cccccCCcccccC
Q gi|33300828|re   79 ANTLRHPSFSRFR   91 (91)
Q Consensus        79 antlrhp~fsrfr   91 (91)
                        |||||||||||||
T Consensus        81 antlrhpkfsrfr   93 (93)
T gi|114796395|e   81 ANTLRHPKFSRFR   93 (93)
T ss_pred              cccccCCcccccC
Confidence          9999999999998
```

No	Hit	Prob	E-value	P-value	Score	SS	Cols	Query HMM	Template HMM
1	d1qpca_ d.144.1.7 (A:) Lymphoc	99.7	4.5E-17	3.2E-21	154.3	10.2	99	203-320	56-157 (272)
2	d1jpaa_ d.144.1.7 (A:) ephb2 r	99.7	4.3E-17	3.1E-21	156.8	8.8	99	203-321	75-177 (299)
3	d1uwha_ d.144.1.7 (A:) B-Raf k	99.7	5.1E-17	3.7E-21	154.8	7.7	100	203-322	52-154 (276)
4	d1opja_ d.144.1.7 (A:) Abelson	99.7	6.2E-17	4.5E-21	154.8	8.3	100	203-321	61-164 (287)
5	d1mp8a_ d.144.1.7 (A:) Focal a	99.6	9.9E-17	7.2E-21	151.3	8.6	100	203-322	56-158 (273)
6	d1sm2a_ d.144.1.7 (A:) Tyrosin	99.6	1.2E-16	8.8E-21	150.3	8.8	99	203-321	48-150 (263)
7	d1u59a_ d.144.1.7 (A:) Tyrosin	99.6	2.4E-16	1.7E-20	150.9	9.5	99	203-321	57-158 (285)
8	d1xbba_ d.144.1.7 (A:) Tyrosin	99.6	2.2E-16	1.6E-20	150.2	8.6	97	203-320	56-155 (277)
9	d1vjya_ d.144.1.7 (A:) Type I	99.6	2.6E-16	1.9E-20	151.3	8.8	98	204-320	46-156 (303)
10	d1mqba_ d.144.1.7 (A:) epha2 r	99.6	4.4E-16	3.2E-20	148.0	8.7	193	203-422	57-272 (283)
...									
64	d1j7la_ d.144.1.6 (A:) Type II	97.3	0.00014	1E-08	65.0	6.3	33	292-324	184-216 (263)
65	d1nd4a_ d.144.1.6 (A:) Aminogl	96.7	0.0012	8.5E-08	58.5	6.6	31	292-322	176-206 (255)
66	d1nw1a_ d.144.1.8 (A:) Choline	96.6	0.0011	7.8E-08	63.9	5.8	37	203-239	92-128 (395)
67	d2pula1 d.144.1.6 (A:5-396) Me	95.6	0.0071	5.2E-07	58.3	6.4	32	290-322	222-253 (392)
68	d1a4pa_ a.39.1.2 (A:) Calcycli	91.7	0.12	8.9E-06	40.0	5.4	62	140-202	18-80 (92)
69	d1ksoa_ a.39.1.2 (A:) Calcycli	91.2	0.17	1.2E-05	39.5	5.8	56	147-203	28-83 (93)
70	d1e8aa_ a.39.1.2 (A:) Calcycli	90.5	0.23	1.7E-05	38.3	6.0	56	147-203	27-82 (87)

>1bdmA structureX: 1bdm.pdb
MKAPVRVAVTGAAGQIGYSLLFRIAAGEMLGKDQPVILQLLEIPQAMKALEGVVMELEDCAFPLLAGLEATDDPDVAFKDADYALLVG
AAPRLQVNGKIFTEQGRALAEVAKKDVKVLVVGNPANTNALIAYKNAPGLNPRNFTAMTRLDHNRAKAQLAKKTGTGVDRIRRMTV
WGNHSSIMFPDLFHAEVDGRPALELVDMEWYEKVFIPTVAQRGAAIIQARGASSAASAANAAIEHIRDWALGTPEGDWVSMAVPSQ
GEYGIPEGIVYSFPVTAKDGAYRVVEGLEINEFARKRMEITELLDEMEQVKAL--GLI*

>TvLDH sequence:TvLDH
MSEAAHVLITGAAGQIGYILSHWIASGELYGDRQVYLHLLDIPPAMNRLTALTMELEDCAFPHLAGFVATTDPKAAFKDIDCAFLVASM
PLKPGQVRADLISSNSVIFKNTGEYLSKWAKPSVKVLVIGNPDNTNCEIAMLHAKNLKPENFSSLSMLDQNRAYYEVASKLGVDVKDV
HDIIVWGNHGESMVADLTQATFTKEGKTQKVVDVLDHDYVFDTFFKKIGHRAWDILEHRGFTSAASPTKAAIQHMKAWLFGTAPGE
VLSMGIPVPEGNPYGIKPGVVFSFPCNVDKEGKIHVVEGFKVNDWLREKLDFAQGG*

Modeller

Green: Template
Red: New structure

Method 1 (Naive)

1. Among all the output structures of HHalign, select all templates that have the probability to be a true positive higher than 0.9. Take them as inputs of Modeller.

2. Among all the output structures of Modeller, select the one with highest model score to be the “simulated” structure of input protein sequence.

Method 2 (similar to EM Algorithm)

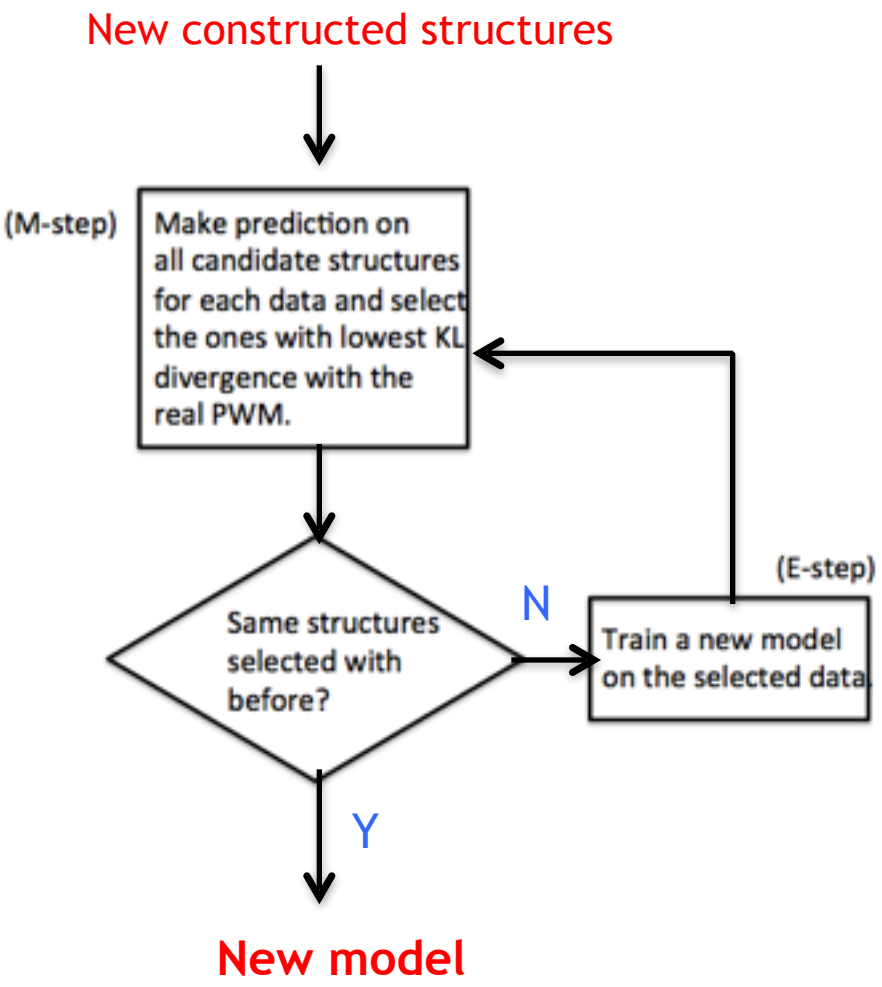
1. Among all the output structures of HHalign,

select all templates that have the probability to be a true positive higher than 0.9. Take them as inputs of Modeller.

2. Among all the output structures of Modeller, select the one has minimal KL divergence result with true PWM on the proposed model.

3. Run the model on selected data and adjust the optimal parameters.

4. Repeat 2~3 until the model is always selecting the same structures. (***Converge!***)



Application

- Make prediction on interaction when mutation caused by diseases hap
    - Mutation on proteins (trans)
    - Mutation on DNA (cis)

Acknowledgements

- Dr. Mohammed AlQuraishi
- Prof. Peter Sorger
- Saroja Somasundaram
- Samuel Cho
- All LSP/HiTS Lab members

Thank you!