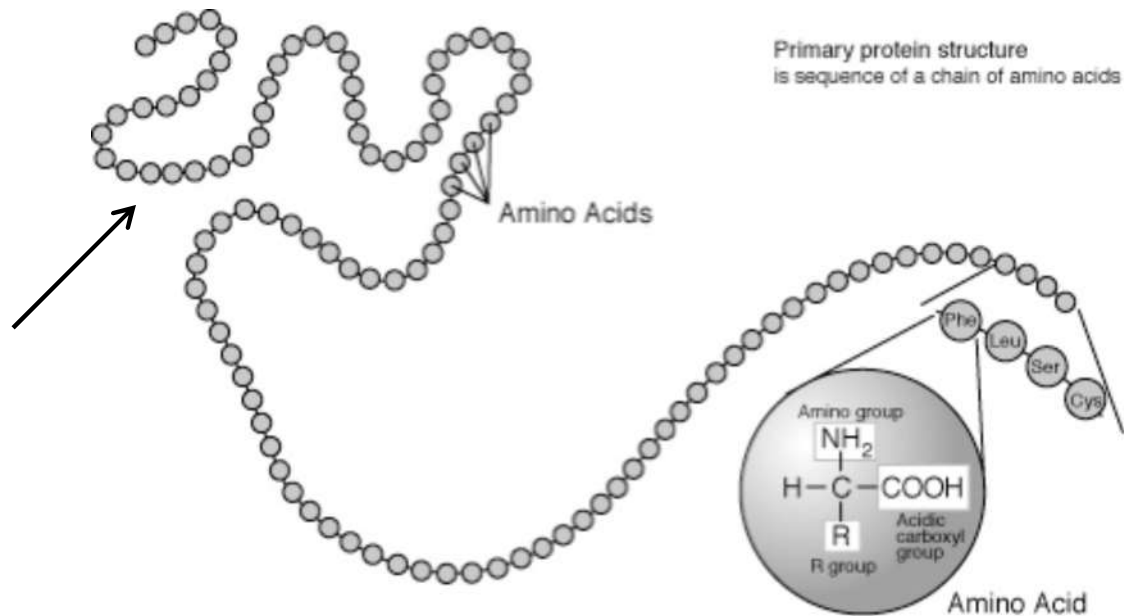


Case Studies of Machine-Learning in Structure Biology

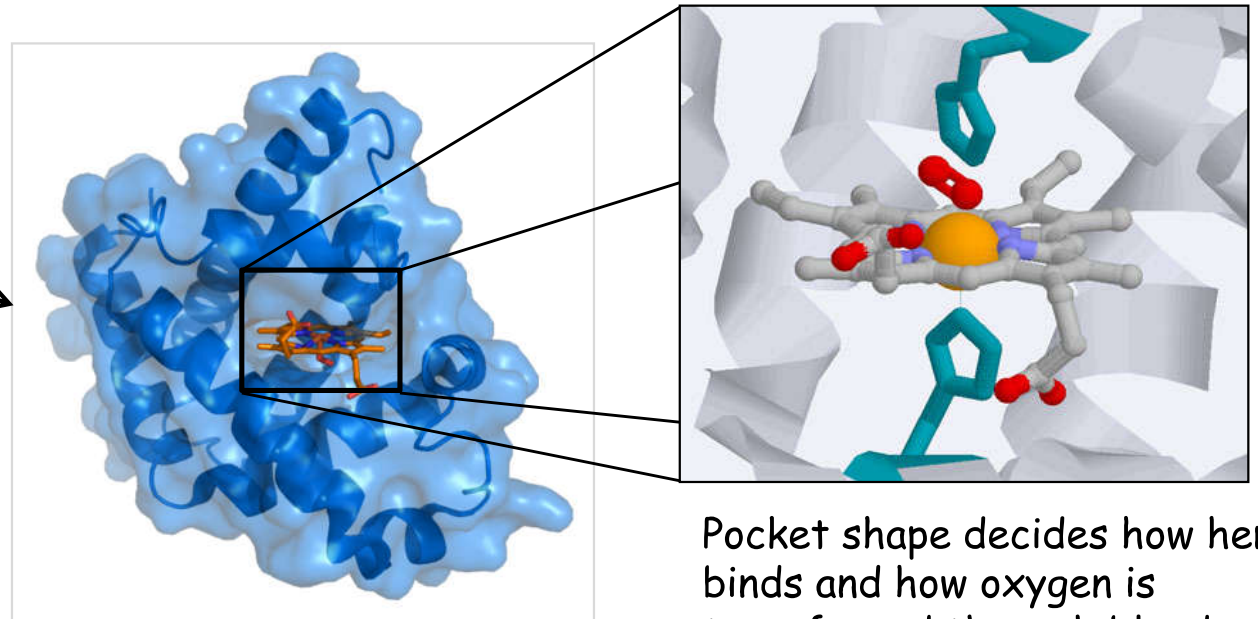
1. Protein Secondary Structure Prediction
2. Protein Contact Prediction
3. Disease-Associated Mutation Prediction

What is protein?

Protein is a 1D chain of amino acids. ~1M different proteins regulate life process in human body



Protein functions only when it folds into a unique shape



3D structure of myoglobin

1.1. What is protein secondary structure?

1, Primary amino acid sequences (1D)

MVLSEGEWQLVLHVWAKVEADVAGHGGQDILIRLFKSHPETLEKFDRV
KHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGHHAELEKPLAQSHA
TKHKIPIKYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKDI
AAKYKELGYQG

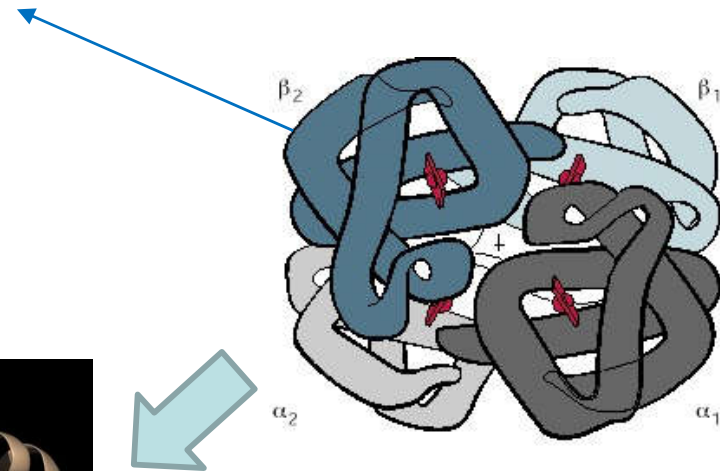
2, Secondary structure



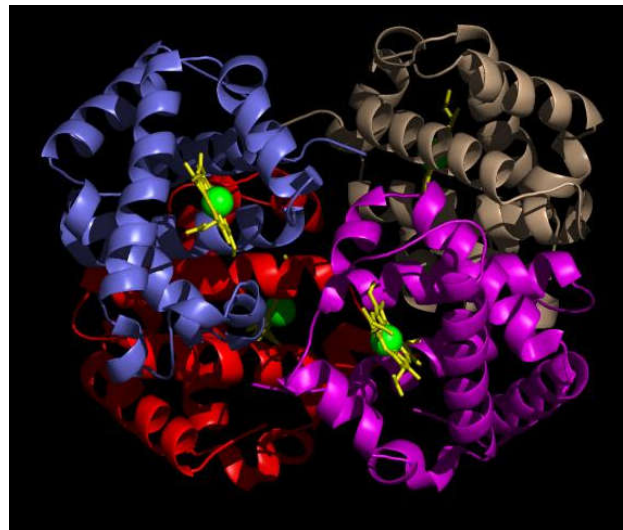
HHHHHHHHHHHLLLLLEEEEEEEEEEELLLLLLEEEEEEEEEEELLLLLLEEEEEEEEEEELLLLHHHHHLLLLHHHHHHHHH

1.1. What is protein secondary structure?

3, Tertiary structure



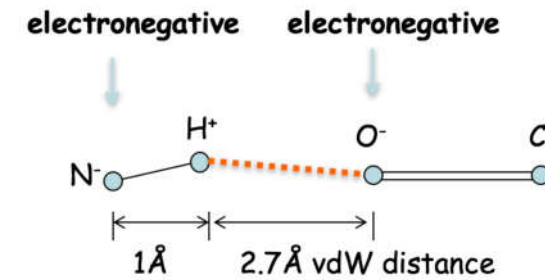
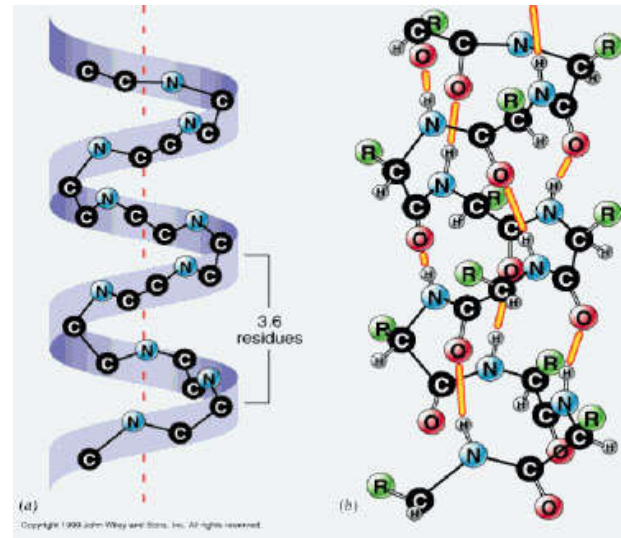
4, Quaternary structure



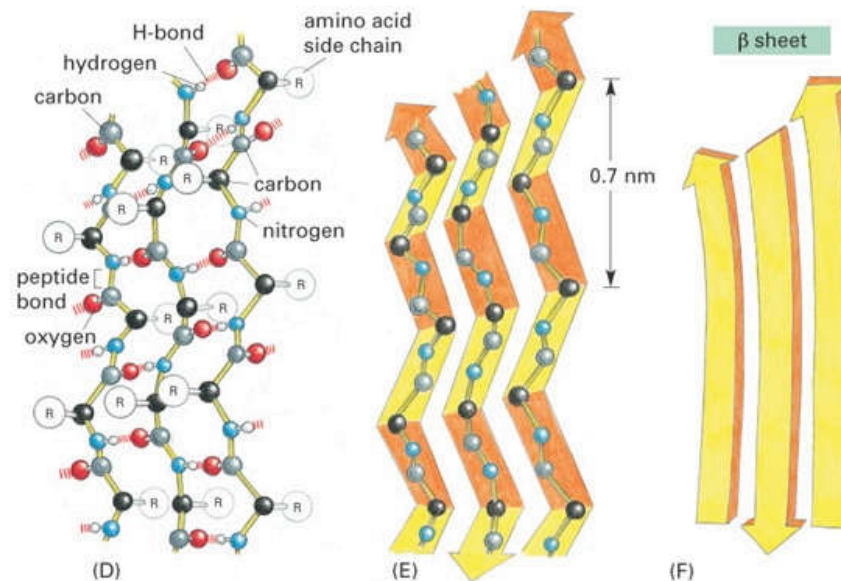
1.2. Hydrogen-bond

(Secondary structure is specified by H-bonding)

H-bond in α -helix



H-bond in β -sheet



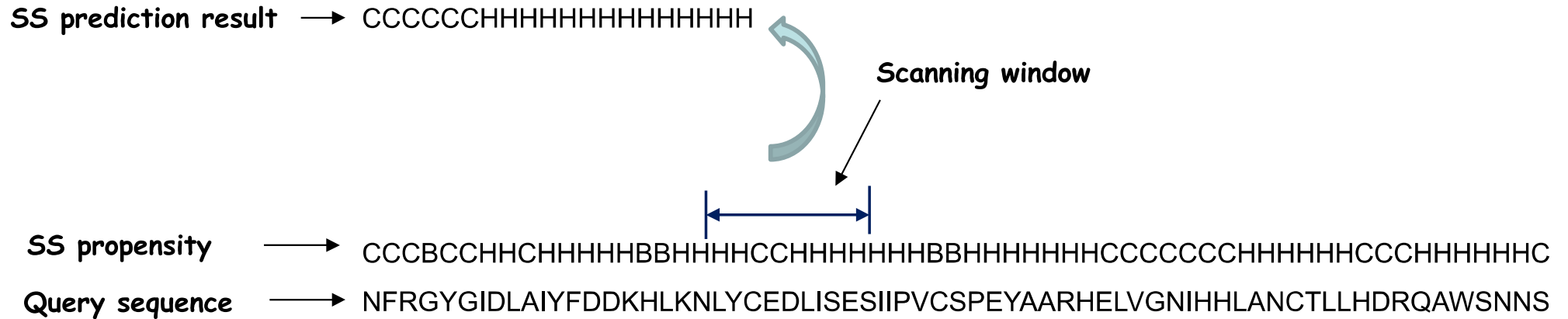
1.3. How to predict second structure from sequence? (Former effort: Chou-Fasman method)

SS propensity of amino acids:

Helical Residues ^b	P_{α}	β -Sheet Residues ^c	P_{β}
Glu ⁽⁻⁾	1.53	Met	1.67
Ala	1.45	Val	1.65
Leu	1.34	Ile	1.60
His ⁽⁺⁾	1.24	Cys	1.30
Met	1.20	Tyr	1.29
Gln	1.17	Phe	1.28
Trp	1.14	Gln	1.23
Val	1.14	Leu	1.22
Phe	1.12	Thr	1.20
Lys ⁽⁺⁾	1.07	Trp	1.19
Ile	1.00	Ala	0.97
Asp ⁽⁻⁾	0.98	Arg ⁽⁺⁾	0.90
Thr	0.82	Gly	0.81
Ser	0.79	Asp ⁽⁻⁾	0.80
Arg ⁽⁺⁾	0.79	Lys ⁽⁺⁾	0.74
Cys	0.77	Ser	0.72
Asn	0.73	His ⁽⁺⁾	0.71
Tyr	0.61	Asn	0.65
Pro	0.59	Pro	0.62
Gly	0.53	Glu ⁽⁻⁾	0.26

Predicting SS based on simple statistics

Chou-Fasman method



Accuracy of SS prediction:

$$Q3 = \frac{\text{\#residues with correct predicted SS}}{\text{\#total residues}}$$

Average accuracy: 50-60%

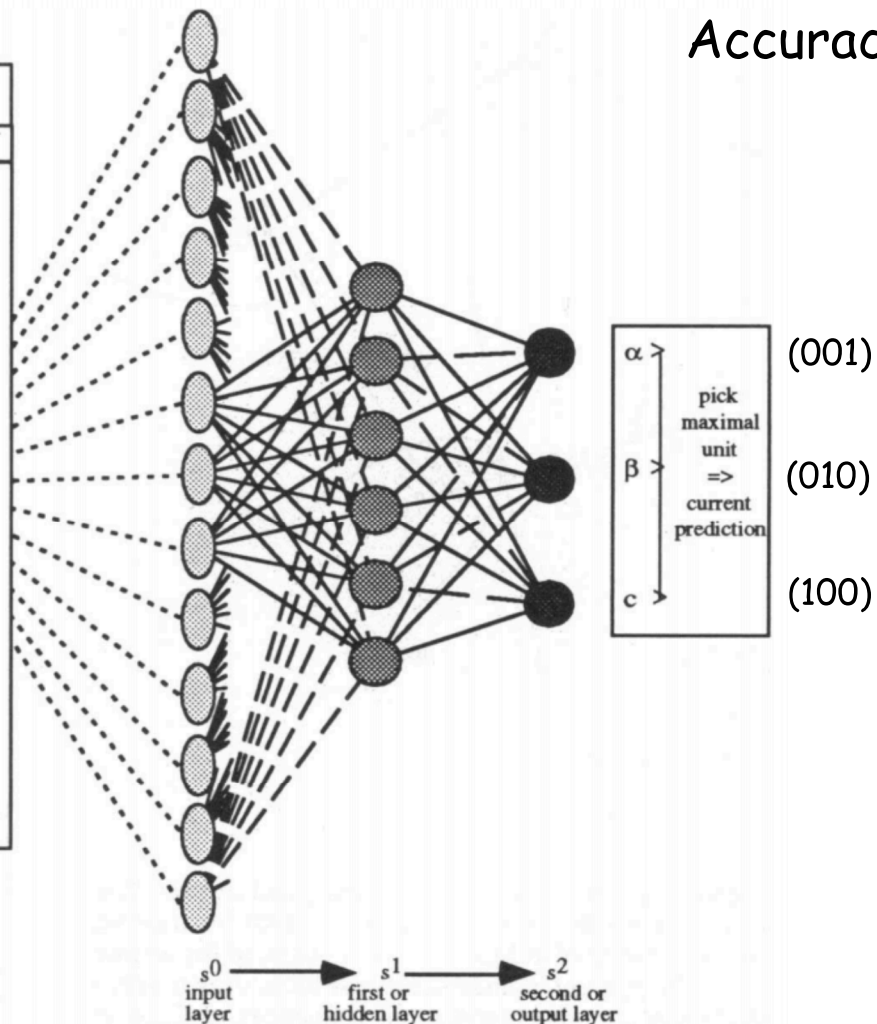
Better than random: 47%

(32% α -helix, 21% β -strand, 47% loop)

1.3. Former effort on SSP: PHD

Protein	Alignments	profile table
		GSAPD NT EKQ C VH IR LM YFW
:	:: :: ::	
G	GG GG	5.....
Y	YY YY5..
I	II EE2..3..
Y	YY YY5..
D	DD DD5.....
P	PP PP	...5.....
E	AE AA	..3...2.....
D	VV EE1..2...2...
G	GG GG	5.....
D	DD DD5.....
P	PP PP	...5.....
D	DT DD4..1.....
D	NQ NN1..3...1.....
G	GN GG	4....1.....
V	VI VV4..1.....
N	EP KK	...1..1.12.....
P	PP PP	...5.....
G	GG GG	5.....
T	TT TT5.....
D	EK S A	.11.1..11.....
F	FF FF5.....
:	:: :: ::	

Accuracy: 75%



Using sequence profile instead of single sequence as input of network training

Rost, B. & Sander, C., Prediction of protein secondary structure at better than 70 % Accuracy, Journal of Molecular Biology, (1993) 232, 584-599.

1.3. Former effort on SSP: PSIPRED

Accuracy: 80%

Raw profile from PSI-BLAST Log File

Position-based scoring matrix used

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
-3	-4	-4	-4	-3	-4	-4	-4	-2	-1	-1	-4	-1	8	-5	-3	-3	0	2	-2
0	-1	-1	3	-4	3	4	1	-1	-4	-4	0	-3	-4	-2	-1	-2	-4	-3	-3
0	-1	2	1	-3	4	0	-1	-2	-4	-3	1	-2	-4	-2	2	0	-4	-3	-3
-2	-3	-4	-5	-2	-3	-4	-6	-4	0	6	0	0	-1	-4	-3	-2	-4	-2	0
0	-3	-1	-2	-3	0	-2	4	-3	-3	0	-2	-2	-4	-3	3	1	-4	-4	-3
0	2	0	4	-4	1	2	1	-2	-4	-4	0	-3	-4	-3	1	-2	-5	-4	-4
-1	5	3	-2	-4	-1	-1	1	-2	-1	-4	1	-3	-4	-3	1	-2	-5	-4	-4
-2	-3	-4	-5	-3	-3	-4	-5	-4	3	4	-1	1	2	-4	-3	-2	-3	-1	0
-2	3	2	-2	-4	2	1	-3	-2	-3	-3	1	1	-4	-3	2	1	-4	-3	-1
0	2	3	1	-4	0	0	0	-2	-4	-4	1	-3	-4	-3	2	0	-5	-4	-4
5	-3	-3	-3	-2	-3	-3	-2	-3	1	-2	-3	-2	1	-3	0	1	-4	-2	0
-1	-4	-5	-5	-3	-4	-4	-5	-4	3	3	-4	2	3	-5	-3	-2	5	-1	2
0	3	3	0	-4	3	0	1	-2	-4	-4	1	-3	-4	-3	1	-1	-4	-3	-4
-1	0	1	0	-4	1	-1	-1	-2	-4	-3	5	-2	0	-3	0	-2	-4	0	-3
-2	-3	-1	-5	-3	-3	-4	-5	-4	3	4	0	4	2	-4	-3	-2	-3	-2	0
0	3	0	-2	-3	-1	0	0	-2	0	0	1	0	-1	-3	2	0	-4	-3	0
-1	1	3	-2	-4	0	-2	4	-2	-4	-4	0	-3	0	-3	0	0	-3	0	-4

Window of 15 rows

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
0.4	0.3	0.3	0.3	0.2	0.9	0.3	0.3	0.4	0.4	0.4	0.3	0.4	0.9	0.1	0.4	0.4	0.5	0.7	0.4
0.3	0.2	0.3	0.8	0.4	0.3	0.7	0.1	0.6	0.2	0.4	0.3	0.5	0.2	0.1	0.4	0.8	0.2	0.3	0.2
0.1	0.1	0.4	0.3	0.5	0.1	0.1	0.3	0.1	0.1	0.4	0.2	0.4	0.9	0.3	0.4	0.4	0.9	0.3	0.6
0.6	0.3	0.3	0.1	0.3	0.5	0.5	0.2	0.1	0.4	0.4	0.3	0.6	0.9	0.1	0.5	0.1	0.5	0.7	0.4
.

15 x 20 scaled inputs to 1st network

1st Network
315 inputs
75 hidden units
3 outputs

Window of 15 x 3
outputs fed to 2nd
network

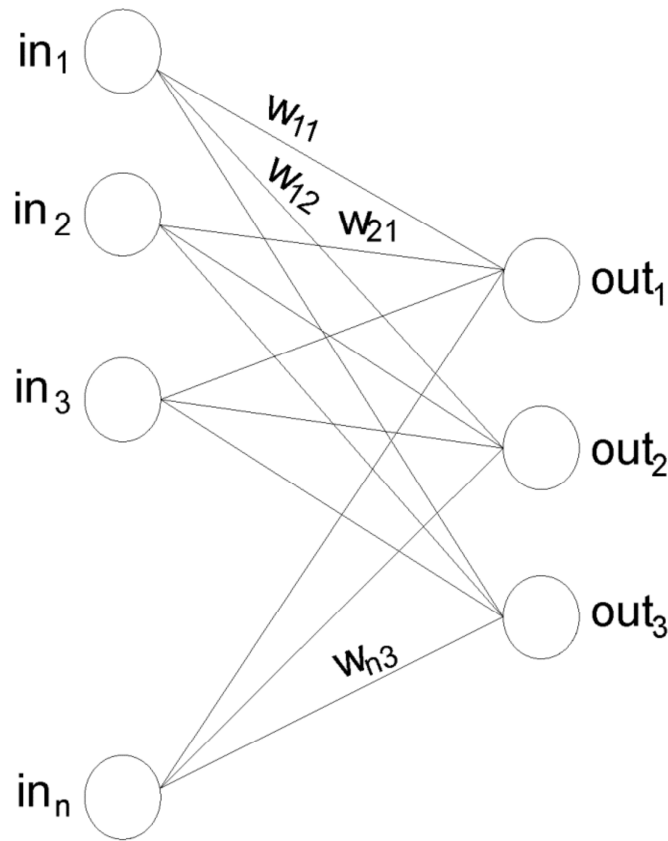
2nd Network
60 inputs
60 hidden units
3 outputs

Final 3-state
Prediction

Major difference between PHD and PSIPRED is the use of PSI-Blast for profile construction

Jones, D., Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. (1999) 292, 195-202.

What is neural network?

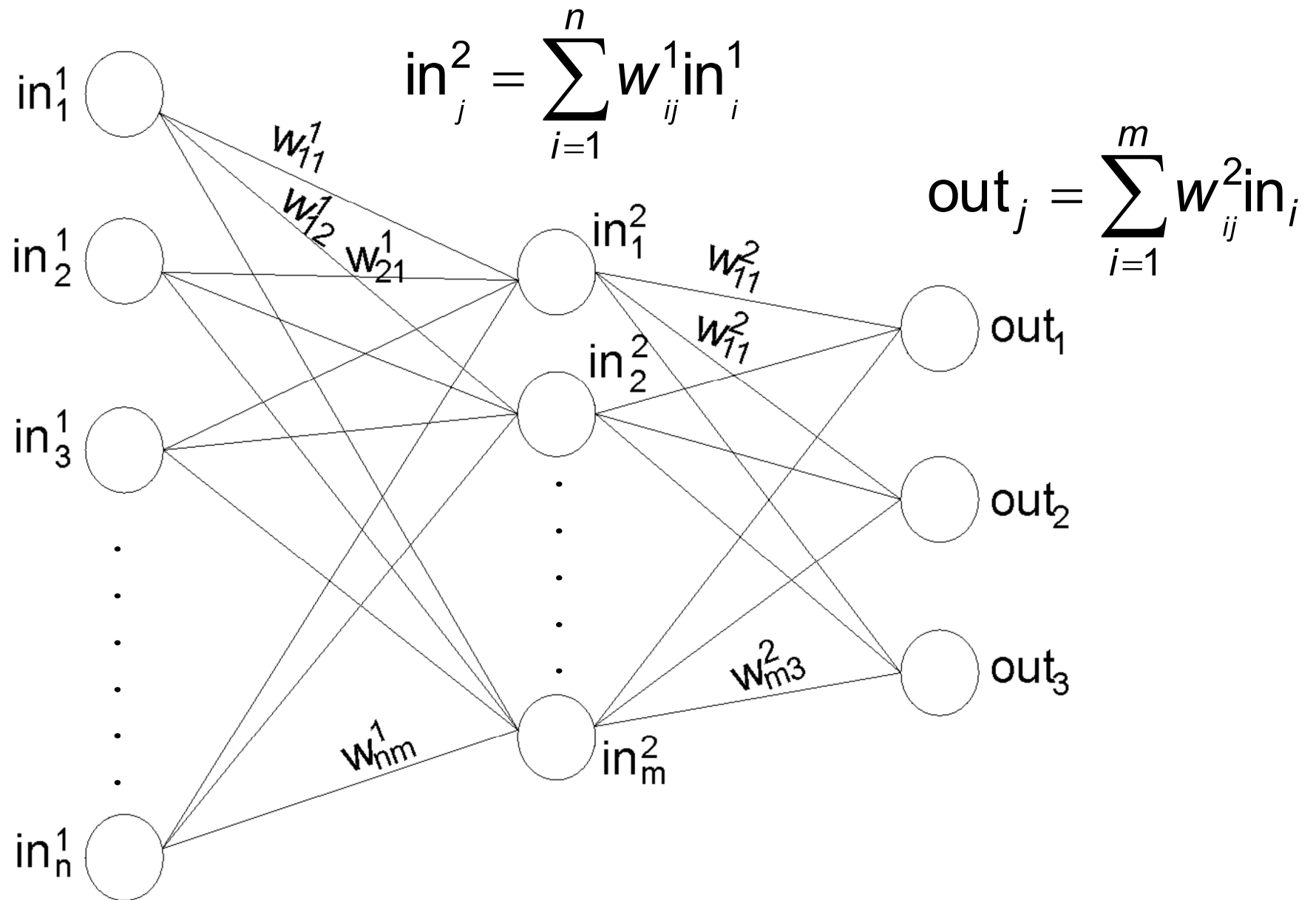


$$out_j = \sum_{i=1}^n w_{ij} in_i$$

The principle of neural network is to adjust the weights (w_{ij}) iteratively so that output (out_j) is close to the true answer (T).

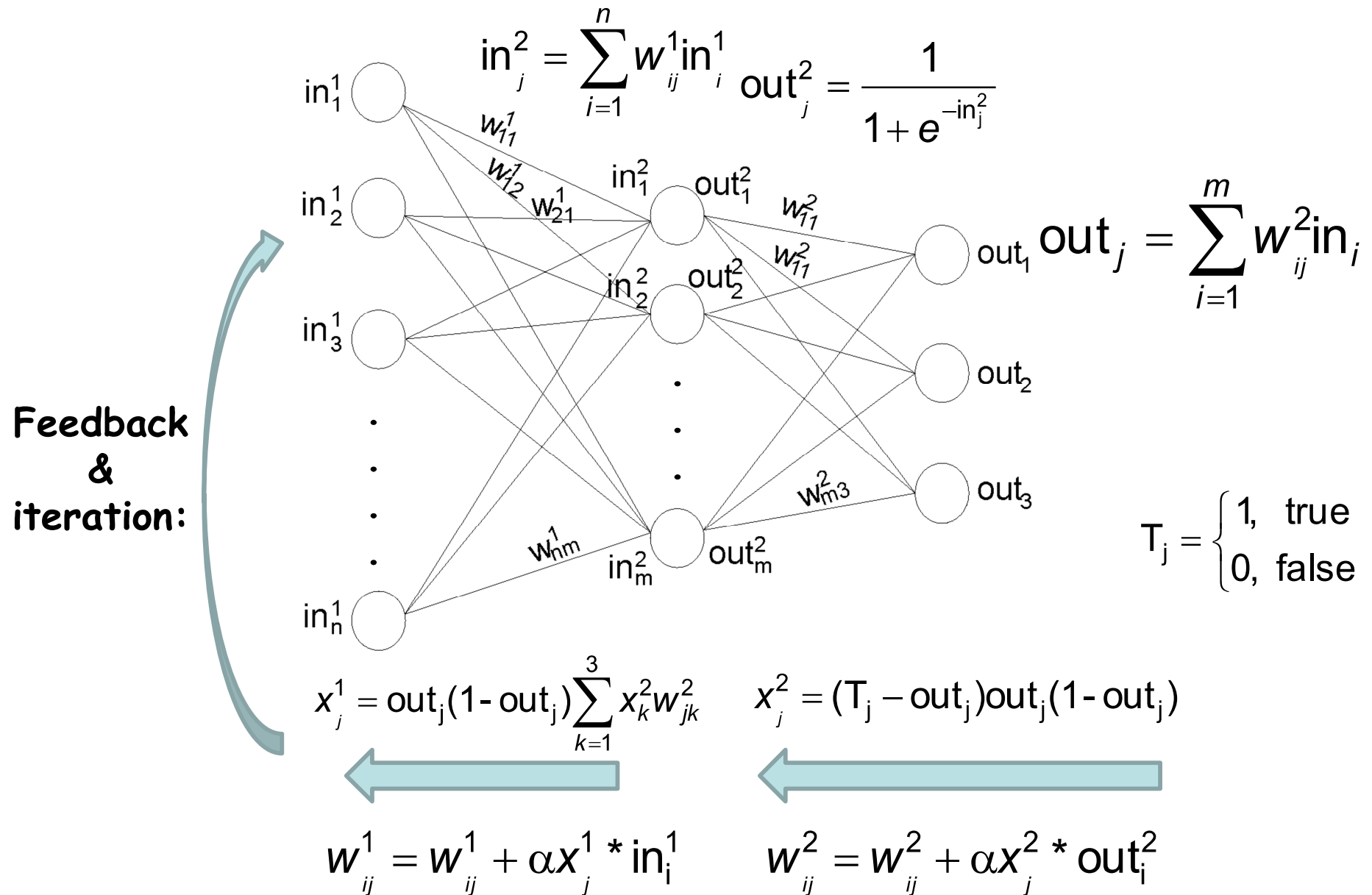
$$w_{ij}^{new} = w_{ij}^{old} + \alpha f(error)$$

A two layer network

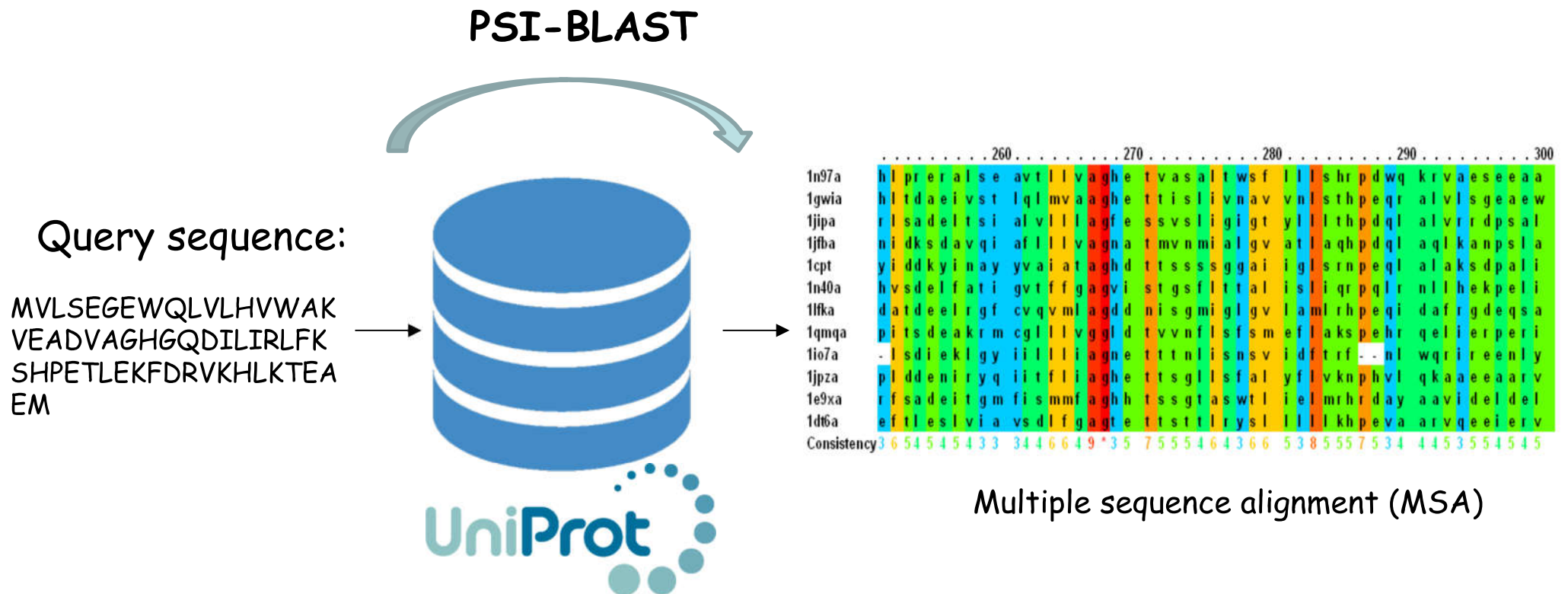


1.4. State of the art: PSSpred

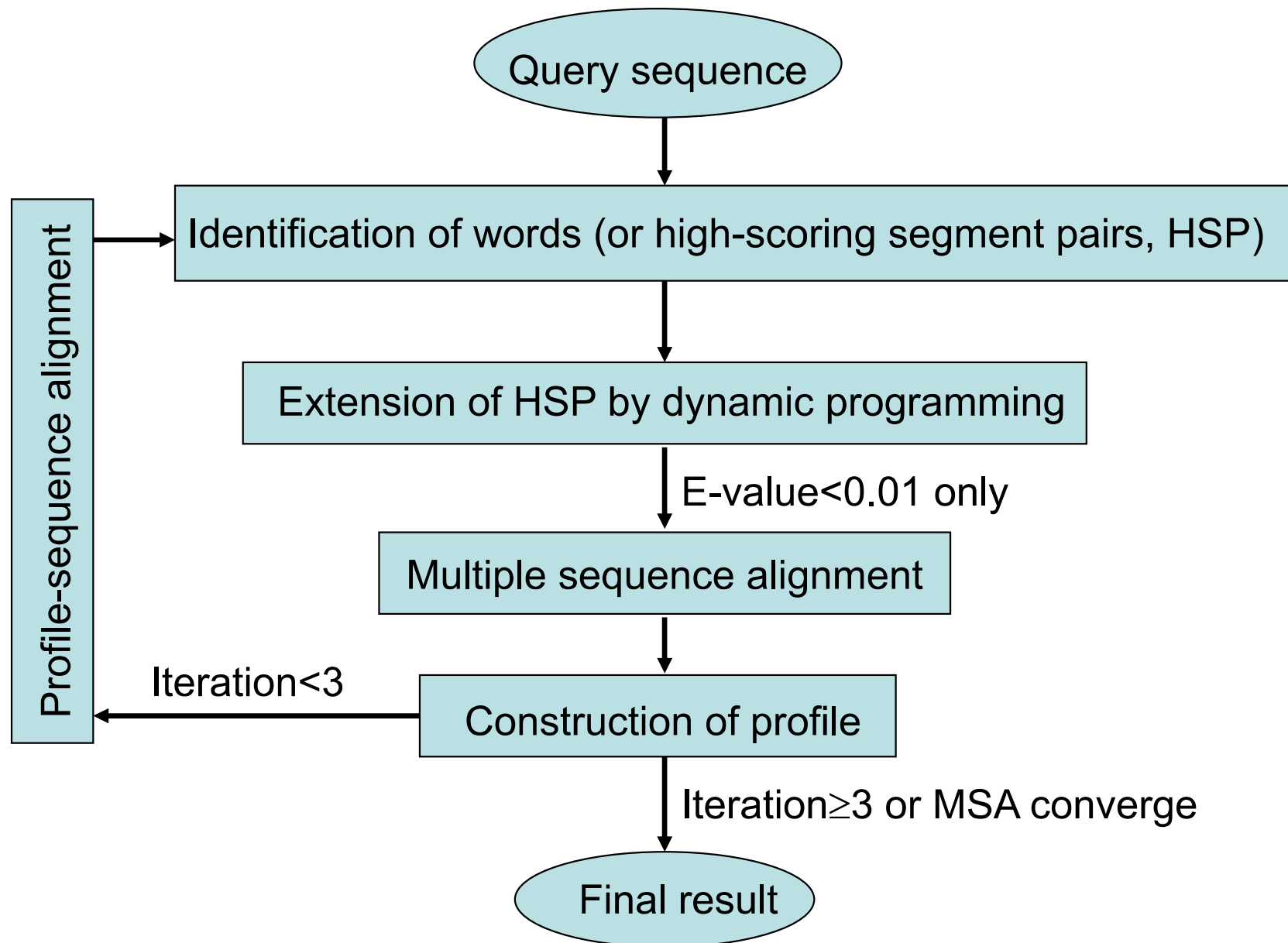
(<http://zhanglab.ccmb.med.umich.edu/PSSpred>)



1.4. PSSpred feature collection



PSI-Blast Pipeline: Iterative Profile-sequence Alignment Algorithm



PSSM: $in = 1/(1 + e^{-x})$ $\mathbf{x} = \log(\mathbf{Q}_{ij}/\mathbf{P}_i)$

MTX: $in = 1/(1 + e^{-x/100})$ $\mathbf{x} = \log(\mathbf{Q}_{ij}/\mathbf{P}_i)$

PROF_w: $in = 1/(1 + e^{-x})$ $\mathbf{x} = \sum_{a=1}^{20} \sum_{k=1}^{f(a,j)} w(k)B(A_i, a)$

FREQ_{cw}: $in = 1/[1 + e^{-25(x - \langle x \rangle)}]$ $\mathbf{x} = \sum_{k=1}^{f(A_i, j)} w(k)$

FREQ_{cwQ}: $in = 1/[1 + e^{-30(x - \langle x \rangle)}]$ $\mathbf{x} = \sum_{k=1}^{f(A_i, j)} w(k)$

1.4. PSSpred training parameters

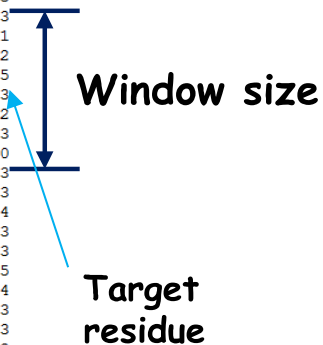
Seven PSSpred programs

1. mtx_pssm_freqccw_profw_12
2. mtx_freqccw_profw_freqccwG_15
3. mtx_freqccw_profw_freqccwG_12
4. mtx_freqccw_profw_12
5. mtx_freqccw_profw_18
6. mtx_profw_freqccwG_18
7. mtx_profw_12

```

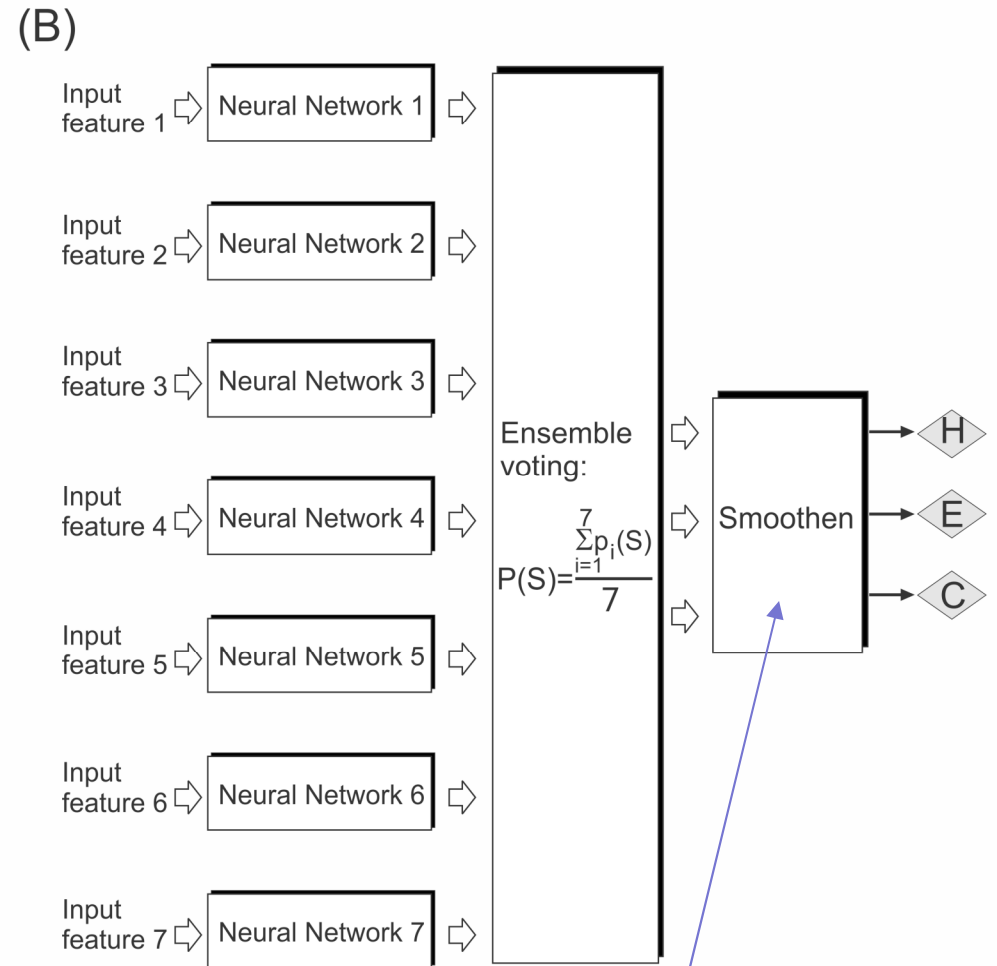
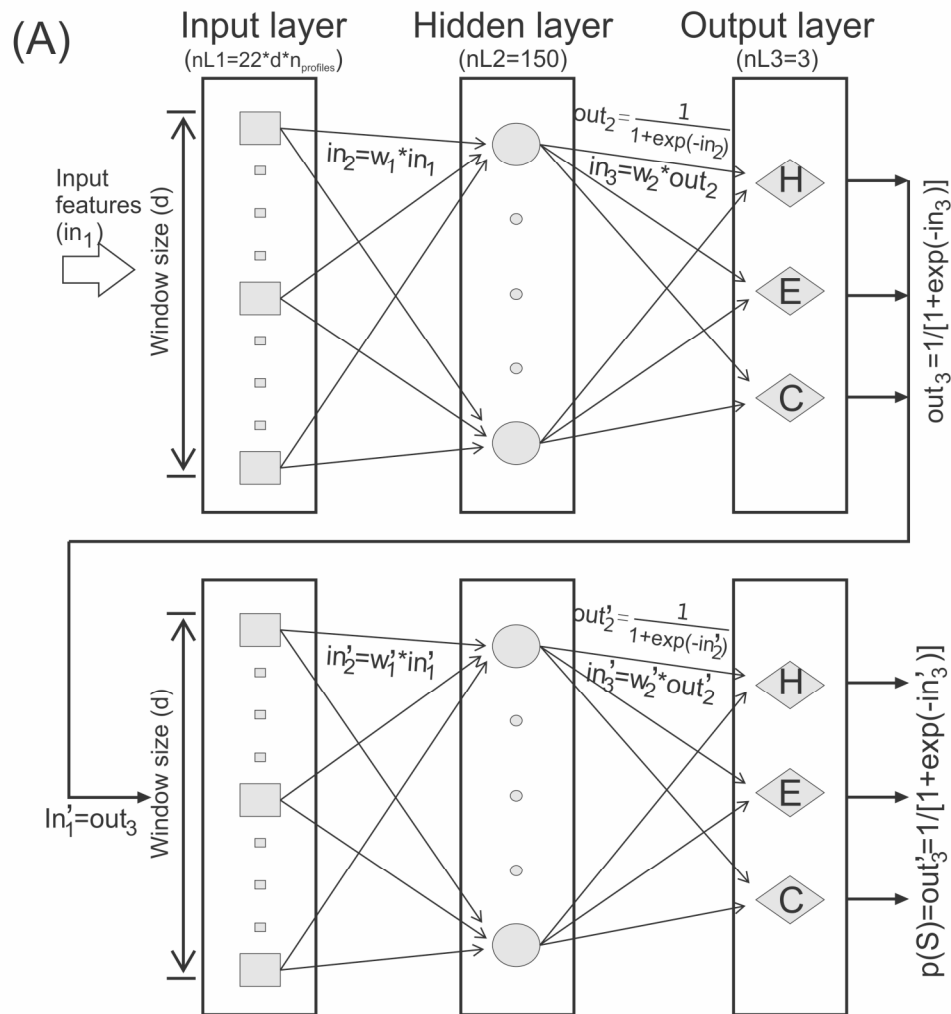
1 M -2 -2 -3 -4 -2 -1 -3 -3 -2 1 2 -2 8 0 -3 -2 -1 -2 -2 0
2 K -2 4 -1 -2 -4 1 0 -2 -1 -3 -3 5 -2 -4 -2 -1 -1 -4 -2 -3
3 I -2 -4 -4 -4 -2 -3 -4 -5 -4 6 1 -3 1 -1 -3 -3 -1 -3 -2 2
4 P -1 -3 -3 -2 -4 -2 -2 -3 -3 -4 -2 -3 -4 8 -1 -2 -4 -4 -3
5 K -1 4 -1 -1 -4 1 0 -2 -1 -3 -3 5 -2 -4 -2 -1 -1 -4 -2 -3
6 I -2 -3 -4 -4 -2 -3 -4 -4 -4 4 4 -3 1 0 -4 -3 -2 -3 -2 1
7 Y -2 -2 -3 -4 -3 -2 -3 -4 1 -2 -2 -2 -2 3 -4 -2 -2 2 8 -2
8 V -1 -3 -4 -4 -1 -3 -3 -4 -4 3 0 -3 0 -1 -3 -2 -1 -4 -2 5
9 E -1 -1 -1 1 -4 2 6 -3 -1 -4 -4 0 -3 -4 -2 -1 -1 -4 -3 -3
10 G 2 -2 2 -1 -2 -2 -2 5 -2 -3 -3 -2 -3 -3 -2 0 -1 -3 -3 -2
11 E -2 -1 2 1 -4 1 6 -2 0 -4 -4 0 -3 -4 -2 -1 -1 -4 -3 -3
12 L -2 0 -2 -2 -3 -1 1 -3 -2 2 3 3 0 -1 -3 -2 -1 -3 -2 0
13 N -2 -1 5 -1 -3 -1 -1 1 6 -3 -3 -1 -2 2 -3 -1 -2 -2 0 -3
14 D -2 -1 0 5 -4 1 5 -2 -1 -4 -4 0 -3 -4 -2 -1 -2 -4 -3 -3
15 G -1 -2 3 3 -4 -1 1 4 -1 -4 -4 -1 -3 -4 -2 -1 -2 -4 -3 -4
16 D -2 4 2 3 -3 0 3 -2 -1 -3 -3 1 -2 -3 -2 -1 -1 -3 -2 -3
17 R -1 4 -1 -2 -4 1 0 -2 -1 -3 -3 5 -2 -4 -2 -1 -1 -4 -2 -3
18 V -1 -3 -4 -4 -1 -3 -3 -4 -4 3 1 -3 0 -1 -3 -2 -1 -4 -2 5
19 A 3 -3 -3 -3 -1 -2 -2 -2 -3 1 -1 -2 0 -2 -2 -1 -1 -3 -2 4
20 I -2 -4 -4 -4 -2 -3 -4 -4 -4 5 1 -3 1 -1 -3 -3 -1 -3 -2 3
21 E -1 -1 -1 1 -4 1 6 -2 -1 -4 -3 0 -3 -4 -2 1 -1 -4 -3 -3
22 K -2 0 5 0 -4 1 3 -2 0 -4 -4 3 -2 -4 -2 0 -1 -4 -3 -3
23 D -1 -1 2 5 -3 0 3 1 -1 -3 -3 -1 -3 -3 -2 0 -1 -3 -3 -3
24 G 0 -2 -1 -1 -2 -2 -2 6 -2 -3 -3 -2 -3 -3 -2 -1 -2 -2 -3 -3
25 N -1 1 3 -1 -4 1 0 -2 -1 -3 -3 5 -2 -4 -2 0 -1 -4 -3 -3
26 A 2 -2 -2 -2 -2 -1 1 -2 -2 2 -1 1 -1 -2 -2 -1 -1 -3 -2 3
27 I -2 -4 -4 -4 -2 -3 -4 -5 -4 6 1 -3 1 -1 -3 -3 -1 -3 -2 2
28 I -2 3 -2 -3 -3 -1 -2 -4 -2 4 0 2 0 -2 -3 -2 -1 -3 -2 1
29 F -3 -3 -4 -4 -3 -4 -4 -4 -2 -1 0 -4 0 7 -4 -3 -3 0 3 -1
30 L -2 -3 -4 -4 -2 -3 -4 -4 -3 1 5 -3 2 0 -4 -3 -2 -2 -2 0
31 E -2 -1 0 6 -4 0 4 -2 -1 -4 -4 0 -3 -4 -2 -1 -2 -4 -3 -4
32 K -1 1 -1 0 -3 1 3 -2 -1 -3 -3 5 -2 -4 -2 1 -1 -4 -2 -3

```



Predictors	Training features	Window size	Number of iterations
PSSpred1	PSSM MTX PROF_W FREQ_CW	12	44
PSSpred2	MTX PROF_W FREQ_CW FREQ_CWG	15	38
PSSpred3	MTX PROF_W FREQ_CW FREQ_CWG	12	62
PSSpred4	MTX PROF_W FREQ_CW	12	63
PSSpred5	MTX PROF_W FREQ_CW	18	54
PSSpred6	MTX PROF_W FREQ_CWG	18	47
PSSpred7	MTX PROF_W	12	84

1.4. Pipeline of PSSpred



Island removal:

---x---
---xx---


Number of training proteins:

- 5,527 non-redundant proteins

1.4. State of the art: PSSpred

Output of PSSpred

Winner takes all, with $\text{conf} = 10 * [P(S_1) - P(S_2)]$



1	N	C	0.015	0.019	0.972
2	F	C	0.010	0.363	0.704
3	V	E	0.012	0.705	0.354
4	R	E	0.011	0.814	0.250
5	F	E	0.006	1.014	0.063
6	V	E	0.003	1.034	0.043
7	I	E	0.007	0.995	0.076
8	E	C	0.013	0.261	0.738
9	G	C	0.009	0.167	0.825
10	R	E	0.006	0.964	0.113
11	R	E	0.006	1.020	0.050
12	V	E	0.010	0.998	0.066
13	G	E	0.007	0.965	0.103
14	W	E	0.007	0.851	0.193
15	V	E	0.015	0.754	0.359

Possibility of
alpha



...

...

Possibility of
beta

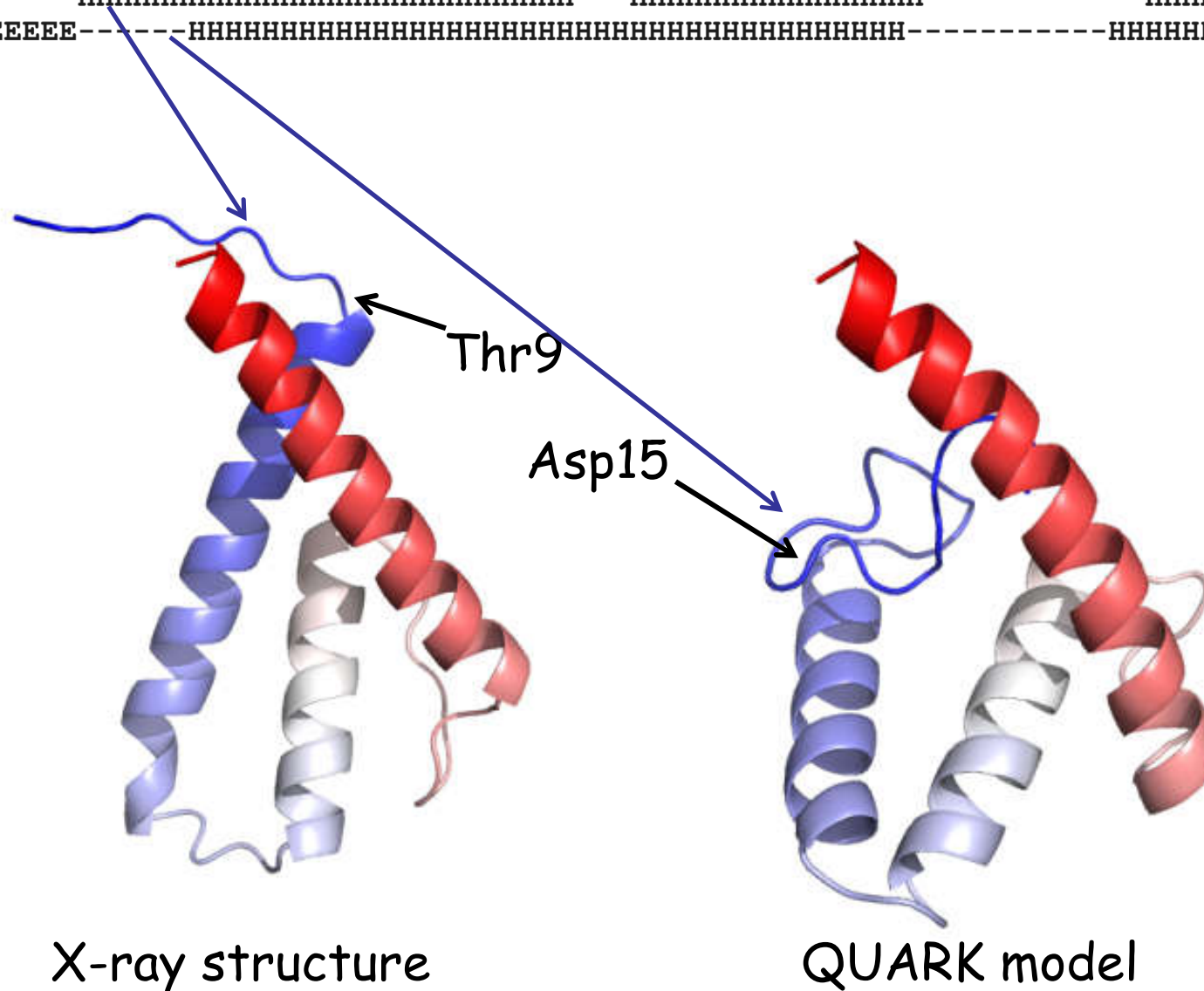


Possibility of
coil



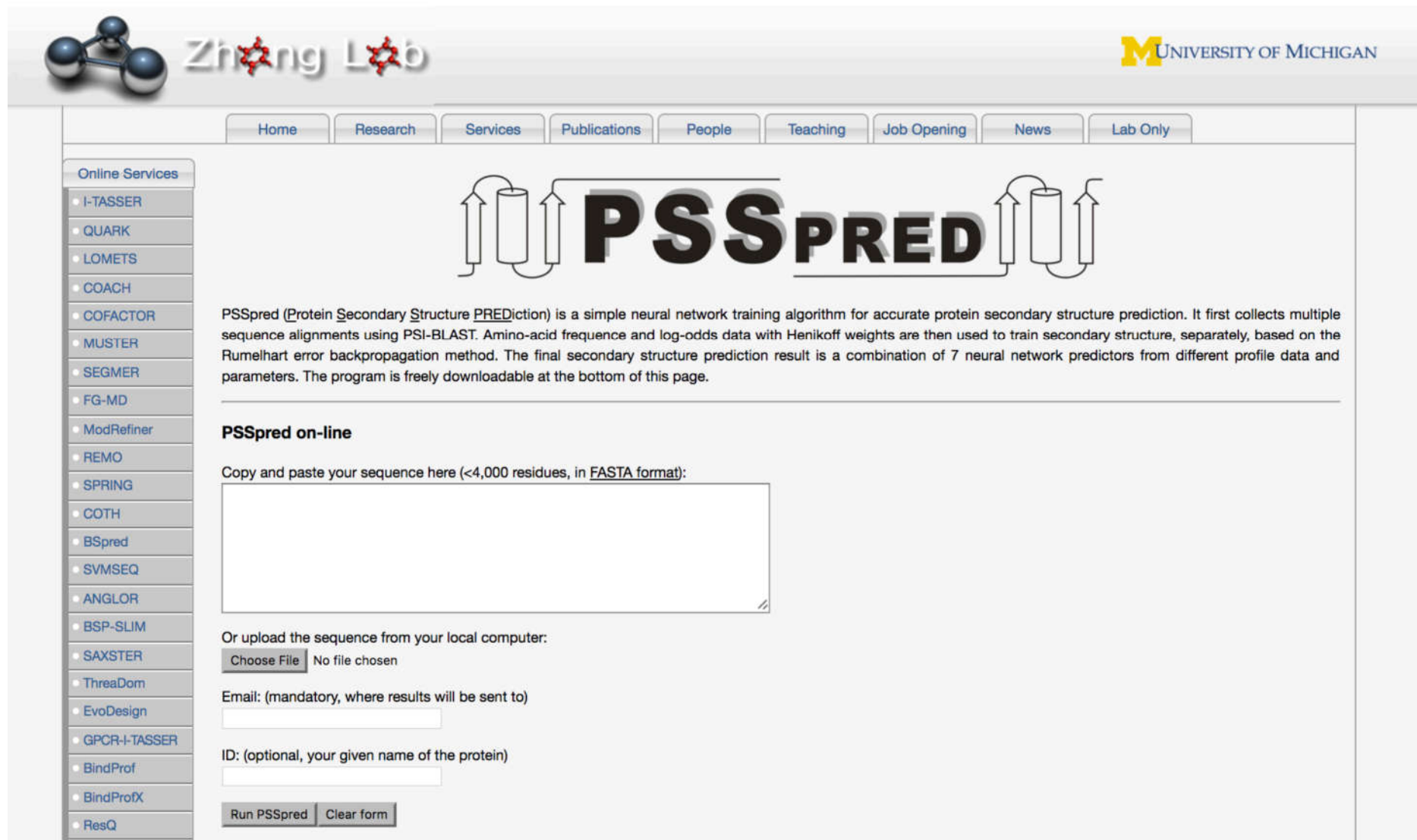
Result: Average accuracy =84% on 600 non-redundant proteins, which represents the start of the art of secondary structure prediction. This is close to the experimental uncertainty of hydrogen-bond and SS definition: ~90%.

Impact of SS prediction on 3D structure prediction (Target T0820-D1 in CASP11)

[illegible]

TM=0.449
RMSD=8.5Å

The on-line server and standalone program of PSSpred is available at <http://zhanglab.ccmb.med.umich.edu/PSSpred/>



The image shows the PSSpred web interface. At the top, there is a header with the 'Zhang Lab' logo on the left and the 'UNIVERSITY OF MICHIGAN' logo on the right. Below the header is a navigation bar with buttons for 'Home', 'Research', 'Services', 'Publications', 'People', 'Teaching', 'Job Opening', 'News', and 'Lab Only'. On the left side, there is a vertical menu titled 'Online Services' with a list of tools: I-TASSER, QUARK, LOMETS, COACH, COFACTOR, MUSTER, SEGMENT, FG-MD, ModRefiner, REMO, SPRING, COTH, BSpred, SVMSEQ, ANGLOR, BSP-SLIM, SAXSTER, ThreaDom, EvoDesign, GPCR-I-TASSER, BindProf, BindProfX, and ResQ. The main content area features the 'PSSPRED' logo, which consists of the word 'PSSPRED' in a bold, sans-serif font, flanked by two diagrams of protein secondary structure elements (alpha-helices and beta-strands). Below the logo, a paragraph describes the algorithm: 'PSSpred (Protein Secondary Structure PREDiction) is a simple neural network training algorithm for accurate protein secondary structure prediction. It first collects multiple sequence alignments using PSI-BLAST. Amino-acid frequency and log-odds data with Henikoff weights are then used to train secondary structure, separately, based on the Rumelhart error backpropagation method. The final secondary structure prediction result is a combination of 7 neural network predictors from different profile data and parameters. The program is freely downloadable at the bottom of this page.' Below this text, there is a section titled 'PSSpred on-line'. It contains a text input field for the sequence, with the instruction 'Copy and paste your sequence here (<4,000 residues, in FASTA format):'. Below the input field, there is a link 'Or upload the sequence from your local computer:' followed by a 'Choose File' button and the text 'No file chosen'. Below this, there is an 'Email:' field with the instruction '(mandatory, where results will be sent to)'. Below the email field, there is an 'ID:' field with the instruction '(optional, your given name of the protein)'. At the bottom of the form, there are two buttons: 'Run PSSpred' and 'Clear form'.

Online Services

- I-TASSER
- QUARK
- LOMETS
- COACH
- COFACTOR
- MUSTER
- SEGMENT
- FG-MD
- ModRefiner
- REMO
- SPRING
- COTH
- BSpred
- SVMSEQ
- ANGLOR
- BSP-SLIM
- SAXSTER
- ThreaDom
- EvoDesign
- GPCR-I-TASSER
- BindProf
- BindProfX
- ResQ

PSSPRED

PSSpred (Protein Secondary Structure PREDiction) is a simple neural network training algorithm for accurate protein secondary structure prediction. It first collects multiple sequence alignments using PSI-BLAST. Amino-acid frequency and log-odds data with Henikoff weights are then used to train secondary structure, separately, based on the Rumelhart error backpropagation method. The final secondary structure prediction result is a combination of 7 neural network predictors from different profile data and parameters. The program is freely downloadable at the bottom of this page.

PSSpred on-line

Copy and paste your sequence here (<4,000 residues, in FASTA format):

Or upload the sequence from your local computer:

Choose File No file chosen

Email: (mandatory, where results will be sent to)

ID: (optional, your given name of the protein)

Run PSSpred Clear form

We will discuss it further in Practical Section

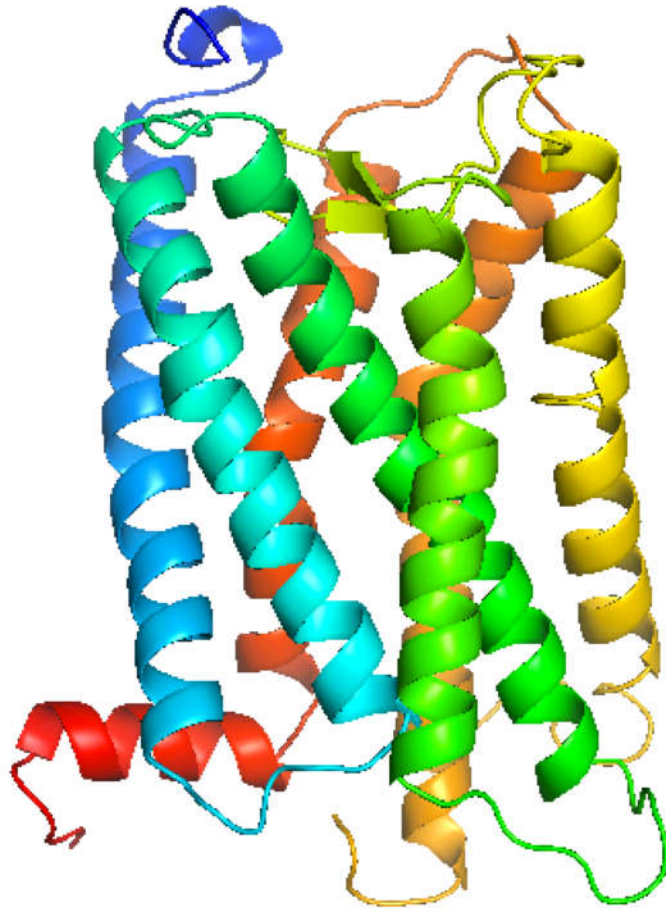
Conclusions

1. Machine-learning approach could generate the best SS prediction, significantly better than statistical or physical approaches
2. Accuracy of NN-based secondary prediction is approaching to its limit of experimental uncertainty. Accordingly, CASP stopped SS prediction competition in CASP5 (2003)
3. This study shows that combining multiple predictor algorithms can still give (statistically significant) improvement over individual predictors

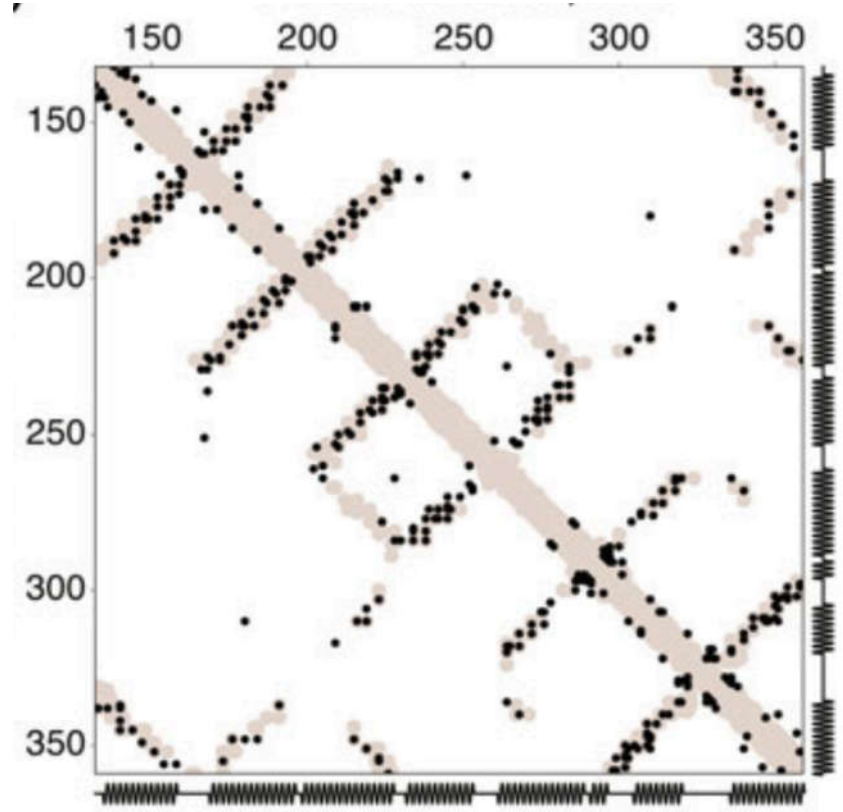
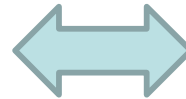
Case Studies of Machine-Learning in Structure Biology

1. Protein Secondary Structure Prediction
2. Protein Contact Prediction
3. Disease-Associated Mutation Prediction

2.1 Protein contact-map dictates the 3D fold

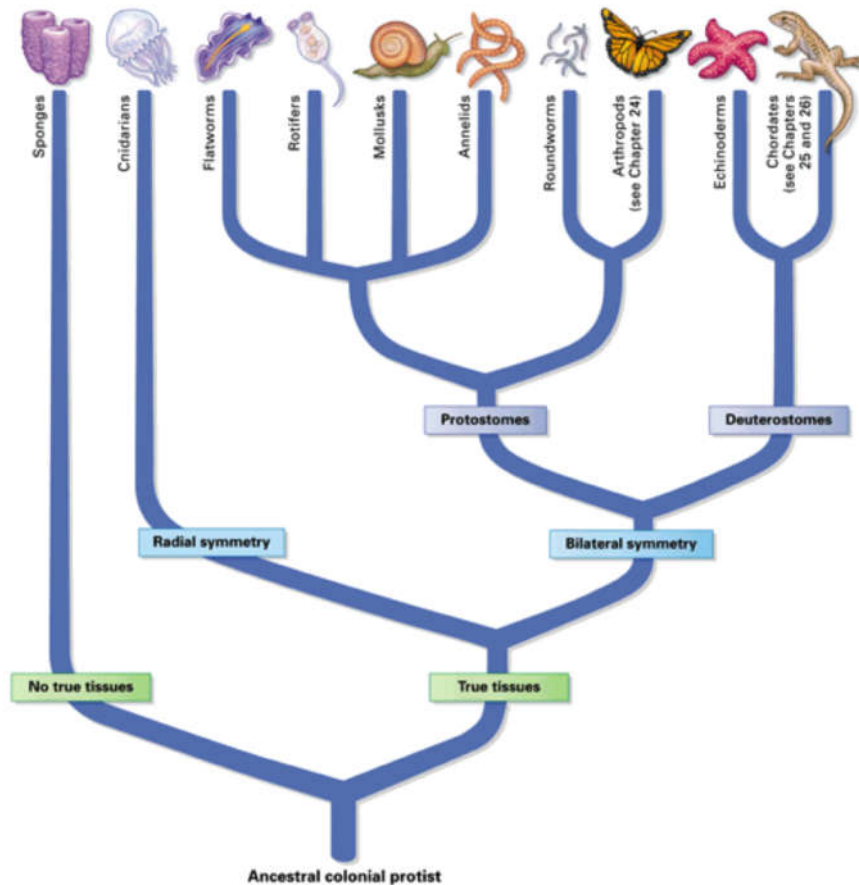


3D structure

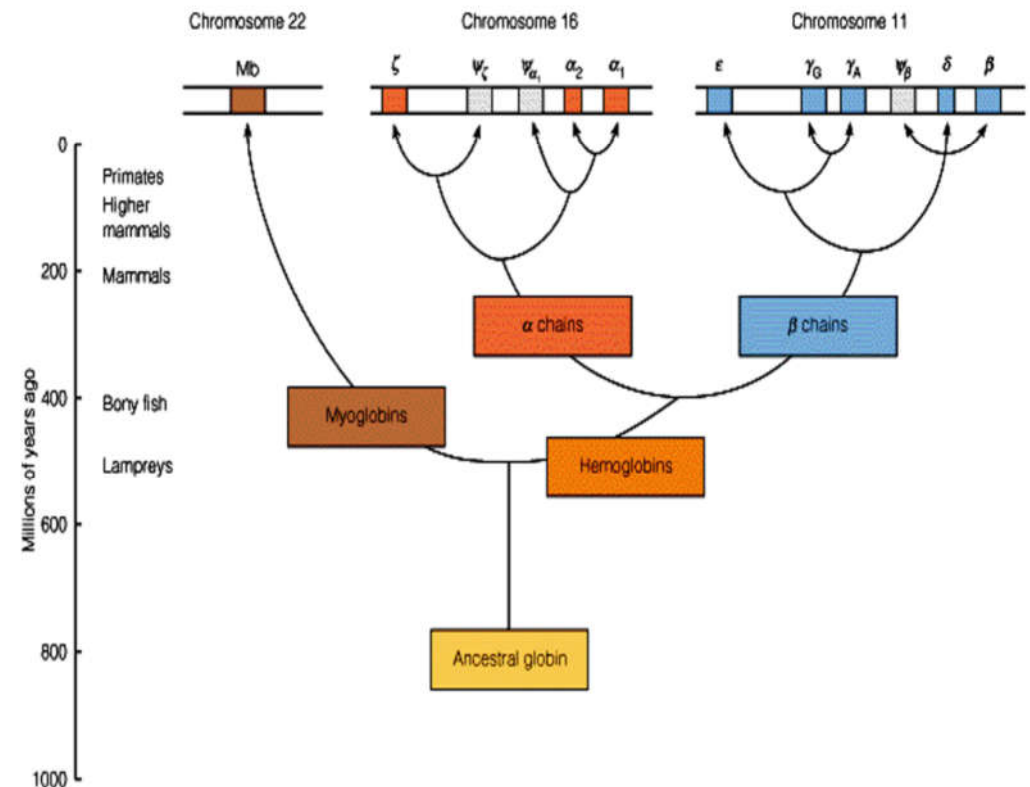


2D contact-map

2.2. Deriving contact-map from co-evolution coupling

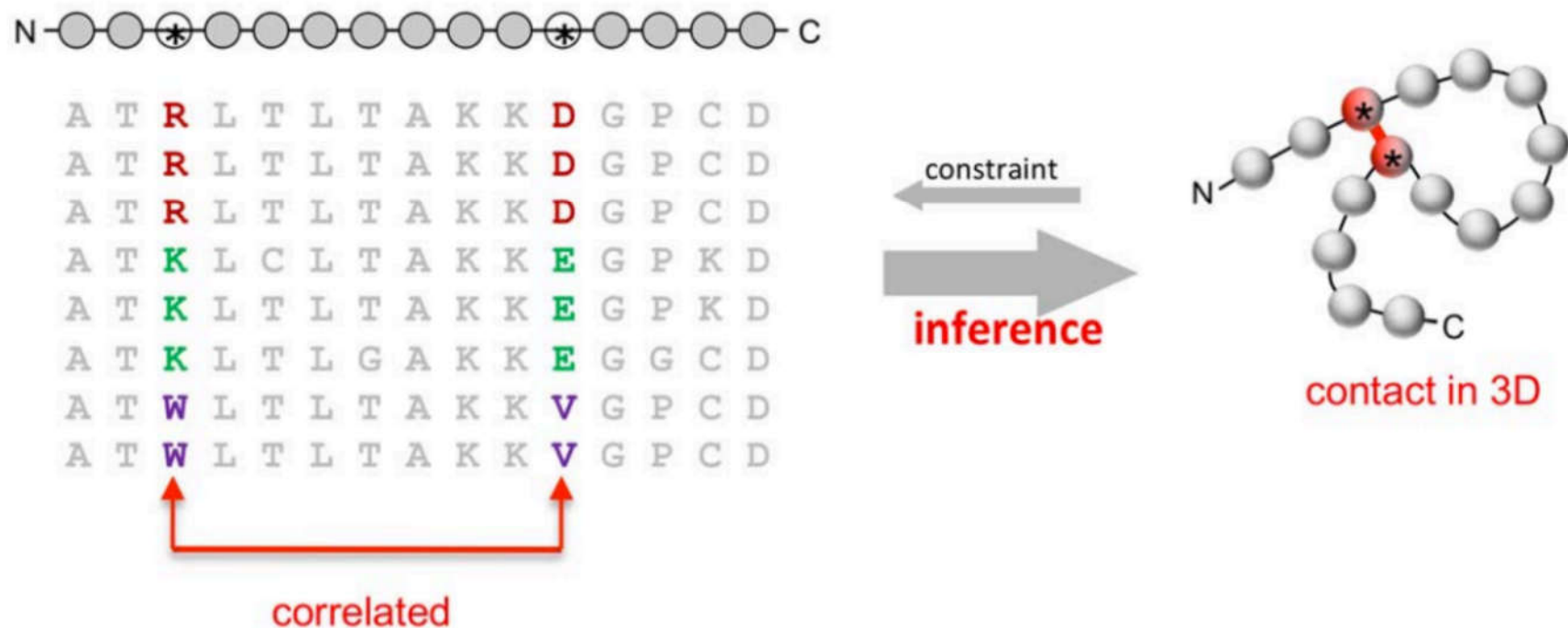


Globin evolution and expression



2.2. Deriving contact-map from co-evolution coupling

Assumption: Spatially contacted residues usually mutate cooperatively



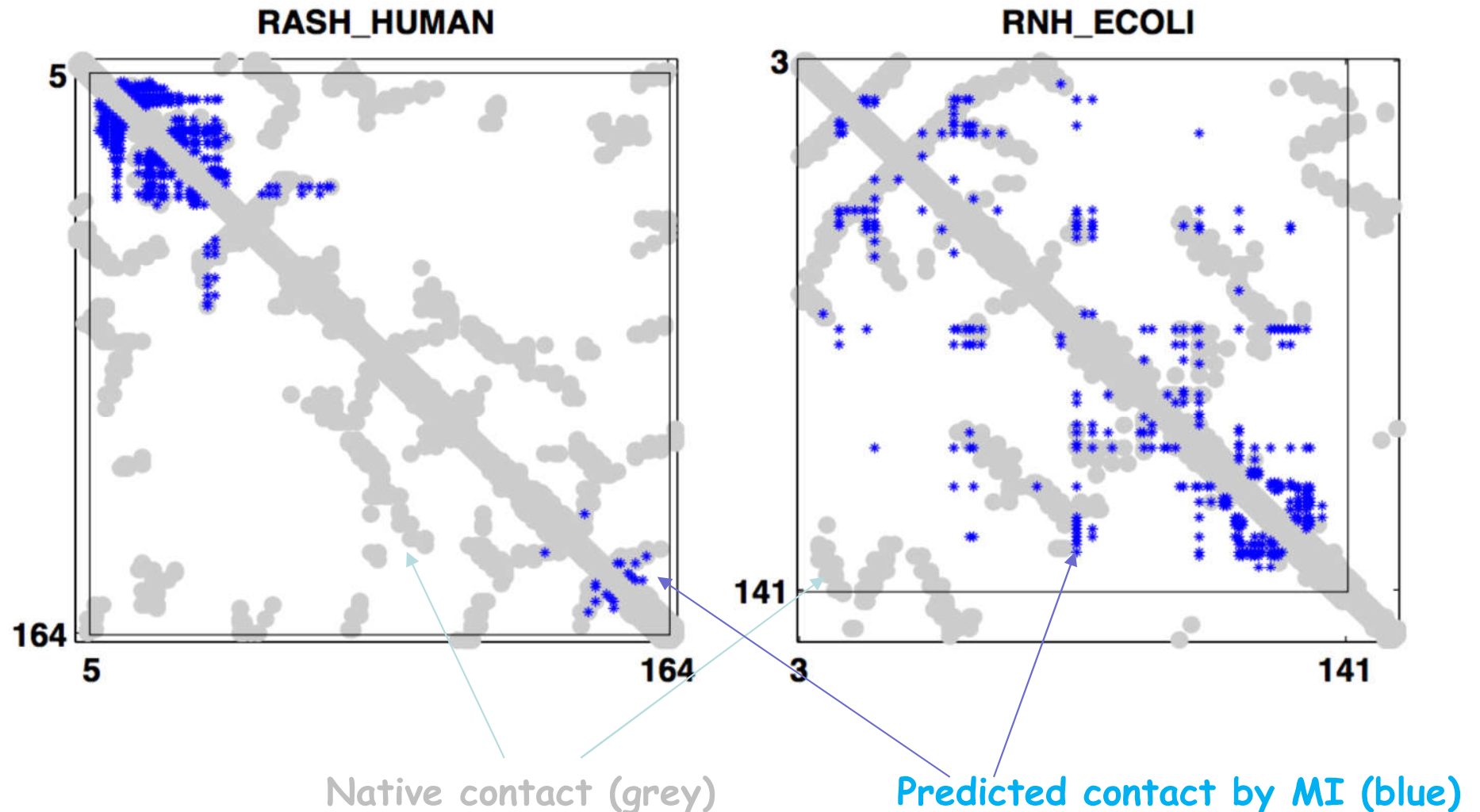
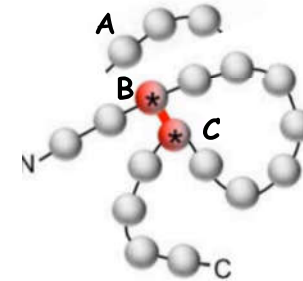
Contacts can be derived by mutual information of A_i, A_j :

$$MI_{ij} = \sum_{A_i, A_j=1}^q f_{ij}(A_i, A_j) \ln \left(\frac{f_{ij}(A_i, A_j)}{f_i(A_i) f_j(A_j)} \right)$$

Problem of MI-based co-evolution contact prediction

Transitivity issues:

If $A \leftrightarrow B$ and $B \leftrightarrow C$, $A \leftrightarrow C$;
but Residues A and C are not in contact



2.2. Deriving contact-map from Coevolution coupling

Maximum entropy model (Evfold by Marks et al):

Define:
$$P_{ij}(A_i, A_j) \equiv \sum_{\{A_k=1, \dots, q\}^{k \neq i, j}} P(A_1, \dots, A_L) = f_{ij}(A_i, A_j)$$

Unknown multivariate distribution

Request: Maximizing entropy of $P(A_1, \dots, A_L)$ consistent with data $f_{ij}(A_i, A_j)$

We have:
$$P(A_1, \dots, A_L) = \frac{1}{Z} \exp \left\{ \sum_{1 \leq i \leq j \leq L} e_{ij}(A_i, A_j) + \sum_{1 \leq i \leq L} h_i(A_i) \right\}$$

Calculating $e_{ij}(A_i, A_j)$:

Direct coupling

$$C_{ij}(A_i, A_j) = f_{ij}(A_i, A_j) - f_i(A_i)f_j(A_j)$$

20*20*L*L-dimension

$$e_{ij}(A_i, A_j) = -(C^{-1})_{ij}(A_i, A_j)$$

Contacts are predicted from DI_{ij} :

$$P_{ij}^{Dir}(A_i, A_j) = \frac{1}{Z} \exp \left\{ e_{ij}(A_i, A_j) + \tilde{h}_i(A_i) + \tilde{h}_j(A_j) \right\}$$

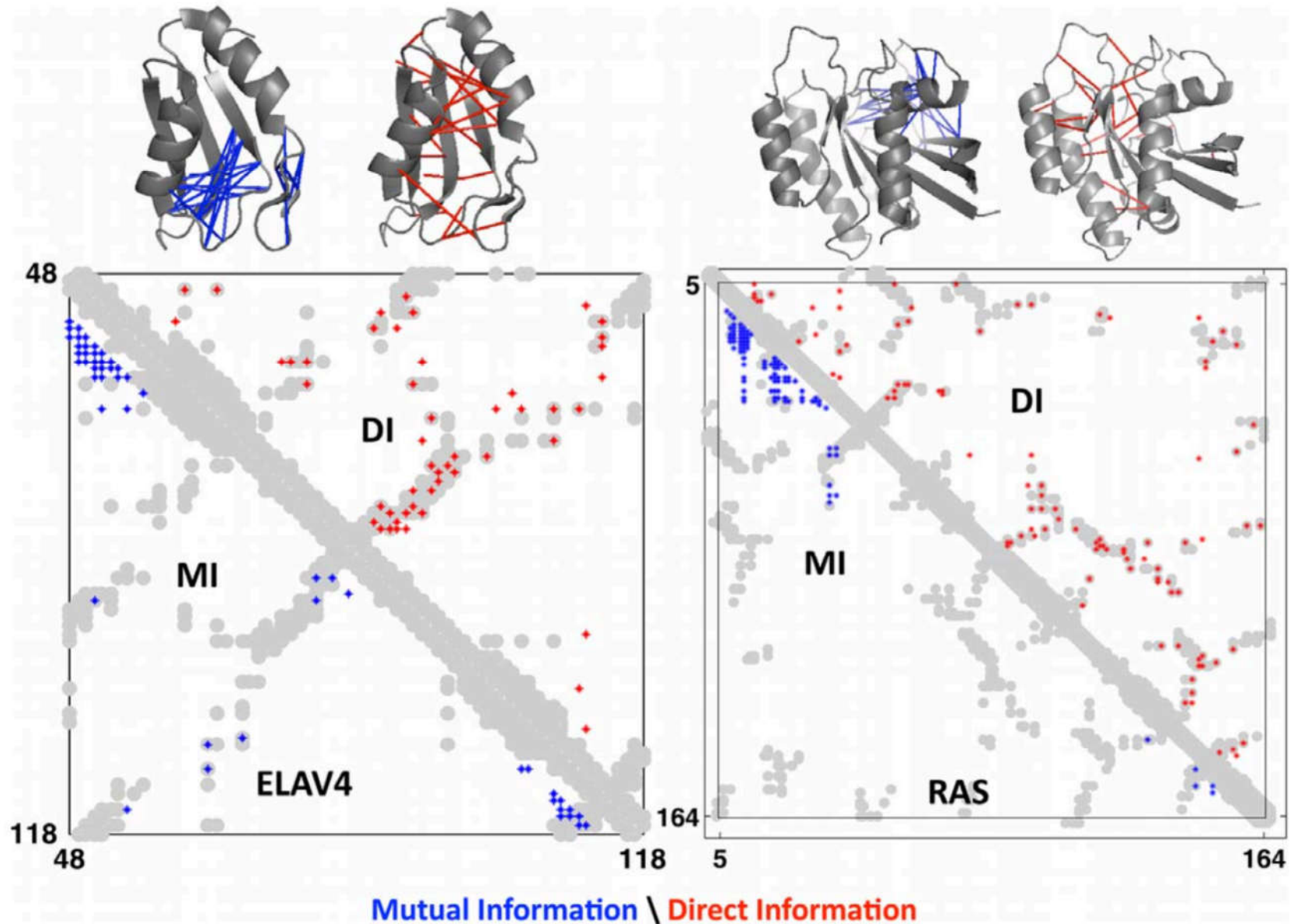
$$DI_{ij} = \sum_{A_i, A_j=1}^q P_{ij}^{Dir}(A_i, A_j) \ln \left(\frac{P_{ij}^{Dir}(A_i, A_j)}{f_i(A_i)f_j(A_j)} \right)$$

2.2. Deriving contact-map from Coevolution coupling

Sparse inverse covariance estimation (PSICOV by Jones):

$$\left\{ \begin{array}{l} PC_{ij} = S_{ij}^{\text{contact}} - \frac{\bar{S}_{(i-)}^{\text{contact}} \bar{S}_{(-j)}^{\text{contact}}}{\bar{S}^{\text{contact}}} \\ S_{ij}^{\text{contact}} = \sum_{ab} |\Theta_{ij}^{ab}| \\ S_{ij}^{ab} = E(x_i^a x_j^b) - E(x_i^a) E(x_j^b) = f(A_i B_j) - f(A_i) f(B_j) \\ \Theta_{ij}^{ab} = (S^{-1})_{ij}(a, b) \end{array} \right.$$

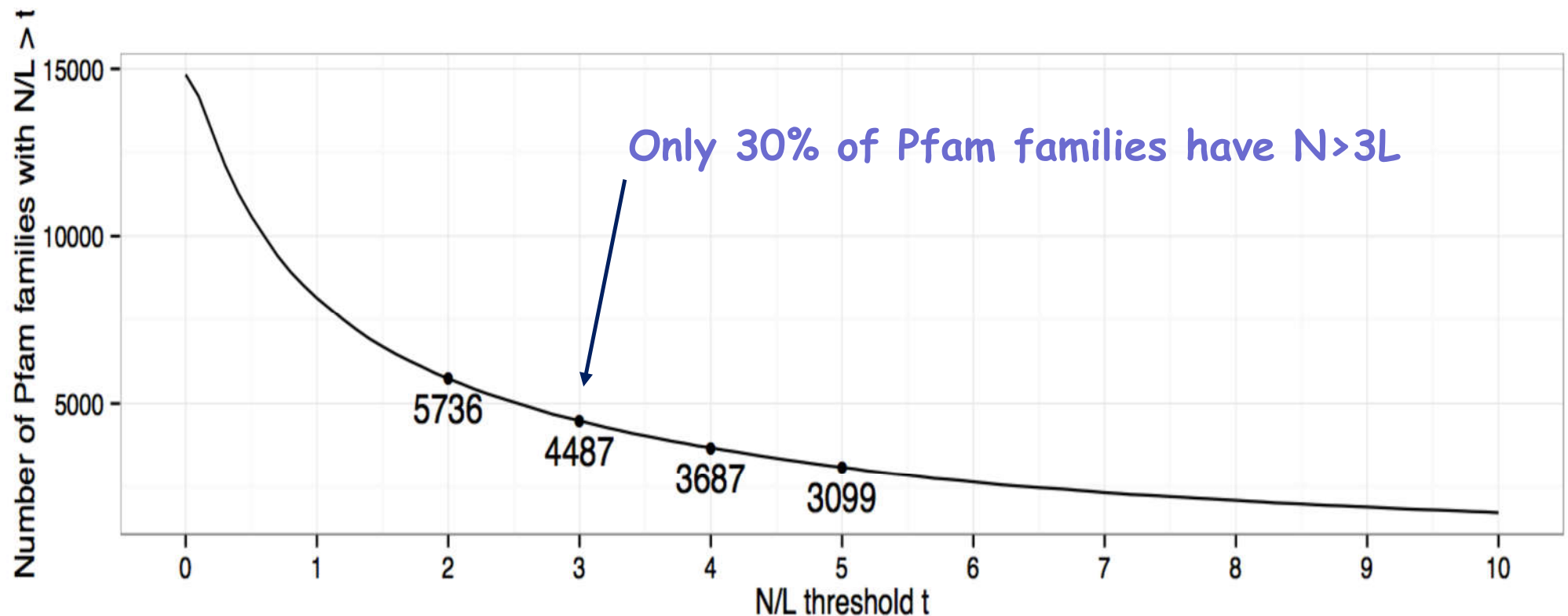
Direct coupling works better than MI due to noise removal



Problem of co-evolution contact prediction

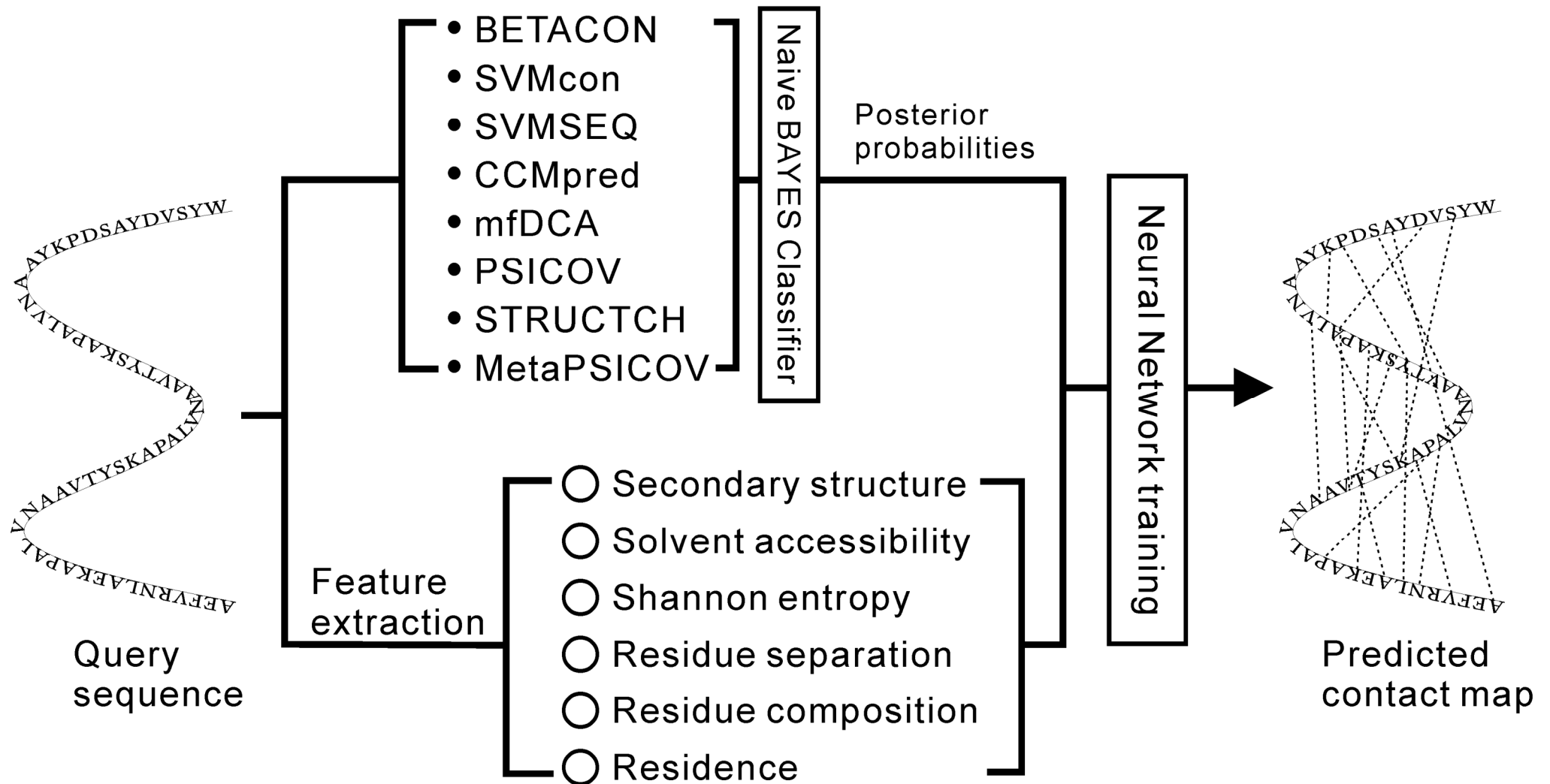
Coevolution based contact predictor works well only when a sufficient number of homologous sequence can be detected, i.e. $N > \sim 3 * L$

	260										270										280										290										300									
1n97a	h	l	p	r	e	r	a	l	s	e	a	v	t	l	l	v	a	g	h	e	t	v	a	s	a	l	t	w	s	f	l	l	l	s	h	r	p	d	w	q	k	r	v	a	e	s	e	e	a	a
1gwia	h	l	t	d	a	e	i	v	s	t	l	q	l	m	v	a	a	g	h	e	t	t	i	s	l	i	v	n	a	v	v	n	l	s	t	h	p	e	q	r	a	l	v	l	s	g	e	a	e	w
1ljpa	r	l	s	a	d	e	l	t	s	i	a	l	v	l	l	l	a	g	f	e	s	s	v	s	l	i	g	i	g	t	y	l	l	l	t	h	p	d	q	l	a	l	v	r	r	d	p	s	a	l
1jfb	n	i	d	k	s	d	a	v	q	i	a	f	l	l	l	v	a	g	n	a	t	m	v	n	m	i	a	l	g	v	a	t	l	a	q	h	p	d	q	l	a	q	l	k	a	n	p	s	l	a
1cpt	y	i	d	d	k	y	i	n	a	y	y	v	a	i	a	t	a	g	h	d	t	t	s	s	s	s	g	g	a	i	i	g	l	s	r	n	p	e	q	l	a	l	a	k	s	d	p	a	i	
1n40a	h	v	s	d	e	l	f	a	t	i	g	v	t	f	f	g	a	g	v	i	s	t	g	s	f	l	t	t	a	l	i	s	l	i	q	r	p	q	l	r	n	l	l	h	e	k	p	e	l	i
1lfka	d	a	t	d	e	e	l	r	g	f	c	v	q	v	m	l	a	g	d	d	n	i	s	g	m	i	g	l	g	v	l	a	m	l	r	h	p	e	q	i	d	a	f	r	g	d	e	q	s	a
1qmqa	p	i	t	s	d	e	a	k	r	m	c	g	l	l	l	v	g	g	d	t	v	v	n	f	l	s	f	s	m	e	f	l	a	k	s	p	e	h	r	q	e	l	i	e	r	p	e	r	i	
1io7a	-	l	s	d	i	e	k	l	g	y	i	i	l	l	i	a	g	n	e	t	t	n	l	i	s	n	s	v	i	d	f	t	r	f	-	-	n	l	w	q	r	i	e	e	n	l	y			
1jpza	p	l	d	d	e	n	i	r	y	q	i	t	t	f	l	i	a	g	h	e	t	t	s	g	l	s	f	a	l	y	f	l	v	k	n	p	h	v	l	q	k	a	a	e	e	a	r	v		
1e9xa	r	f	s	a	d	e	i	t	g	m	f	i	s	m	m	f	a	g	h	h	t	s	s	g	t	a	s	w	t	l	i	e	l	m	r	h	r	d	a	y	a	a	v	i	d	e	l	d		
1dt6a	e	f	t	l	e	s	l	v	i	a	v	s	d	l	f	g	a	g	t	e	t	t	s	t	l	t	r	y	s	l	l	l	l	l	k	h	p	e	v	a	a	r	v	q	e	e	i	e	r	v
Consistency	3	6	5	4	5	4	5	4	3	3	3	4	4	6	6	4	9	3	5	7	5	5	5	4	6	4	3	6	6	5	3	8	5	5	5	7	5	3	4	4	5	3	5	5	4	5	4	5		



Family statistics for 14831 Pfam families

2.3 NeBcon: A new machine-learning approach to contact prediction (by combining neural-network training and naïve Bayes classifier)

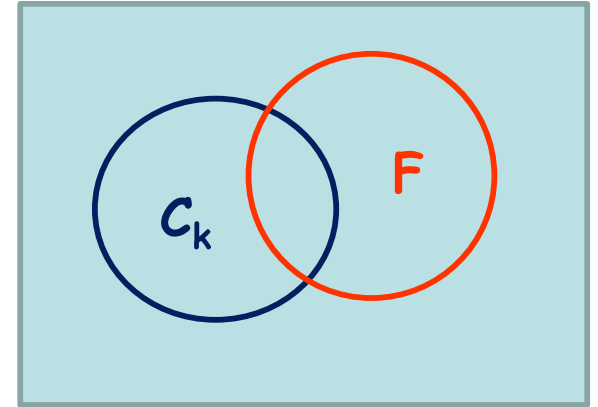


What is naïve Bayes classifier?

Given:

$F=(f_1, f_2, \dots, f_n)$: n specific features

C_k : k 'th possible outcome



Naïve Bayes classifier theorem:

$$P(C_k|f_1, f_2, \dots, f_n) = P(C_k|F) = \frac{P(C_k)P(F|C_k)}{P(F)}$$

$$P(C_k|F) \propto P(C_k)P(F|C_k) = P(C_k)P(f_1|f_2, \dots, f_n, C_k)P(f_2|f_3, \dots, f_n, C_k) \dots P(f_n|C_k)$$

Under naïve assumption (ie, all features are independent): $P(f_i|f_{i+1}, \dots, f_n, C_k) = P(f_i|C_k)$

$$P(C_k|F) \propto P(C_k) \prod_{i=1}^n P(f_i|C_k)$$

$$P(C_k|F) = \frac{P(C_k) \prod_{i=1}^n P(f_i|C_k)}{P(F)}$$

An example of application of naïve Bayes classifier

Training data:

Sex	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Target sample:

height (feet)	weight (lbs)	foot size(inches)
6	130	8

Question: is this target a male or female?

Solution:

Naïve approach: by vote and consensus

- Height: male
- Weight: female
- Foot size: female

Conclusion: female

An example of application of naïve Bayes classifier

Training data:

Sex	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Target sample:

height (feet)	weight (lbs)	foot size(inches)
6	130	8

\uparrow \uparrow \uparrow
 f_1 f_2 f_3

Question: is this target a male or female?

Solution:

$$P(C_k|F) = \frac{P(C_k) \prod_{i=1}^n P(f_i|C_k)}{P(F)}$$

$$\text{posterior (male)} = \frac{P(\text{male}) p(\text{height} | \text{male}) p(\text{weight} | \text{male}) p(\text{foot size} | \text{male})}{\text{evidence}}$$

$$P(\text{male}) = 0.5$$

$$p(\text{height} | \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6 - \mu)^2}{2\sigma^2}\right) \approx 1.5789$$

$$p(\text{weight} | \text{male}) = 5.9881 \cdot 10^{-6}$$

$$p(\text{foot size} | \text{male}) = 1.3112 \cdot 10^{-3}$$

$$\text{posterior numerator (male)} = \text{their product} = 6.1984 \cdot 10^{-9}$$

$$\text{posterior (female)} = \frac{P(\text{female}) p(\text{height} | \text{female}) p(\text{weight} | \text{female}) p(\text{foot size} | \text{female})}{\text{evidence}}$$

$$P(\text{female}) = 0.5$$

$$p(\text{height} | \text{female}) = 2.2346 \cdot 10^{-1}$$

$$p(\text{weight} | \text{female}) = 1.6789 \cdot 10^{-2}$$

$$p(\text{foot size} | \text{female}) = 2.8669 \cdot 10^{-1}$$

$$\text{posterior numerator (female)} = \text{their product} = 5.3778 \cdot 10^{-4}$$

Suppose sample follow Gaussian distribution

The major advantage of NBC over consensus is that the NBC combination considers specific distribution of individual features.

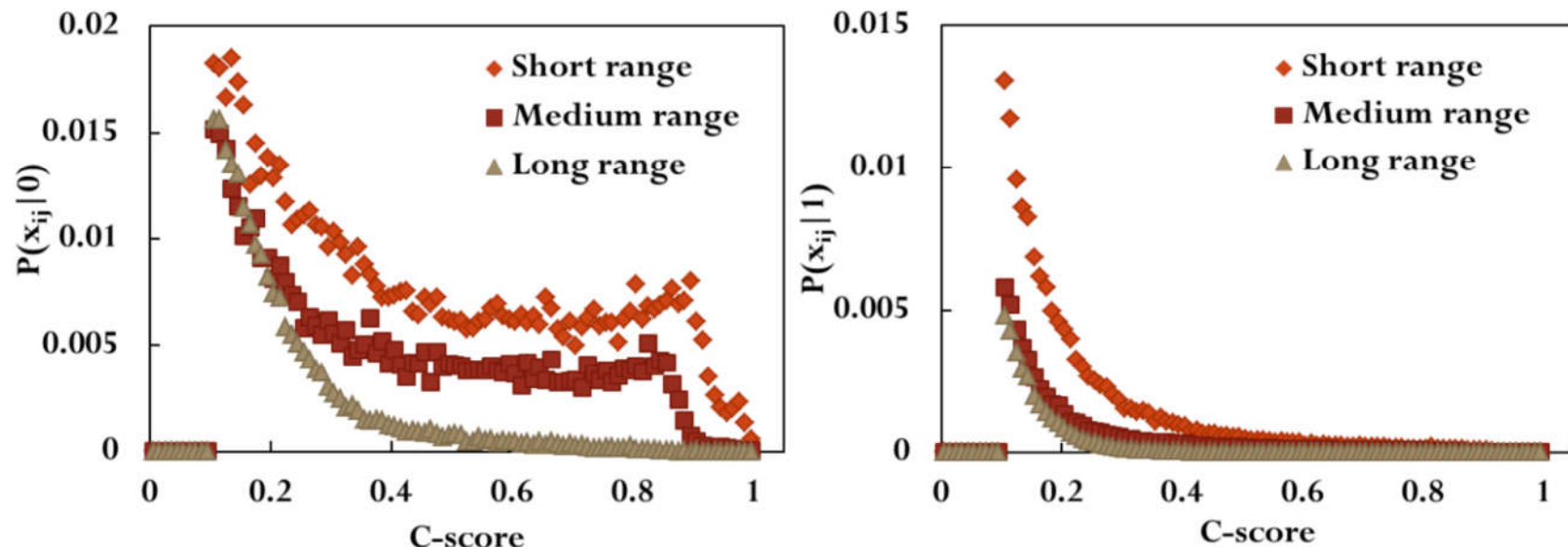
2.3 NeBcon: Combining neural-network training and naïve Bayes classifier for protein contact map prediction

Feature type-I: posterior probability of meta-predictors (121 features):

$$P(C|X_{ij}) = \frac{P(C) \prod_{m=1}^N P(X_{ij}^m|C)}{P(X_{ij})} = \frac{P(C) \prod_{m=1}^N P(X_{ij}^m|C)}{P(0) \prod_{m=1}^N P(X_{ij}^m|0) + P(1) \prod_{m=1}^N P(X_{ij}^m|1)}$$

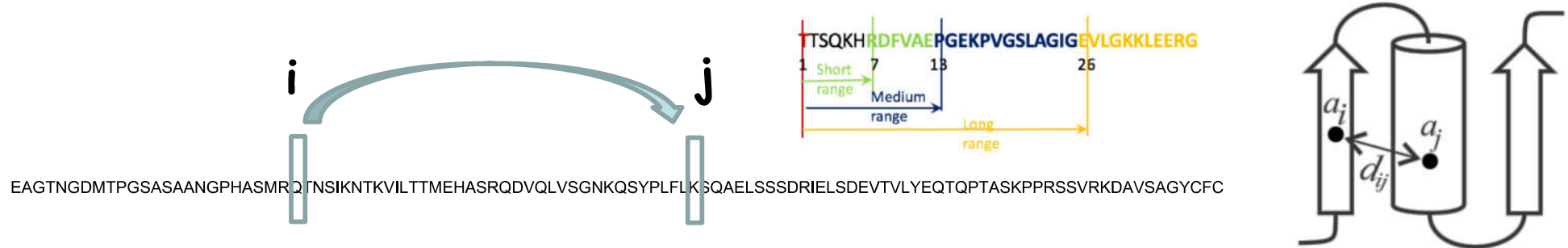
Conditional probabilities:

One of the examples (SVMSEQ) for 8 contact predictors:



2.3 NeBcon: Combining neural-network training and naïve Bayes classifier for protein contact map prediction

Feature type-II: inherent physicochemical feature collection (596 features):



1. $X_i=0,1$ when i 'th residue within sequence (22=11x2 features)
2. Secondary structure by PSSpred (66=11x2x3 features)
3. Solvent accessibility of target residues (22=11x2 features)
4. Shannon entropy of i 'column in PsiBlast MSA (22=11x2 features):

$$x_i = \sum_{k=1}^{21} p_k^i \ln p_k^i$$

5. Sequence separation (2 features): $x=|i-j|$
6. Sequence profile (462=11x2x21 features)


6. Sequence profile (462=11x2x21 features):

Query sequence:

EAGTNGDMTPGSASAANGPHASMRQTNSIKNTKVILTTMEHAS
RQDVQLVSGNKQSYPLFLKSQAELSSSDRIELSDEVTVLYEQTQ
PTASKPPRSSVRKDAVSAGYCF

Sequence profile:

PSI-BLAST



POS	PROBE	CONSENSUS	PROFILE																				
			A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	E G V L	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	L L S P	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
3	V V V V	V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
4	K E A T	A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
5	A P L P	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
6	G G G G	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
7	S S Q E	D	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
8	S S T P	S	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
9	V L V A	V	5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
10	K R R S	R	0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
11	M L I I	I	0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
12	S S T S	S	4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
13	C C C C	C	3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
14	K S Q R	K	1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9
15	A A G S	A	10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
16	T S D S	S	4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
17	G G S Q	G	5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9
18	Y F L S	F	-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
19	T T R L	T	1	-2	0	1	0	0	0	2	2	2	3	1	1	3	1	7	2	1	-2	9	9
20	F F . L	F	-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
21	S S . D	S	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
22	S . . S	S	2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
23	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
24	. . . D	D	1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
25	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
26	. A G N	A	6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
27	Y N Y T	Y	0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
28	E D D Y	D	2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
29	L M A L	L	3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
30	Y N A W	N	4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
.
48	S G N S	S	4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
49	S S N Y	S	2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

i-th window

j-th window

Neural Network Training

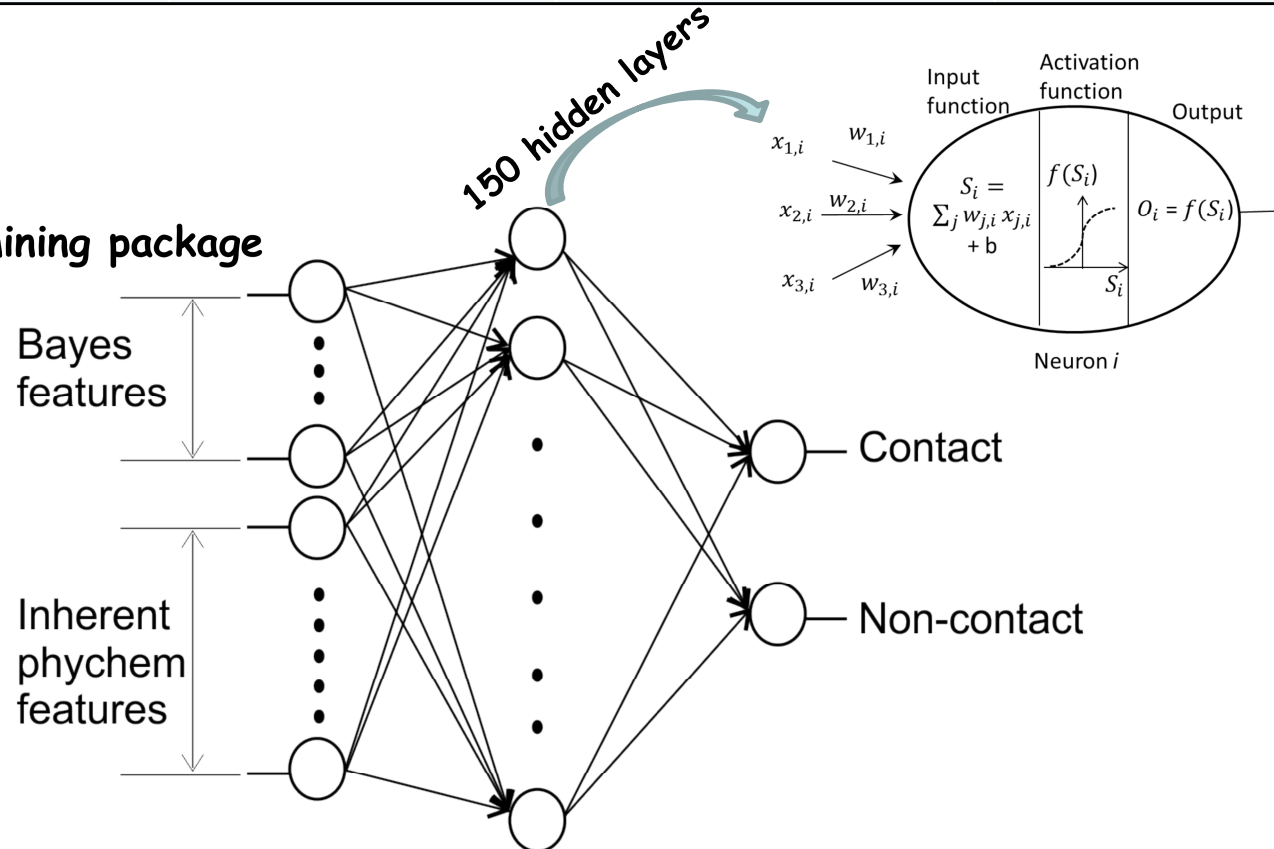
Training set:

517 non-homologous proteins containing:

	Short-range $ i-j < 7$	Medium range $6 < i-j < 24$	Long-range $23 < i-j $
#true contacts	20,636	26,798	87,200
#residue pairs	407,036	757,315	209,080

Training package:

Weka data mining package



Test Results

Test protein set:

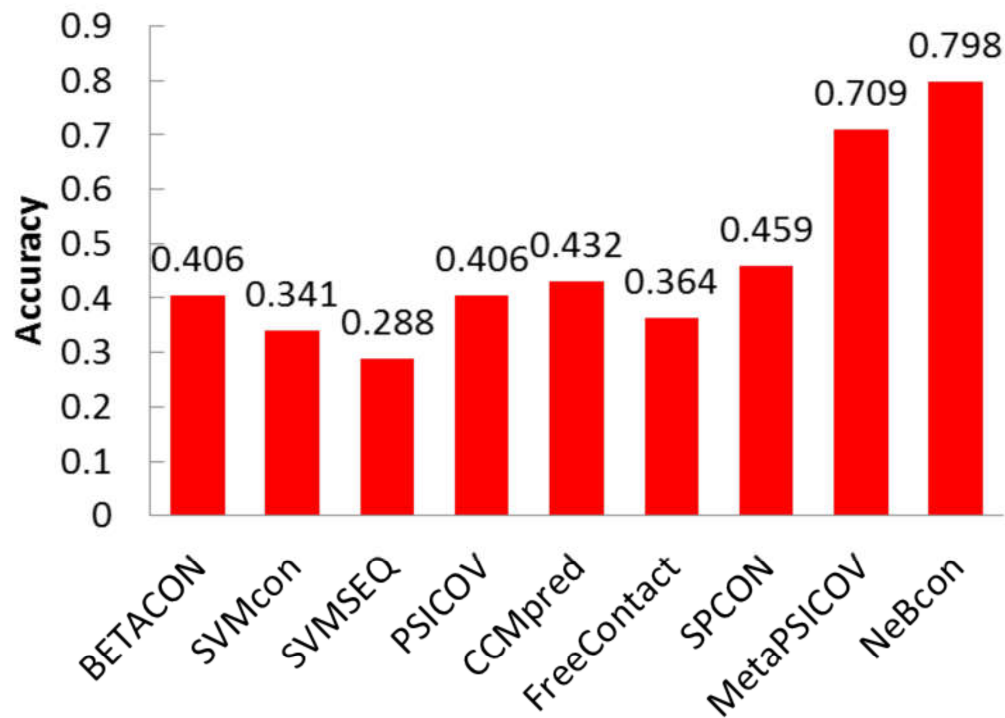
98 proteins containing 3850, 5849, 13792 short-, medium- and long-range contacts

Accuracy of the prediction: $Acc = N_{corr}/N_T$

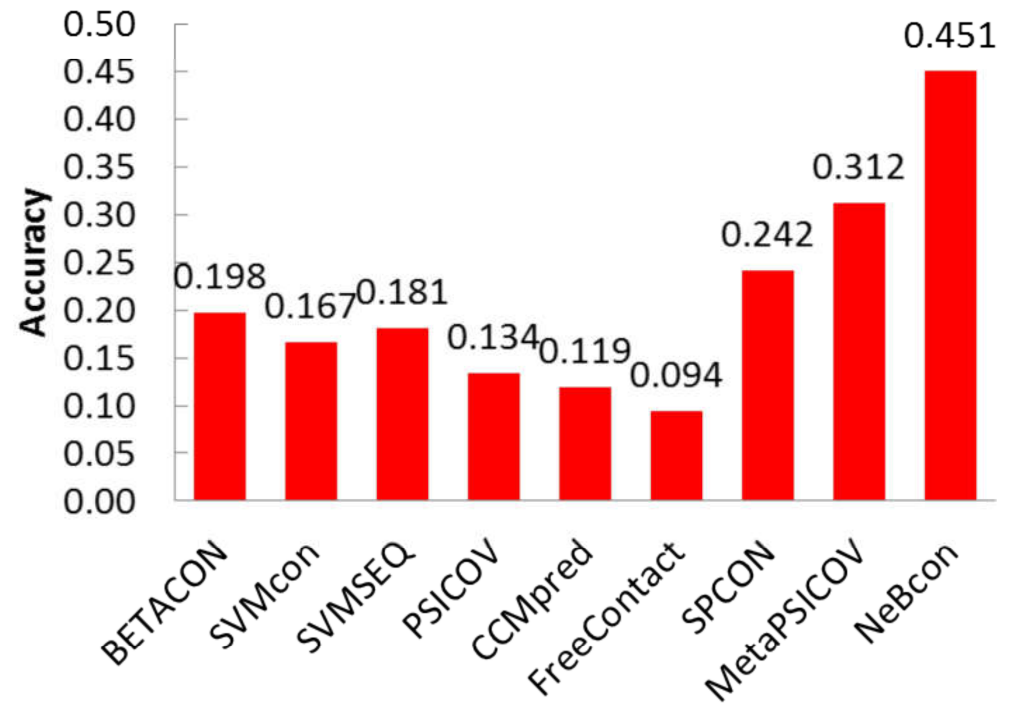
• N_{corr} = # of correctly predicted contacts in the contact map

• N_T = # of predicted contacts in the contact map

Results:



50 easy targets
Top L/5 long range



48 hard targets
Top L/5 long range

Test Results

Combine all target together:

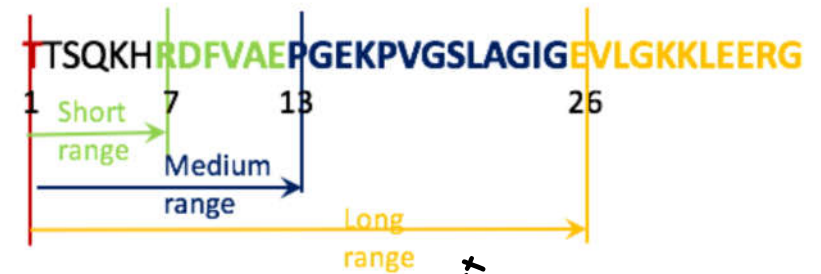
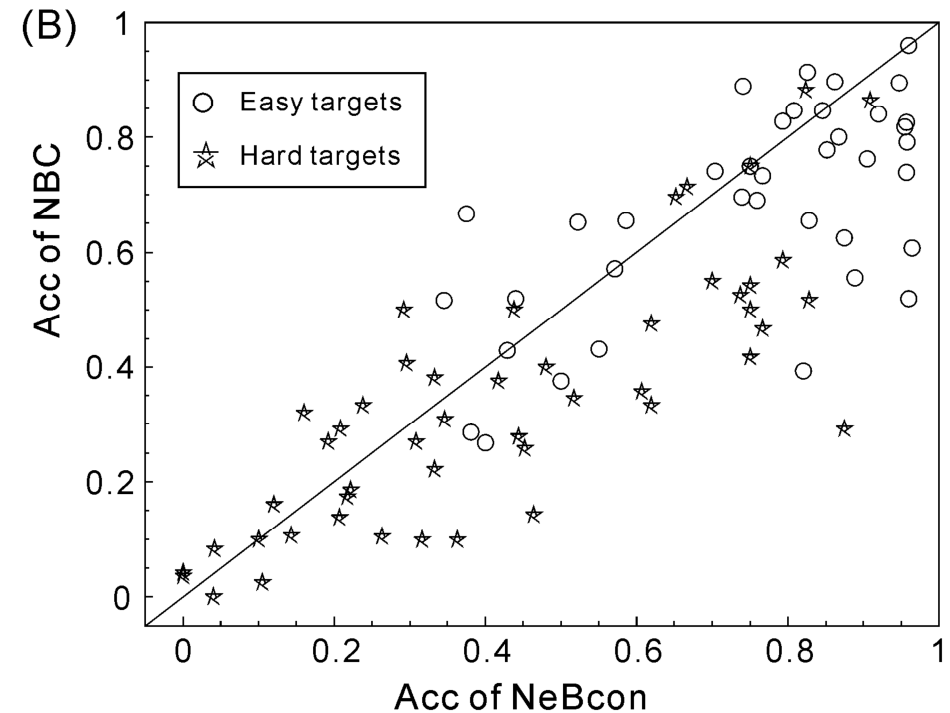
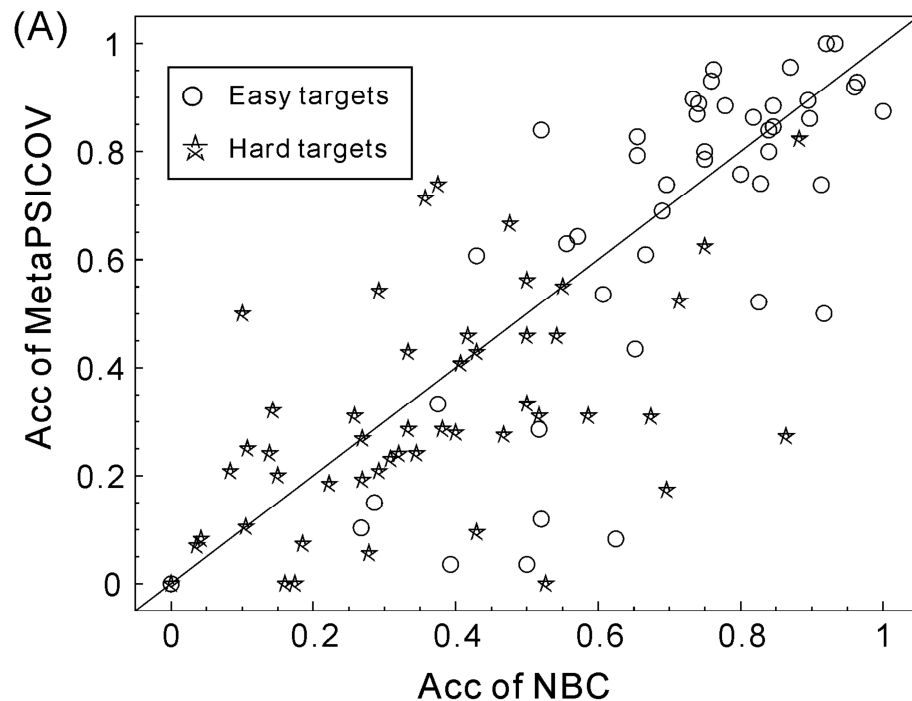


Table 2. Average accuracy of top $L/5$ contact predictions by different methods on 98 test proteins

Methods	Short (6–11)	Medium (12–24)	Long (>24)
BETACON	0.540 ($1 \cdot 10^{-9}$)	0.430 ($3 \cdot 10^{-10}$)	0.310 ($2 \cdot 10^{-12}$)
SVMSEQ	0.475 ($2 \cdot 10^{-12}$)	0.393 ($2 \cdot 10^{-12}$)	0.255 ($2 \cdot 10^{-12}$)
SVMcon	0.564 ($4 \cdot 10^{-9}$)	0.455 ($1 \cdot 10^{-8}$)	0.236 ($2 \cdot 10^{-12}$)
PSICOV	0.204 ($2 \cdot 10^{-12}$)	0.246 ($2 \cdot 10^{-12}$)	0.262 ($2 \cdot 10^{-12}$)
CCMpred	0.206 ($2 \cdot 10^{-12}$)	0.238 ($2 \cdot 10^{-12}$)	0.278 ($2 \cdot 10^{-12}$)
FreeContact	0.234 ($2 \cdot 10^{-12}$)	0.278 ($2 \cdot 10^{-12}$)	0.232 ($2 \cdot 10^{-12}$)
STRUCTCH	0.605 ($3 \cdot 10^{-4}$)	0.487 ($4 \cdot 10^{-5}$)	0.353 ($2 \cdot 10^{-12}$)
MetaPSICOV	0.576 ($5 \cdot 10^{-6}$)	0.572 ($5 \cdot 10^{-1}$)	0.515 ($2 \cdot 10^{-7}$)
NeBcon	0.651	0.574	0.628

NeBcon significantly outperforms all individual contact predictors

Comparison of NeBcon to the best predictor



1. Bayes combination contributes to overall performance
2. NN training increases accuracy for hard targets that have low number of homologous sequences

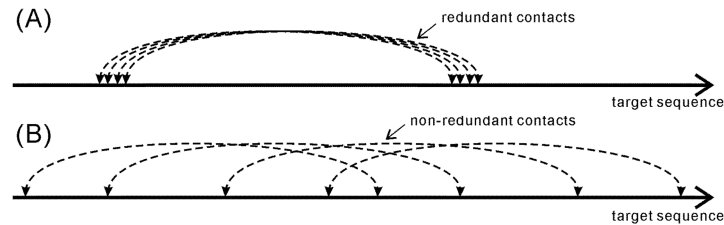
Testing results on CASP targets

Contact prediction on the free-modeling (FM) targets in CASP

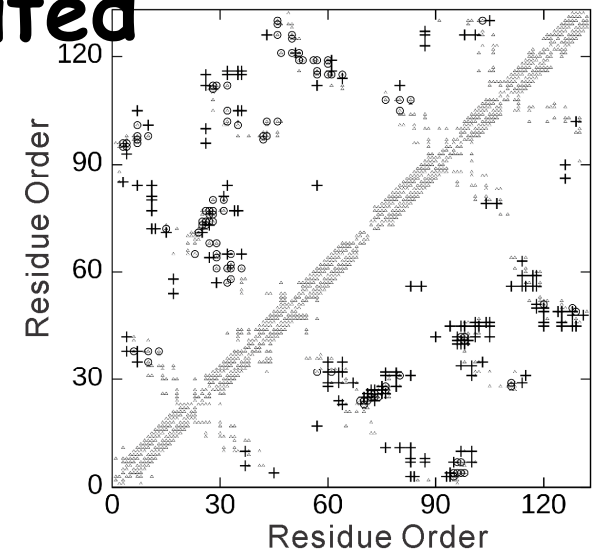
CASP10 (20 FM targets)		CASP11 (33 FM targets)	
Methods	Accuracy (p-value)	Methods	Accuracy (p-value)
NeBcon	0.4659	NeBcon	0.3763
Multicon	0.4058 (2.5×10^{-1})	MetaPSICOV	0.3632 (3.9×10^{-1})
Distill_roll	0.2804 (7.5×10^{-3})	Pcons-net	0.2482 (1.0×10^{-3})
Distill	0.2448 (3.5×10^{-3})	Shen-group	0.2330 (4.5×10^{-3})
Multicon-Const	0.2252 (4.0×10^{-4})	UCI-IGB-CMpro	0.2199 (4.8×10^{-3})
IGBteam	0.2038 (1.3×10^{-4})	RBO_Aleph	0.1990 (2.9×10^{-3})
SAM-T08-server	0.1924 (1.8×10^{-4})	LEE	0.1988 (1.1×10^{-3})
MetaPSICOV	0.1721 (4.5×10^{-5})	Multicom-Clust	0.1831 (4.1×10^{-4})
RaptorX-Roll	0.1573 (6.2×10^{-5})	RaptorX-Contact	0.1605 (7.2×10^{-5})
Multicon-Novel	0.1514 (7.6×10^{-6})	Multicon-Const	0.1559 (4.7×10^{-5})
ZHOU-SPARKS-X	0.0864 (3.0×10^{-6})	Distill	0.0782 (2.7×10^{-6})

Evenness of contact-map distributed

Shannon entropy of predicted contact-map



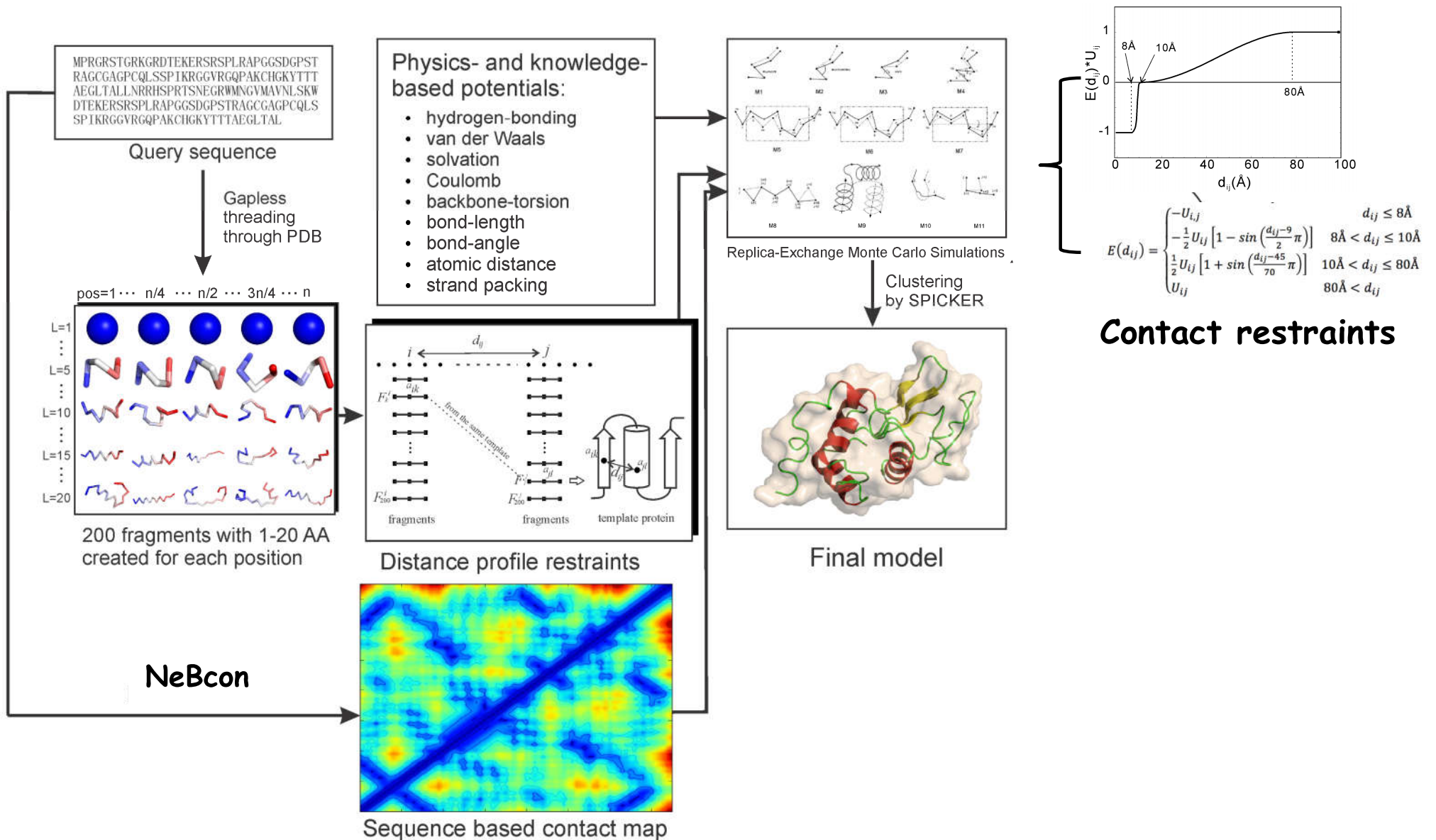
$$H = - \sum_i^{100} p_i \log_2 p_i$$



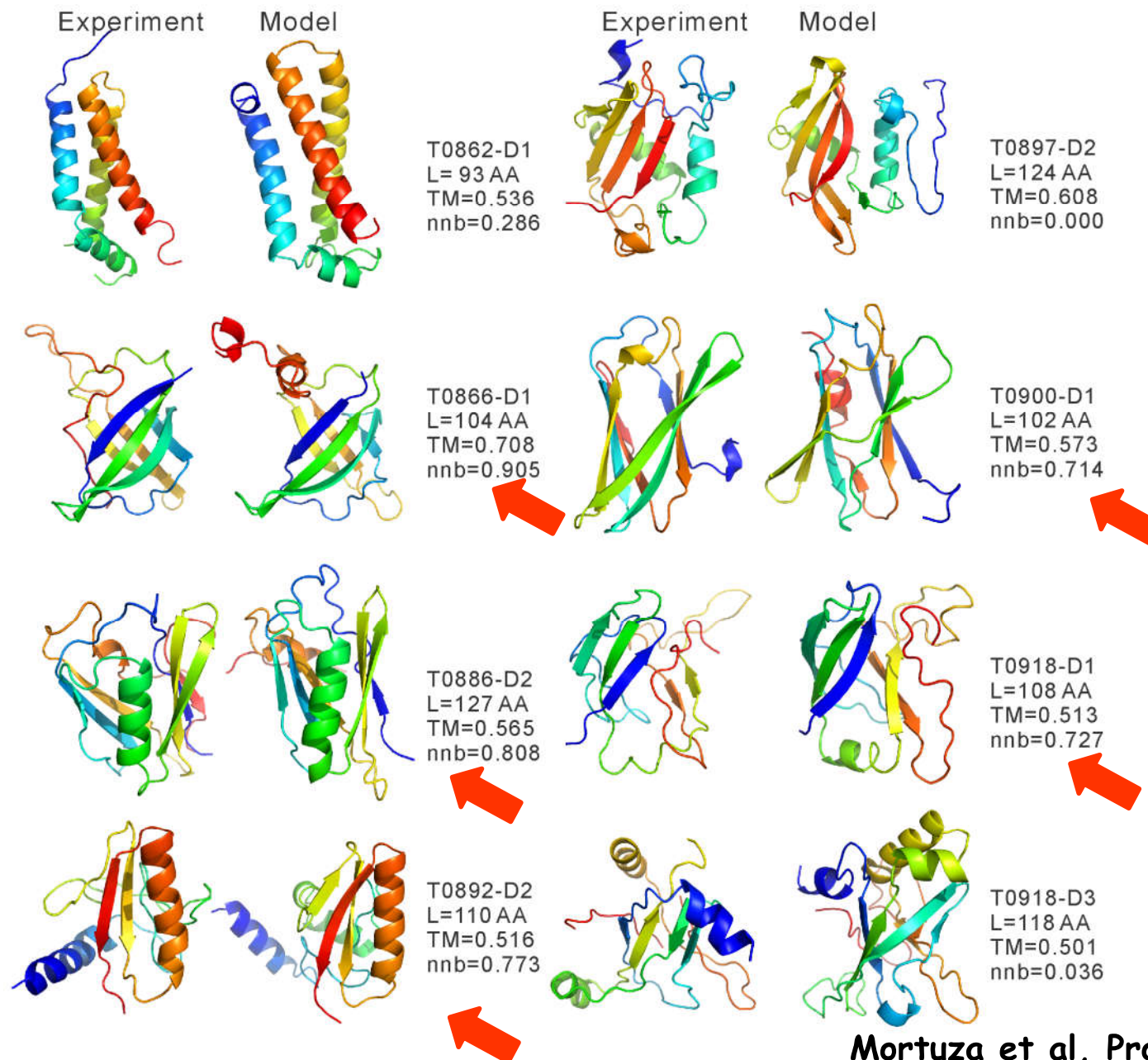
Methods	Short	Medium	Long	All
BETACON	2.705 (2.2*10 ⁻⁹)	1.953 (5.2*10 ⁻¹⁸)	2.656 (8.4*10 ⁻¹⁶)	3.912 (6.9*10 ⁻²⁵)
SVMSEQ	2.680 (5.6*10 ⁻¹⁵)	2.523 (2.0*10 ⁻⁸)	3.540 (4.9*10 ⁻⁷)	4.146 (5.6*10 ⁻¹³)
SVMcon	2.589 (5.7*10 ⁻¹⁶)	2.402 (5.2*10 ⁻¹³)	3.289 (1.5*10 ⁻¹⁶)	3.962 (1.2*10 ⁻²⁴)
PSICOV	2.676 (2.6*10 ⁻²)	2.726 (2.6*10 ⁻¹)	3.505 (6.2*10 ⁻²)	3.959 (1.23*10 ⁻²)
CCMpred	3.377 (1.3*10 ⁻⁷)	3.472 (8.3*10 ⁻¹³)	4.415 (6.9*10 ⁻⁹)	5.016 (1.1*10 ⁻⁶)
FreeContact	3.245 (2.0*10 ⁻⁴)	3.426 (2.1*10 ⁻¹¹)	4.478 (4.5*10 ⁻¹⁰)	4.977 (5.0*10 ⁻⁶)
STRUCTCH	2.723 (1.5*10 ⁻⁹)	2.647 (3.0*10 ⁻⁴)	3.477 (2.6*10 ⁻⁸)	4.072 (7.7*10 ⁻¹⁷)
MetaPSICOV	2.958 (3.8*10 ⁻¹)	2.709 (1.4*10 ⁻²)	3.552 (4.0*10 ⁻⁵)	4.217 (9.7*10 ⁻⁶)
NeBcon	2.750 (5.6*10 ⁻¹⁰)	2.570 (7.8*10 ⁻¹⁰)	3.665 (6.5*10 ⁻⁵)	4.273 (3.3*10 ⁻⁹)
Native	2.973	2.823	3.815	4.473

Most predictors create contact map with similar evenness as native

C-QUARK: Using contact-map prediction to guide *ab initio* protein structure folding in CASP12



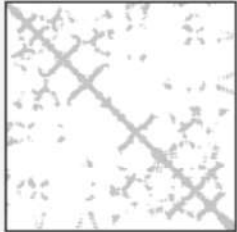
5 of 8 successful *ab initio* folding cases in CASP12 are due to contact prediction



NeBcon is freely available at <http://zhanglab.ccmb.med.umich.edu/NeBcon/>

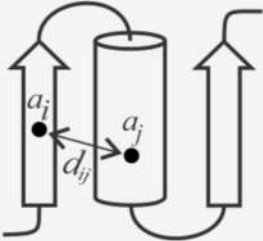
Online Services

- I-TASSER
- QUARK
- LOMETS
- COACH
- COFACTOR
- MUSTER
- SEGMER
- FG-MD
- ModRefiner
- REMO
- SPRING
- COTH
- BSpred
- SVMSEQ
- ANGLOR
- BSP-SLIM
- SAXSTER
- ThreaDom
- EvoDesign
- GPCR-I-TASSER
- BindProf
- BindProfX
- ResQ
- IonCom



NeBcon

Accurate prediction of protein contact maps



NeBcon (Neural-network and Bayes-classifier based contact prediction) is a hierarchical algorithm for sequence-based protein contact map prediction. It first uses the naive Bayes classifier theorem to calculate the posterior probability of eight machine-learning and co-evolution based contact prediction programs (SVMSEQ, BETACON, SVMcon, PSICOV, CCMpred, FreeContact, MetaPSICOV, and STRUCTCH). Final contact maps are then created by neural network machine that trains the posterior probability scores with intrinsic structural features from secondary structure, solvent accessibility, and Shannon entropy of multiple sequence alignments.

NeBcon On-line ([view an example of NeBcon output](#))

Cut and paste your sequence (in [FASTA format](#)) below:

Or upload the sequence from your local computer: No file chosen

Email: (mandatory, where results will be sent to)

ID: (optional, your given name of the protein)

Download package:

The standalone NeBcon package can be downloaded from [NeBconpackage.tar.gz](#). In order to install and run the package, follow the instructions

We will discuss on its application in Pratical Section

Conclusions

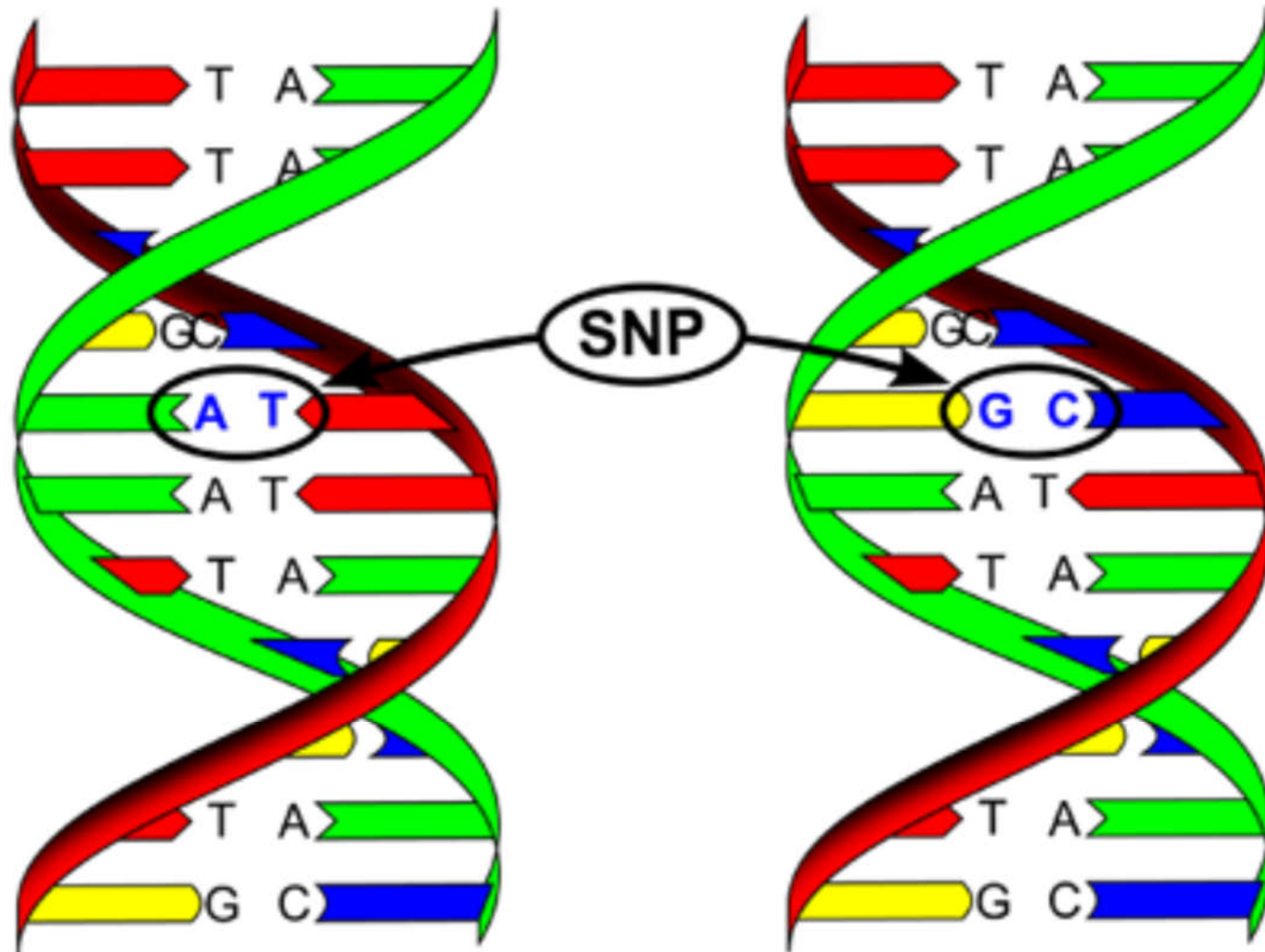
1. Contact prediction can improve accuracy of ab initio protein structure for targets without templates. This is particularly true given (a) sequence library increases; (b) new methods for removing translation correlation
2. Naïve Bayes classifier helps combining multiple contact predictors
3. NN training on inherent protein features improves contact prediction for hard targets

Case Studies of Machine-Learning in Structure Biology

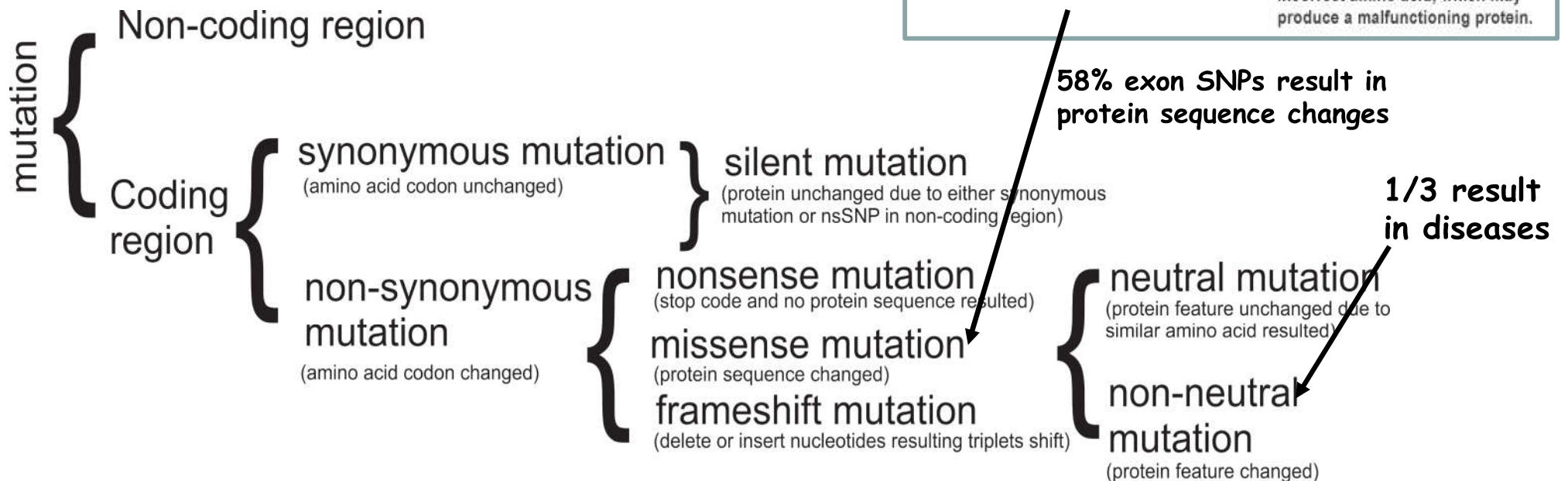
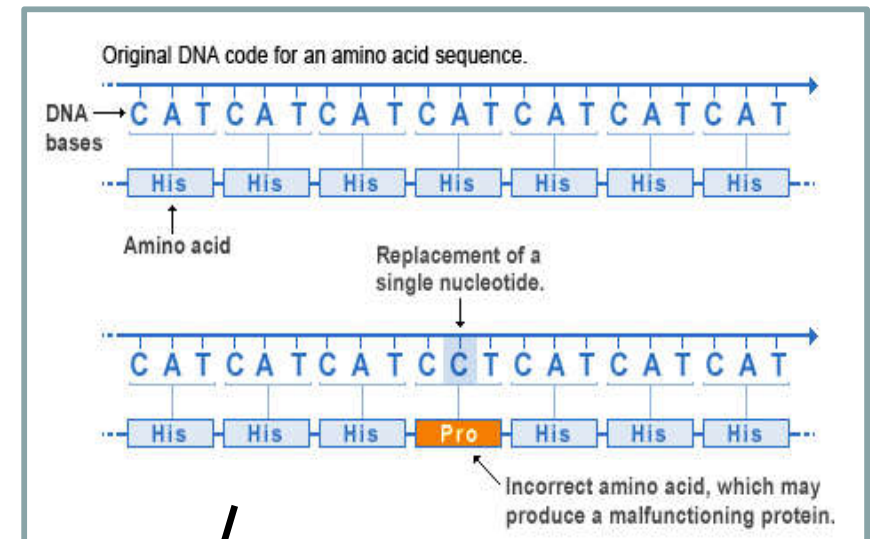
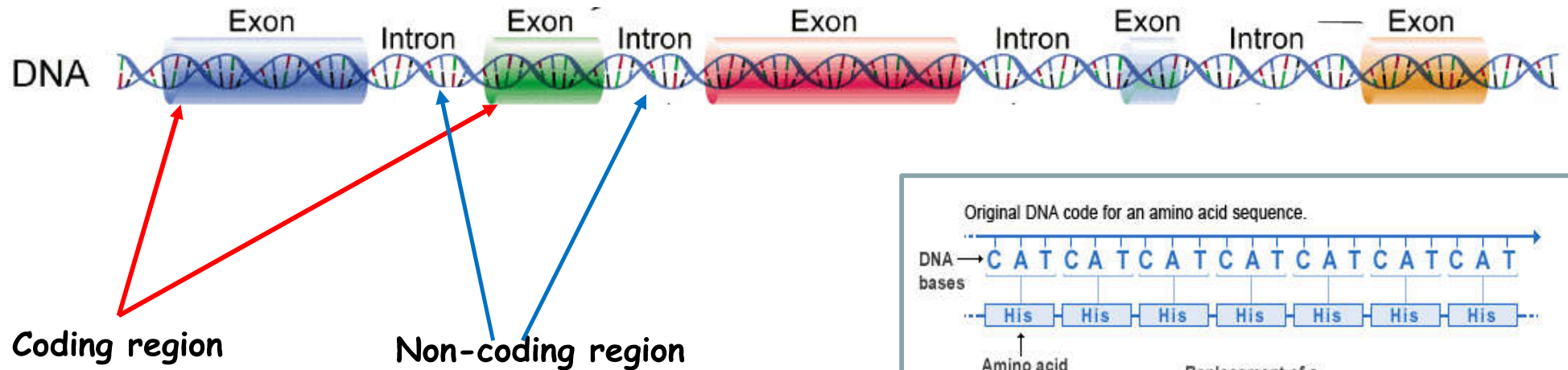
1. Protein Secondary Structure Prediction
2. Protein Contact Prediction
3. Disease-Associated Mutation Prediction

Single-nucleotide polymorphism (SNP)

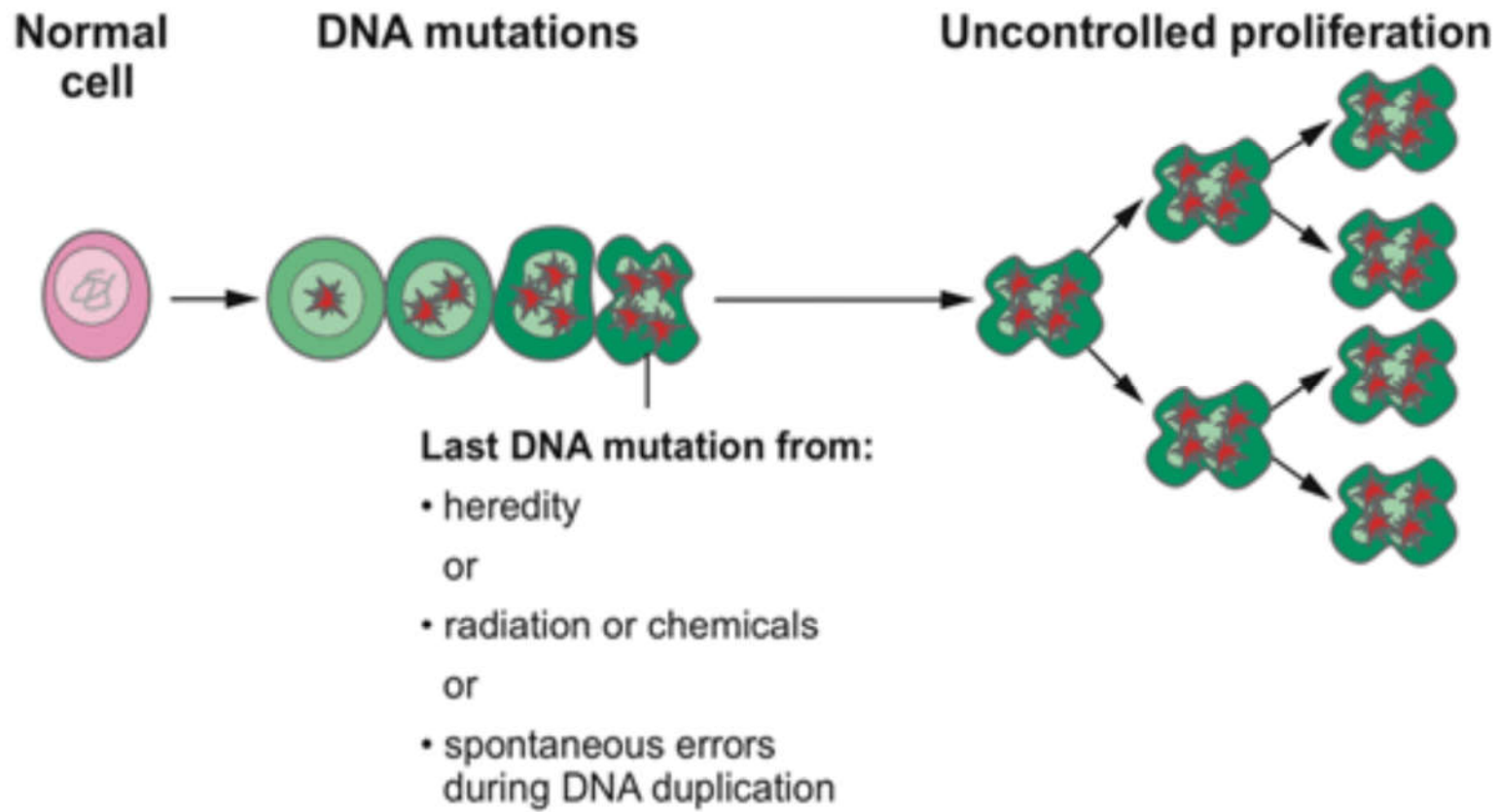
Genome evolution is mainly driven by SNP mutations



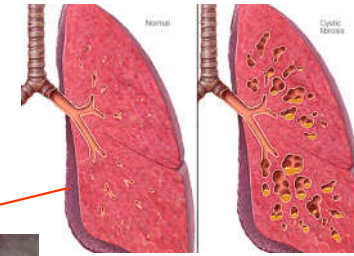
Types of SNP mutations



Cancer Arises From DNA Mutations in Cells



Genetic Diseases



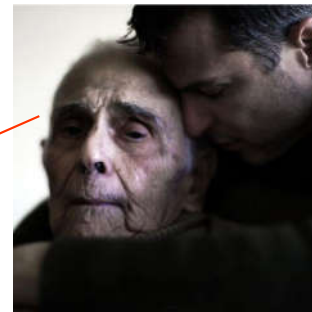
More than 6,000 diseases are due to SNP mutations

- cystic fibrosis (囊胞性纤维症)
- sickle cell anemia (镰状细胞性贫血)
- Marfan syndrome (马方综合征)
- Huntington's disease (亨廷顿氏舞蹈病)
- Hemochromatosis (血色沉着病)



Some serious diseases due to mutation on multiple genes:

- heart disease (心脏病)
- high blood pressure (高血压)
- Alzheimer's disease (老年性痴呆)
- Arthritis (关节炎)
- Diabetes (糖尿病)
- Obesity (肥胖)
- Cancer (癌症)



How to predict what mutations could cause diseases and what could not?

PreDAM: predicting disease-associated mutations based on machine learning



<http://zhanglab.ccmb.med.umich.edu/PreDAM/>

PreDAM: Feature design and extraction

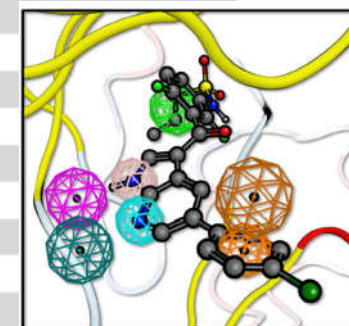
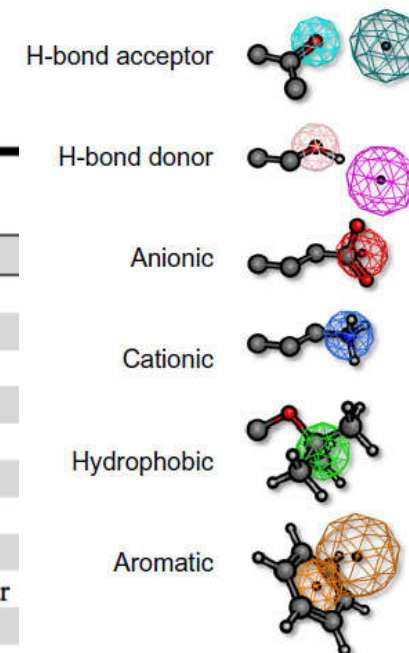
Group-I: Physicochemical properties (Pharmacophore)

Feature Class	No.	Feature	Mean		MCC		p-value (M-W Test) ^c	Description
			DM ^a	NM ^b	cutoff	value		
Physico-chemical properties	Pharmacophore for the wild-type residues							
	1	HP _w	2.909	2.23	2	0.14	6.17E-53	Hydrophobic
	2	noHP _w	4.479	3.98	6	0.11	2.11E-25	Non hydrophobic
	3	AR _w	1.289	0.99	2	0.11	1.13E-29	Aromatic rings
	4	noAR _w	6.09	5.23	6	0.15	1.60E-63	Non aromatic rings
	5	PC _w	0.869	0.80	3	0.03	3.30E-3	Positive charge
	6	NC _w	0.72	0.76	4	0.05	1.93E-2	Negative charge
	7	noC _w	5.79	4.66	5	0.19	1.12E-89	Neutral charge
	8	BP _w	1.67	1.27	2	0.08	6.73E-9	Both wild-type and neighbor AA are polar
	9	OP _w	2.83	2.29	3	0.11	2.33E-39	Either of wild-type and neighbor AA is polar
	10	NP _w	1.87	1.65	5	0.08	2.69E-3	Both wild-type and neighbor are nonpolar t
	11	AC _w	9.12	8.05	11	0.11	1.16E-18	The count of reside being the hydrogen accep...
	12	DO _w	5.59	4.86	8	0.13	1.73E-25	The count of reside being the hydrogen donor
	Pharmacophore for the mutant residues							
	13	HP _m	3.04	2.30	3	0.16	8.54E-62	Hydrophobic
	14	noHP _m	4.52	3.79	5	0.16	4.78E-56	Non hydrophobic
	15	AR _m	1.33	1.03	1	0.10	2.96E-29	Aromatic rings
	16	noAR _m	6.23	5.06	7	0.19	3.5E-104	Non aromatic rings
	17	PC _m	0.83	0.68	2	0.07	2.63E-11	Positive charge
	18	NC _m	0.69	0.69	4	0.03	1.88E-1	Negative charge
	19	noC _m	6.041	4.72	5	0.19	1.7E-105	Neutral charge
	20	BP _m	1.38	1.22	4	0.05	1.1E-1	Both wild-type and neighbor AA are polar
	21	OP _m	3.22	2.37	5	0.17	8.09E-73	Either of wild-type and neighbor AA is polar
	22	NP _m	1.95	1.50	5	0.12	1.75E-14	Both wild-type and neighbor are nonpolar t
	23	AC _m	9.31	7.84	15	0.14	9.92E-30	The count of reside being the hydrogen acceptor
	24	DO _m	5.69	4.79	7	0.14	7.39E-32	The count of reside being the hydrogen donor

↑
Disease
associated
mutations

↑
Neutral
mutations

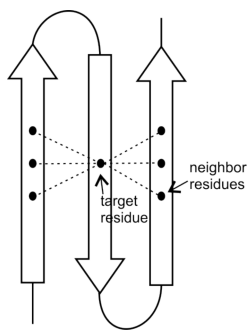
↑
P-value in
Mann-
White test



PreDAM: Feature design and extraction

Group-I: Physicochemical properties (contact environments)

Feature Class	No.	Feature	Mean		MCC		p-value (M-W Test) ^c	Description
			DM ^a	NM ^b	cutoff	value		
Physico-chemical properties	<i>Mutation-induced environmental pharmacophore changes</i>							
	25	cos _{WM}	0.74	0.80	0.89	0.17	5E-66	The cosin for the pharmacophores of wild-type and mutant residues
	26	rms _{WM}	0.47	0.39	0.41	0.17	1.14E-61	The RMSD for the pharmacophores of wild-type and mutant residues
	27	cosN _{WM}	0.94	0.95	0.97	0.06	1.25E-07	The cosin for the neighbor pharmacophores of wild-type and mutant residues
	28	rmsN _{WM}	2.09	1.58	1.68	0.19	6.91E-98	The RMSD for the neighbor pharmacophores of wild-type and mutant residues
	29	cosNS _{WM}	0.92	0.93	1.00	0.08	5.48E-11	The cosin for the neighbor pharmacophores of wild-type and mutant residues related with single residue
	30	rmsNS _{WM}	1.77	1.30	1.84	0.21	1.2E-110	The RMSD for the neighbor pharmacophores of wild-type and mutant residues related with single residue
	31	cosNP _{WM}	0.92	0.93	1.00	0.04	2.08E-1	The cosin for the neighbor pharmacophores of wild-type and mutant residues related with residue paired
	32	rmsNP _{WM}	2.87	2.27	4.24	0.12	6.87E-28	The RMSDfor the neighbor pharmacophores of wild-type and mutant residues related with residue paired
	Other physicochemical properties							
	33	Vol _w	2.83	2.86	1.90	0.09	4.51E-1	The volume of wild-type residue
	34	Vol _m	2.91	2.89	3.16	0.09	1.86E-3	The volume of mutant residue
	35	dVol	0.08	0.03	0.65	0.13	2.13E-05	The volume difference
	36	W _w	132.01	131.84	75.07	0.09	0.19646	The weight of wild-type residue
	37	W _m	136.24	133.09	165.19	0.09	8.7E-05	The weight of mutant residue
	38	dW	4.23	1.26	42.08	0.15	4.78E-4	The molecular weight difference



$$\begin{aligned} \overrightarrow{P_w(i)} &= \sum_{k=1}^{n_{con}^w(i)} \overrightarrow{p_w(k)} = \{P_w^1(i), \dots, P_w^L(i)\} \\ \overrightarrow{P_m(i)} &= \sum_{k=1}^{n_{con}^m(i)} \overrightarrow{p_m(k)} = \{P_m^1(i), \dots, P_m^L(i)\} \end{aligned}$$

$$\begin{cases} \cos(\overrightarrow{P_w(i)}, \overrightarrow{P_m(i)}) = \frac{\overrightarrow{P_w(i)} \cdot \overrightarrow{P_m(i)}}{\|\overrightarrow{P_w(i)}\| \cdot \|\overrightarrow{P_m(i)}\|} \\ rmsd(\overrightarrow{P_w(i)}, \overrightarrow{P_m(i)}) = \sqrt{\frac{\sum_{l=1}^L (P_w^l(i) - P_m^l(i))^2}{L}} \end{cases}$$

PreDAM: Feature design and extraction

Group-II: Evolutionary profiles (PSIBlast, LOMETS, Pfam families):

Feature Class	No.	Feature	Mean		MCC		p-value (M-W Test) ^c	Description
			DM ^a	NM ^b	cutoff	value		
Evolutionary profiles	PSI-BLAST profile scores							
	39	PSIC _w	1.567	0.99	1.33	0.37	0	The PSIC score for wild-type residue
	40	PSIC _m	-0.42	0.17	-0.11	0.39	0	The PSIC score for mutant residue
	41	dPSIC	-1.99	-0.82	-1.10	0.48	0	The PSIC score difference
	42	JSD _w	0.03	0.03	0.04	0.10	4.57E-1	The JSD score for wild-type residue
	43	JSD _m	0.02	0.02	0.03	0.09	2.74E-07	The JSD score for mutant residue
	44	dJSD	-0.00	-0.00	-0.01	0.08	3.19E-06	The JSD score difference
	45	JSD _i	0.47	0.33	0.46	0.29	2.4E-200	The JSD score at mutant position <i>i</i>
	LOMETs profile scores							
	46	tPSIC _w	0.78	0.46	0.74	0.20	7.5E-118	The PSIC score for wild-type residue
	47	tPSIC _m	-0.30	-0.08	0.04	0.15	1.2E-58	The PSIC score for mutant residue
	48	dtPSIC	-1.08	-0.54	-0.67	0.25	1.1E-159	The PSIC score difference
	Pfam profile scores							
	49	Pfam _w	1.83	2.35	1.49	0.29	1.1E-124	The Pfam score for wild-type residue
	50	Pfam _m	3.66	3.10	3.12	0.25	2E-119	The Pfam score for mutant residue
	51	dPfam	1.83	0.75	1.12	0.32	2.6E-190	The Pfam score difference

Jensen-Shannon divergence

Q5E940_BOVIN -----MREDRATWSSYFLKIIQLDDPKCFIVGAMVGRKQHQDIHSLRGK-AVVLNGKTHHKAIRGSLNN--PAL
 RL00_HUMAN -----MREDRATWSSYFLKIIQLDDPKCFIVGAMVGRKQHQDIHSLRGK-AVVLNGKTHHKAIRGSLNN--PAL
 RL00_MOUSE -----MREDRATWSSYFLKIIQLDDPKCFIVGAMVGRKQHQDIHSLRGK-AVVLNGKTHHKAIRGSLNN--PAL
 RL00_RAT -----MREDRATWSSYFLKIIQLDDPKCFIVGAMVGRKQHQDIHSLRGK-AVVLNGKTHHKAIRGSLNN--PAL
 RL00_CHICK -----MREDRATWSSYFLKIIQLDDPKCFIVGAMVGRKQHQDIHSLRGK-AVVLNGKTHHKAIRGSLNN--PAL
 RL00_RABBIT -----MREDRATWSSYFLKIIQLDDPKCFIVGAMVGRKQHQDIHSLRGK-AVVLNGKTHHKAIRGSLNN--PAL
 Q7ZUG3_BRASS -----MREDRATWSSYFLKIIQLDDPKCFIVGAMVGRKQHQDIHSLRGK-AVVLNGKTHHKAIRGSLNN--PAL
 RL00_ICPPO -----MREDRATWSSYFLKIIQLDDPKCFIVGAMVGRKQHQDIHSLRGK-AVVLNGKTHHKAIRGSLNN--PAL
 RL00_DROME -----MYREKAKWAGYFKIVFLDDEPKCFIVGAMVGRKQHQDIHSLRGK-AVVLNGKTHHKAIRGSLNN--PAL
 RL00_DICD1 -----MSLALSKKKLFLKKTKLFTTDMKIVAGAVVGRSGLKIKKSTGI-GAVLNGKTHHKAIRGSLNN--PAL
 Q54LP0_DICD1 -----MSLALSKKKLFLKKTKLFTTDMKIVAGAVVGRSGLKIKKSTGI-GAVLNGKTHHKAIRGSLNN--PAL
 RL00_PLAFB -----MAKLSQKQKHVYKELSLIQDQSKILIVYVGRMAVYKSLRGK-AVVLNGKTHHKAIRGSLNN--PAL
 RL00_SULAC -----HILAVITTKKIAHVVGVIAETKTKTKTIIIAIKIGPADLHIEIKSLRGK-ADIKVTHHKAIRGSLNN--PAL
 RL00_SULTO -----HILAVITTKKIAHVVGVIAETKTKTKTIIIAIKIGPADLHIEIKSLRGK-ADIKVTHHKAIRGSLNN--PAL
 RL00_SULSO -----HILAVITTKKIAHVVGVIAETKTKTKTIIIAIKIGPADLHIEIKSLRGK-ADIKVTHHKAIRGSLNN--PAL
 RL00_AERP2 -----HIVVSLVQNHKKKFLPKKTLMLGKELFKSKKAVVLPADLGGKFFVGVYKSLRGK-AVVLNGKTHHKAIRGSLNN--PAL
 RL00_PYBAC -----HMLATKRRYVTRGTHAKKVIKSEKTLQKIVYVLPADLGGKFFVGVYKSLRGK-AVVLNGKTHHKAIRGSLNN--PAL
 RL00_HETAC -----HAEERHTEHIFQKDEKDKELQSKVYVGVVILATKTHHKAIRGSLNN--PAL
 RL00_HETHA -----HAEERHTEHIFQKDEKDKELQSKVYVGVVILATKTHHKAIRGSLNN--PAL
 RL00_ARCFB -----MAVYSGGKPKKAVKEXKMISSVYVGVVILATKTHHKAIRGSLNN--PAL
 RL00_HETKA -----HAYKAGGKPKKAVKEXKMISSVYVGVVILATKTHHKAIRGSLNN--PAL
 RL00_HETTH -----HAYKAGGKPKKAVKEXKMISSVYVGVVILATKTHHKAIRGSLNN--PAL
 RL00_HETTL -----HITAEERKIAHVVGVIAETKTKTKTIIIAIKIGPADLHIEIKSLRGK-ADIKVTHHKAIRGSLNN--PAL
 RL00_HETVA -----HITAEERKIAHVVGVIAETKTKTKTIIIAIKIGPADLHIEIKSLRGK-ADIKVTHHKAIRGSLNN--PAL
 RL00_HETJA -----HETKAVKPKKAVKEXKMISSVYVGVVILATKTHHKAIRGSLNN--PAL



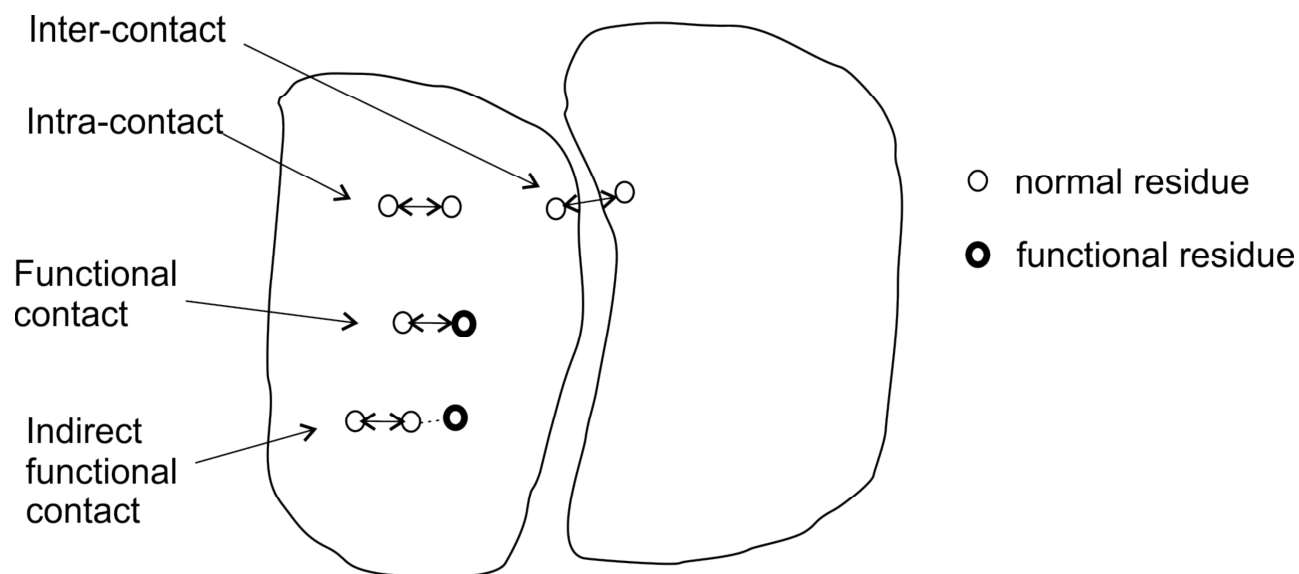
Measuring evolutionary divergence

$$\begin{cases} JSD_{ia} = \lambda p_{ia} \log \frac{p_{ia}}{c_{ia}} + (1 - \lambda) q_a \log \frac{q_a}{c_{ia}} \\ JSD_i = \sum_{a \in AA} JSD_{ia} \end{cases}$$

PreDAM: Feature design and extraction

Group-III: Contact environments with functional residues:

Feature Class	No.	Feature	Mean		MCC		p-value (M-W Test) ^c	Description
			DM ^a	NM ^b	cutoff	value		
Contact environments	Directly contacted residues							
	52	Intra	13.78	11.25	15	0.23	9.5E-137	The number of intramolecular contacts
	53	FunIntra	4.70	3.72	16	0.10	1.35E-15	The number of intramolecular functional contacts
	54	Inter	1.16	0.89	1	0.10	1.96E-15	The number of intermolecular contacts
	55	FunInter	0.30	0.38	5	0.03	3.53E-1	The number of intermolecular functional contacts
	Indirectly contacted residues							
	56	CIntra	57.96	46.18	66	0.24	5E-132	The number of intramolecular indirectly contacts
	57	CFunIntra	18.67	15.00	1	0.09	7.57E-18	The number of intramolecular functional indirectly contacts
	58	CInter	11.54	8.05	11	0.11	8.59E-18	The number of intermolecular indirectly contacts
	59	CFunInter	3.95	3.23	1	0.10	5.39E-13	The number of intermolecular functional indirectly contacts



PreDAM: Feature design and extraction

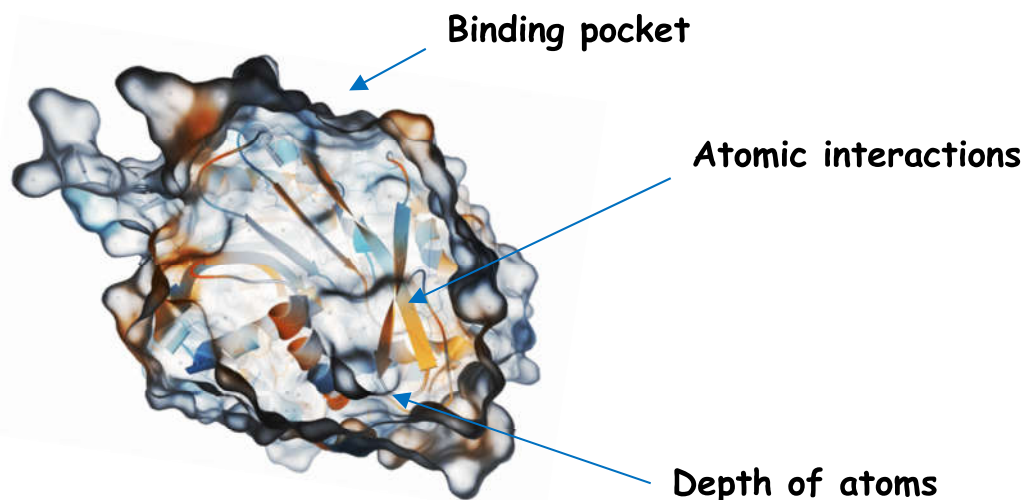
Group-IV: Structure prediction based features:

Feature Class	No.	Feature	Mean		MCC		p-value (M-W Test) ^c	Description
			DM ^a	NM ^b	cutoff	value		
I-TASSER structure models	<i>Protein surface regions favorable for interactions</i>							
	60	CS	0.08	0.04	0.04	0.16	6.46E-56	the ConCavity score for the ligand-binding interaction
	61	Depth	6.72	5.50	5.61	0.24	4.2E-137	The average distance of target atoms/residues to its closest molecule of bulk solvent
	<i>The physics-based energy functions</i>							
	62	ED	649.56	589.01	555.57	0.12	8.93E-18	The EvoDesign score
	63	ddG	1.62	0.62	3.00	0.19	4.27E-61	The free-energy changes upon mutation
	64	VDW _w	-343.44	-322.54	-357.52	0.11	1.39E-9	Van der Waals potential of wild-type residue by CISS-RR
	65	VDW _m	-331.33	-316.17	-351.06	0.09	5.18E-6	Van der Waals potential of mutant residue by CISS-RR
	66	dVDW	12.11	6.37	1.49	0.20	1.43E-96	Van der Waals potential difference
	67	RT _w	460.13	424.66	425.50	0.08	1.75E-11	Rotamer preferences of side-chain conformers by CISS-RR.
	68	CISRR _w	116.69	102.12	55.05	0.10	2.78E-13	CIS-RR score for wild-type residue
	69	CISRR _m	129.06	108.74	67.00	0.13	6.98E-25	CIS-RR score for mutant residue
70	dCISRR	12.37	6.61	4.52	0.17	1.18E-72	CIS-RR score difference	

sequence

FTVSNTNNEFVLISDP
TGGKSIGLLCFRQED
AEAFLAQARLRREL
KTNAAKVPITLDQVYL
LKVEGISFRFLPDPI

I-TASSER



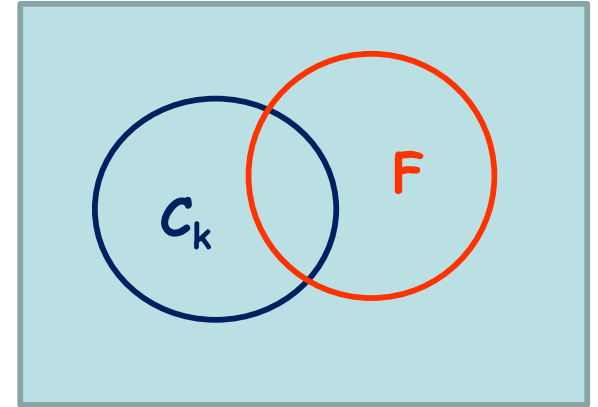
BANN: A new method for NN training on posterior probability

Given:

C_D : disease-associate mutation

C_N : neutral mutation without causing disease

$F=(f_1, f_2, \dots, f_n)$: n specific features



Naïve Bayes classifier theorem:

$$P(C_k|f_1, f_2, \dots, f_n) = P(C_k|F) = \frac{P(C_k)P(F|C_k)}{P(F)}$$

$$P(C_k|F) \propto P(C_k)P(F|C_k) = P(C_k)P(f_1|f_2, \dots, f_n, C_k)P(f_2|f_3, \dots, f_n, C_k) \dots P(f_n|C_k)$$

Under naïve assumption (ie all features are independent): $P(f_i|f_{i+1}, \dots, f_n, C_k) = P(f_i|C_k)$

$$P(C_k|F) \propto P(C_k) \prod_{i=1}^n P(f_i|C_k)$$

$$\log P(C_k|F) \propto \sum_{i=1}^n \log P(f_i|C_k) + \log P(C_k)$$

BANN: A new method for NN training on posterior probability

General form of Bayes classifier for mutation classes:

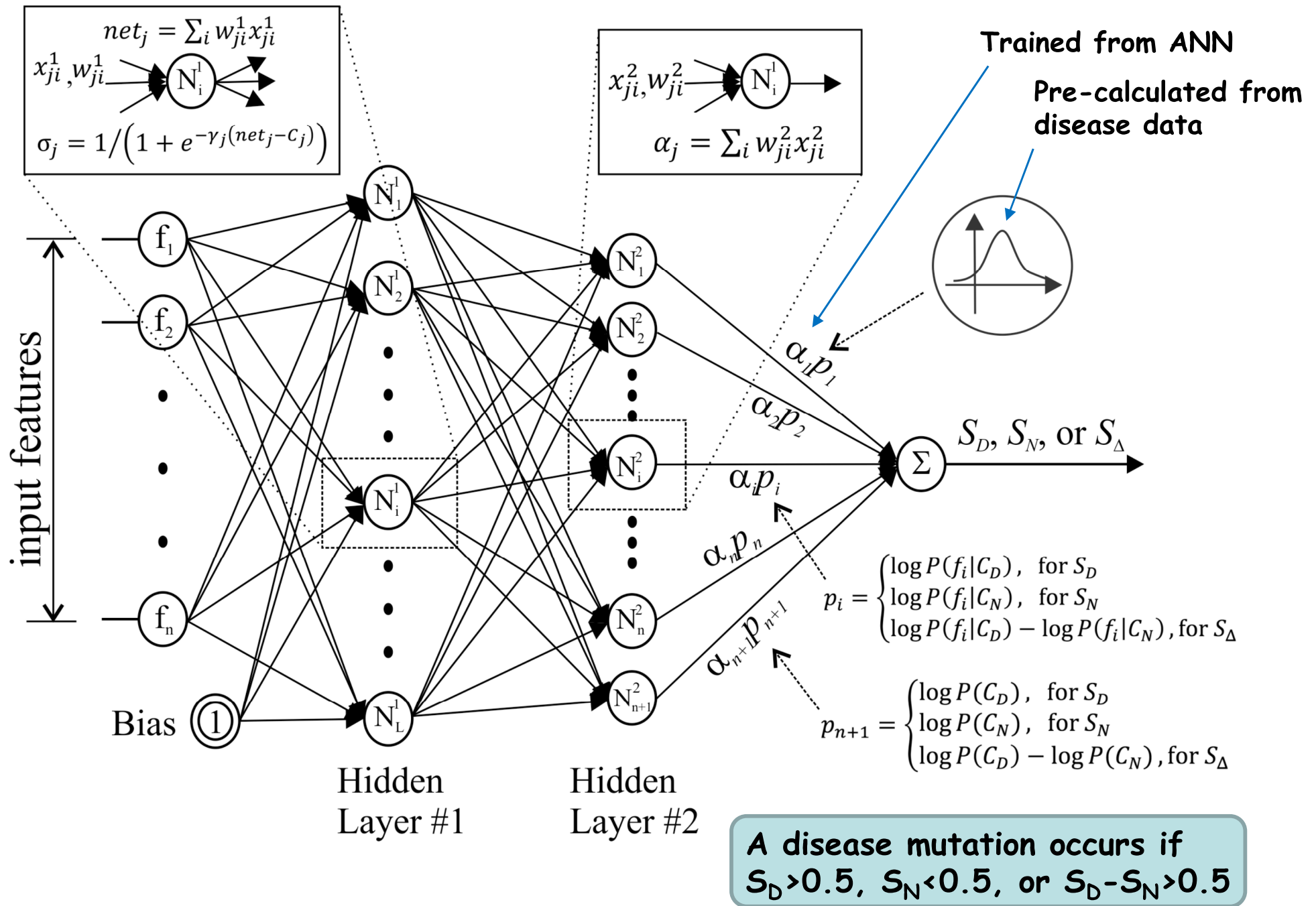
$$\log P(C_k|F) = \sum_{i=1}^n \alpha_i(F) \log P(f_i|C_k) + \alpha_{n+1}(F) \log P(C_k)$$

$$\begin{cases} S_D = \log P(C_D|F) = \sum_{i=1}^n \alpha_i(F) \log P(f_i|C_D) + \alpha_{n+1}(F) \log P(C_D) \\ S_N = \log P(C_N|F) = \sum_{i=1}^n \alpha_i(F) \log P(f_i|C_N) + \alpha_{n+1}(F) \log P(C_N) \end{cases}$$

$$S_\Delta = S_D - S_N =$$

$$\sum_{i=1}^n \alpha_i(F) [\log P(f_i|C_D) - \log P(f_i|C_N)] + \alpha_{n+1}(F) [\log P(C_D) - \log P(C_N)].$$

A new method for NN training on posterior probability



ANN: back propagation training

Error minimization

$$E(\vec{w}, \vec{\gamma}, \vec{C}) \equiv \frac{1}{2N_d} \sum_{d=1}^{N_d} (t_d - o_d)^2 + \mu \sum_{i,j} w_{ji}^2$$

Gradient descent training rules

$$\begin{cases} \Delta w_{ji,2} = \eta \delta_{j,2} x_{ji} - 2\eta \mu w_{ji} \\ \Delta w_{ji,1} = \eta \delta_{j,1} x_{ji} - 2\eta \mu w_{ji} \\ \Delta \gamma_j = \eta \delta_{j,1} (net_j - C_j) \\ \Delta C_j = \eta \delta_{j,1} \gamma_j \end{cases}$$

$\begin{cases} \delta_{j,2} = P_j(t_d - o_d) \\ \delta_{j,1} = \gamma_j o_j (1 - o_j) \sum_{k \in \text{Down}(j)} \delta_{k,2} w_{kj} \end{cases}$

learning rate

Benchmark results

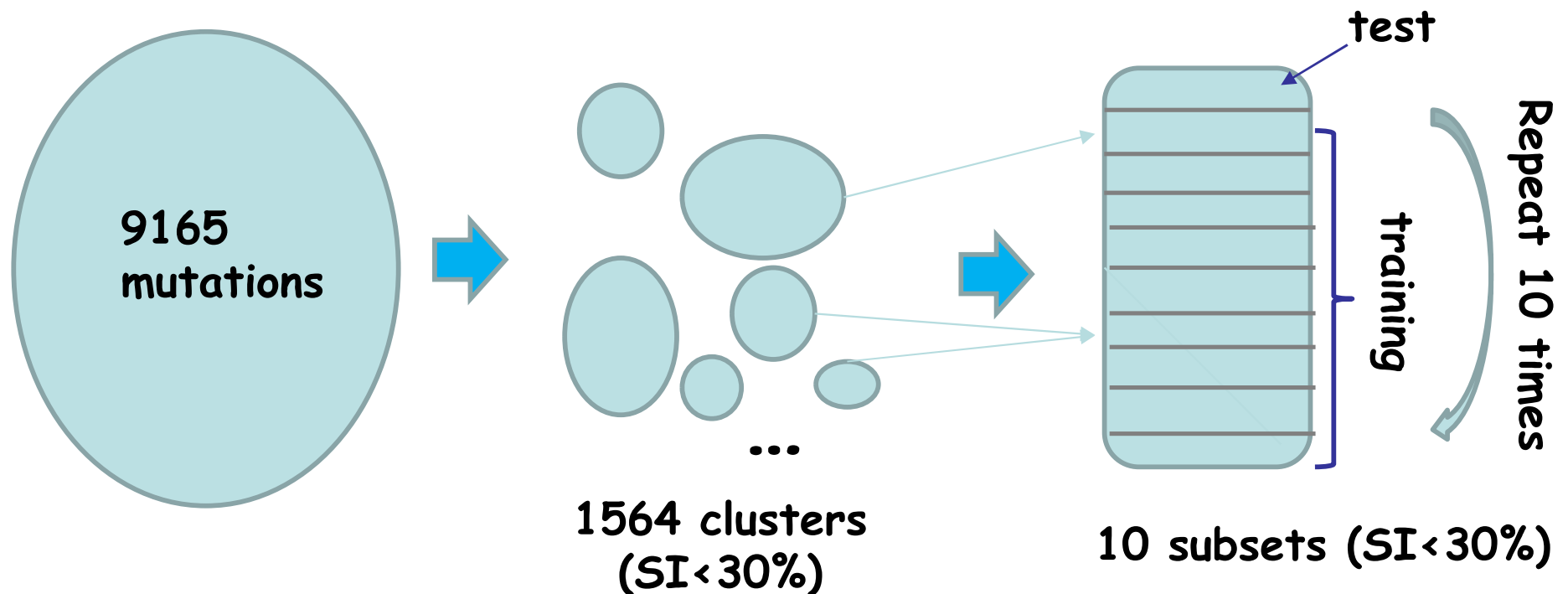
Data sets

Disease-associated mutations: 5,356 SNP mutations in 635 proteins

Neural mutations: 3,809 SNP mutations in 1,645 proteins

Total: 9,165 mutations in 1,974 proteins

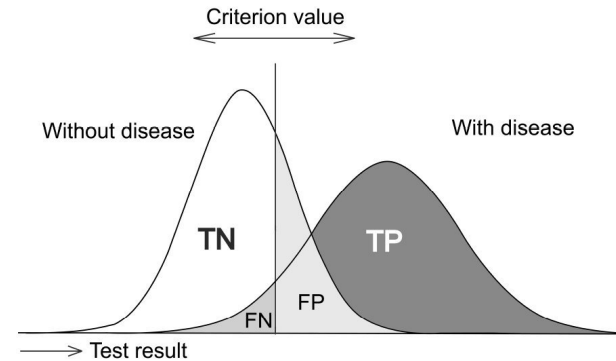
10-fold Cross validation procedure



Benchmark results

Results on different training methods

$$\left\{ \begin{array}{l} MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \\ ACC = \frac{TP + TN}{TP + FN + TN + FP} \\ Sen(+) = \frac{TP}{TP + FN}, \quad Spe(+) = \frac{TP}{TP + FP} \\ Sen(-) = \frac{TN}{TN + FP}, \quad Spe(-) = \frac{TN}{TN + FN} \end{array} \right.$$



	Top 20 features ^a				All features ^a			
	GBC ^b	KNC ^b	SVC ^b	BANN ^b	GBC ^b	KNC ^b	SVC ^b	BANN ^b
MCC	0.48	0.49	0.50	0.52	0.47	0.44	0.50	0.53
ACC	0.75	0.75	0.76	0.77	0.74	0.72	0.76	0.77
SEN ⁺	0.81	0.78	0.80	0.82	0.79	0.73	0.78	0.82
SPE ⁺	0.77	0.79	0.79	0.79	0.77	0.79	0.79	0.80
SEN ⁻	0.67	0.71	0.69	0.69	0.68	0.72	0.70	0.71
SPE ⁻	0.71	0.69	0.72	0.74	0.70	0.65	0.71	0.73

^aRank of features in S1 Table by the p-value. There are two groups: Top20 and All features. Top 20 are the first 20 lowest features.

^bGBC: gradient boosting classifier; KNC: k-nearest neighbor classifier; SVC: support vector classifier; BANN: Bayes-classifier guided ANN (BANN).

BANN is more efficient than other machine-learning methods

Benchmark results

Comparison of PreDAM with other predictors:

Method	MCC	ACC	Positive		Negative	
			Sensitivity	Specificity	Sensitivity	Specificity
SIFT	0.46	0.74	0.92	0.71	0.49	0.81
SNAP2	0.43	0.75	0.87	0.72	0.52	0.75
PolyPhen2	0.47	0.75	0.89	0.73	0.55	0.78
SNAP2+BANN ^a	0.52	0.77	0.84	0.79	0.69	0.75
PolyPhen2+BANN ^b	0.51	0.76	0.84	0.78	0.67	0.75
PreDAM	0.53	0.77	0.82	0.80	0.71	0.73

^aSNAP2+BANN: the twenty of features extracted from SNAP2 with the lowest p-value. The training method used BANN

^bPolyPhen2+BANN: the twenty of features extracted from PolyPhen2 with the lowest p-value. The training method used BANN

- PreDAM output other predictors
- BANN+control > control methods, indicating again BANN is more efficient as machine-learning
- PreDAM > BANN+control, indicating advantage of feature selection in PreDAM

Conclusions

1. Machine learning is an efficient technique to predict disease-associated SNP mutations
2. Bayes-guided neural-network (BANN) training has a higher efficiency than other classifiers and ANN training methods
3. Structure based features can improve the accuracy of disease mutation prediction accuracy