

# 生物信息学：导论与方法

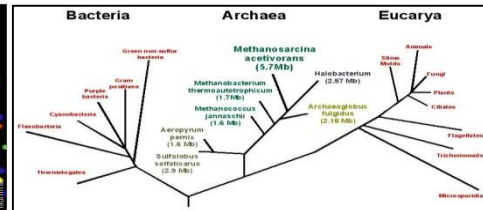
## Bioinformatics: Introduction and Methods

Ge Gao 高歌 & Liping Wei 魏丽萍

Center for Bioinformatics, Peking University



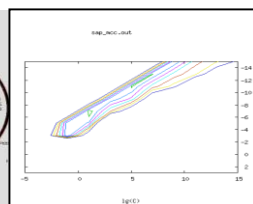
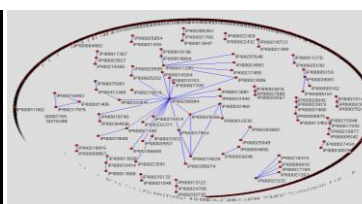
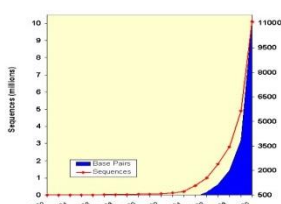
<https://www.coursera.org/course/pkubioinfo>

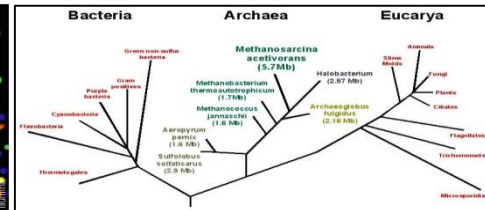


北京大学生物信息学中心 魏丽萍

# Liping Wei, Ph.D.

# Center for Bioinformatics, Peking University

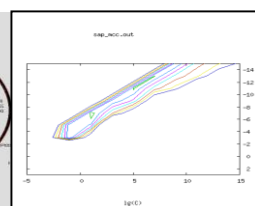
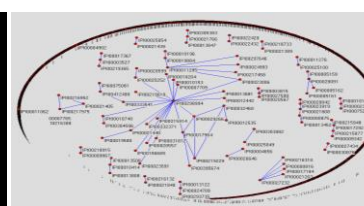
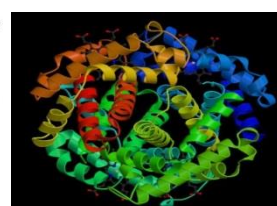
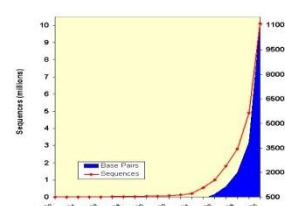




北京大学生物信息学中心 魏丽萍

# Liping Wei, Ph.D.

# Center for Bioinformatics, Peking University

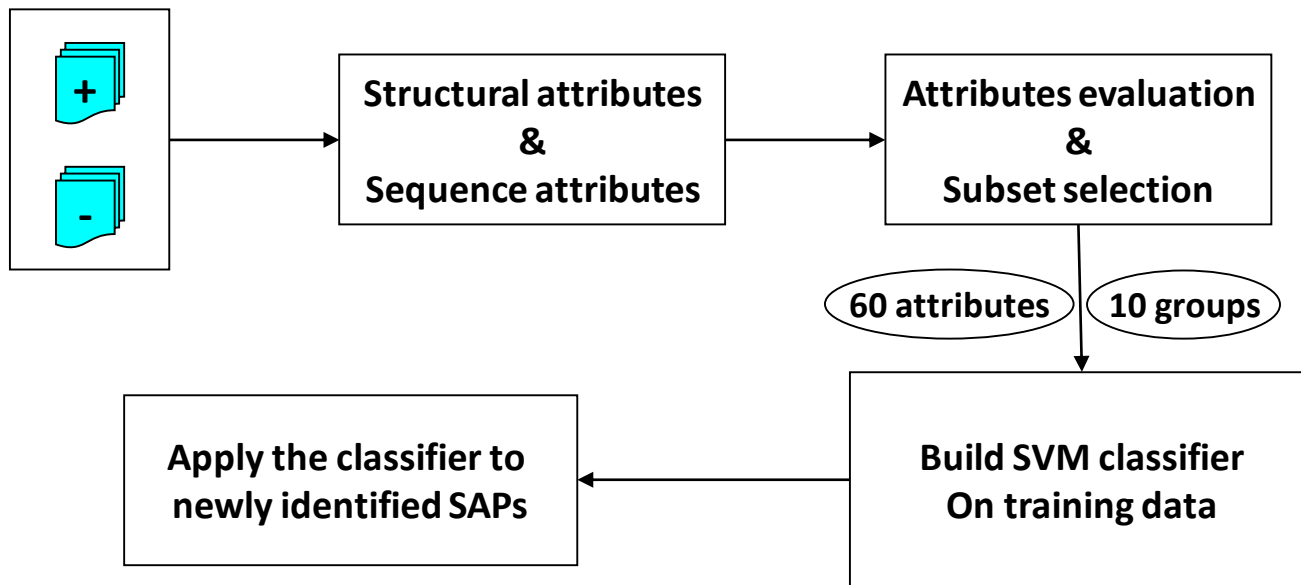


# Single Amino acid Polymorphisms disease-association Predictor (SAPRED)

[www.sapred.cbi.pku.edu.cn](http://www.sapred.cbi.pku.edu.cn)



## Formulate as a supervised classification problem



# PDB – get protein 3D structure

■ <http://www.rcsb.org/pdb/home/home.do>

The screenshot displays the RCSB PDB website. At the top left is the RCSB PDB logo with the text 'PROTEIN DATA BANK'. To its right is a 'PDB-101' badge. Further right, it states 'A MEMBER OF THE PDB EMDatabank' and 'An Information Portal to Biological Macromolecular Structures'. Below this, a timestamp reads 'As of Tuesday May 21, 2013 at 5 PM PDT there are 90810 Structures | PDB Statistics' followed by social media icons. The main navigation area includes a 'Search' button, 'Advanced', and 'Browse' links. A search bar contains the text 'Everything' and a list of filters: 'Author', 'Macromolecule', 'Sequence', 'Ligand', and a help icon. Below the search bar is a text input field with the placeholder 'e.g., PDB ID, molecule name, author' and a search icon. Below the search bar is a link to 'Search History, Previous Results'. The 'Explore Archive' section is expanded, showing a grid of category buttons: 'Organism', 'Taxonomy', 'Exp. Method', 'X-ray Resolution', 'Release Date', 'Polymer Type', 'Enzyme Classification', 'SCOP Classification', 'Protein Symmetry', and 'Protein Stoichiometry'. A 'Show all' link is at the bottom of the grid. To the right of the grid, the 'Organism' category is selected, displaying a list of organisms with their corresponding PDB IDs: Homo sapiens (22611), Escherichia coli (4724), Mus musculus (3874), Saccharomyces cerevisiae (2247), Bos taurus (2145), Rattus norvegicus (1896), Escherichia coli K-12 (1637), and Other (49129).

RCSB PDB PROTEIN DATA BANK

PDB-101

A MEMBER OF THE PDB EMDatabank

An Information Portal to Biological Macromolecular Structures

As of Tuesday May 21, 2013 at 5 PM PDT there are 90810 Structures | PDB Statistics

Search

Advanced

Browse

Everything Author Macromolecule Sequence Ligand ?

e.g., PDB ID, molecule name, author

Search History, Previous Results

↑ Explore Archive Hide

Organism

Taxonomy

Exp. Method

X-ray Resolution

Release Date

Polymer Type

Enzyme Classification

SCOP Classification

Protein Symmetry

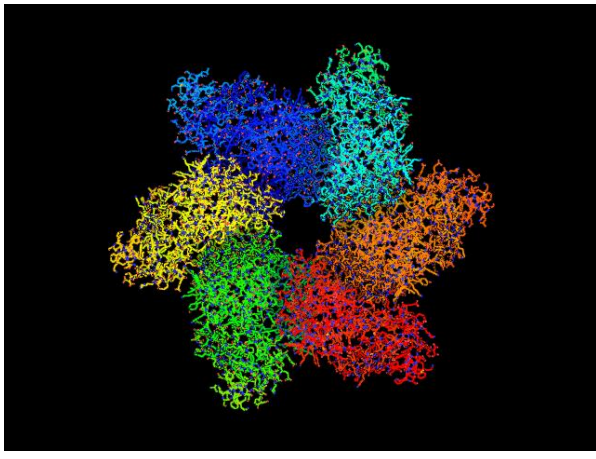
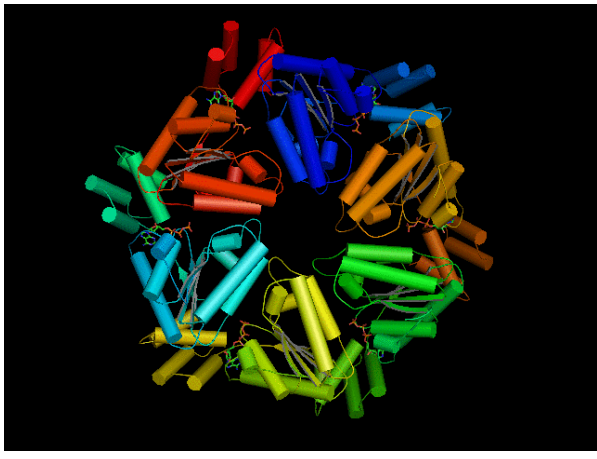
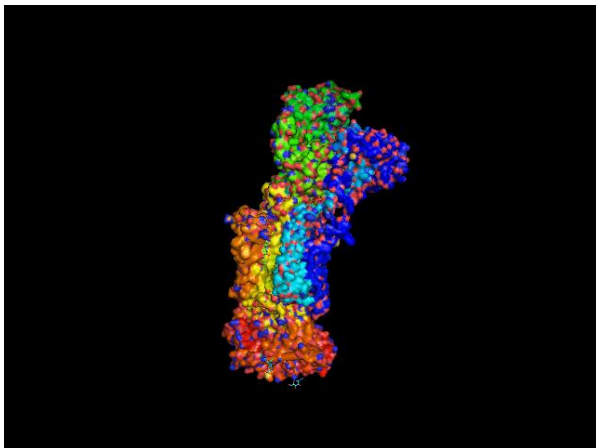
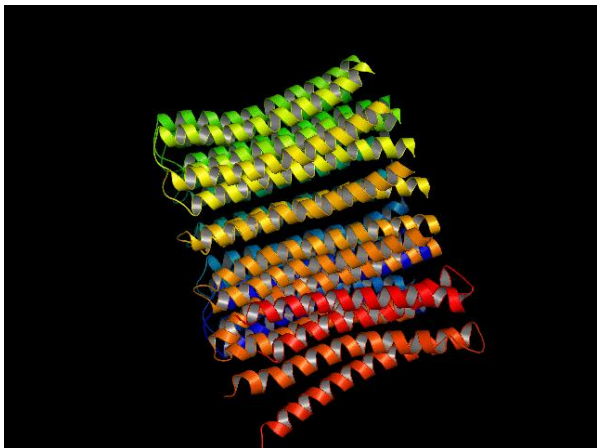
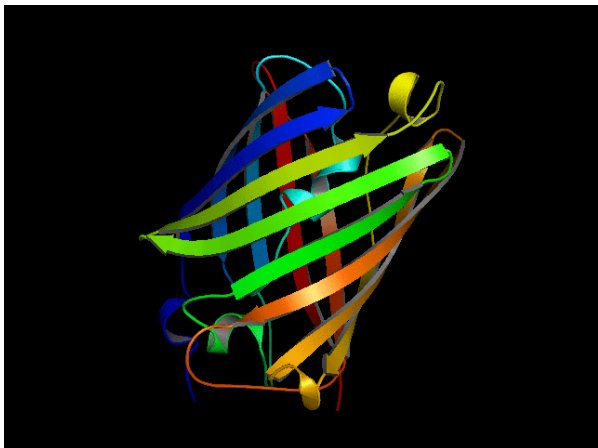
Protein Stoichiometry

Show all

Organism

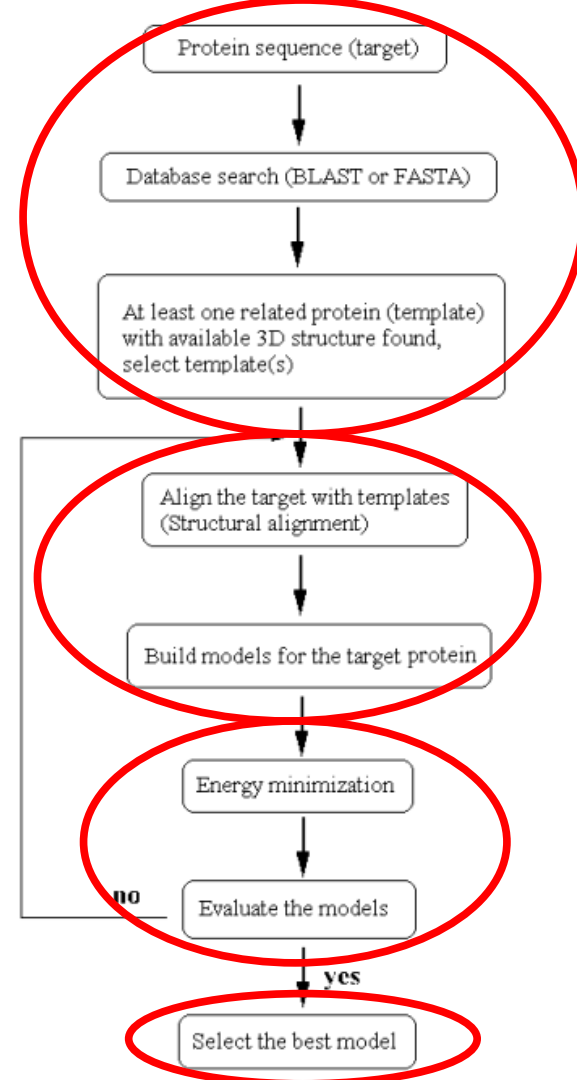
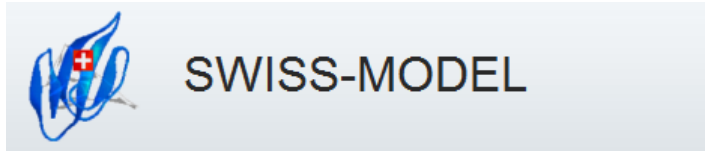
- Homo sapiens (22611)
- Escherichia coli (4724)
- Mus musculus (3874)
- Saccharomyces cerevisiae (2247)
- Bos taurus (2145)
- Rattus norvegicus (1896)
- Escherichia coli K-12 (1637)
- Other (49129)





# Homology Modeling

■ <http://swissmodel.expasy.org/>

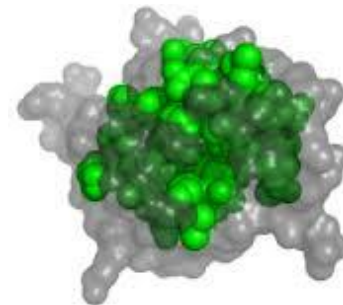
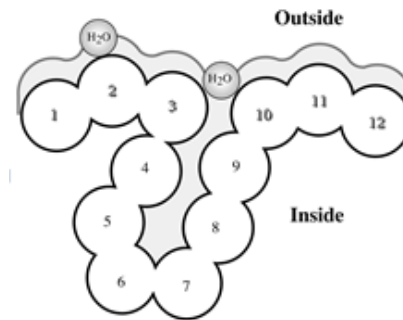


# Biologically-Intuitive features

- Residue frequencies, conservation scores
- Solvent accessibilities and  $C_{\beta}$  density, secondary structure...

## New attributes:

- Structural neighbor profile
- Nearby functional sites
- Disordered regions
- Hydrogen bonds change
- $\beta$ -aggregation
- HLA family





# Structural neighbor profile

## ■ Definition:

- A 20-D vector: take the  $C_\alpha$  of the variant residue as the center, draw a sphere with a specific radius. The residues inside are counted to get the number for each of the 20 kinds of residues. Each number is a component of the vector.

$$\left[ N_{R,a_i} \right] = \sum_{j=1}^L X_j$$

where  $X_j=1$  if  $X_j = a_i$  &  $r_{X_j,c} < R$ ;

otherwise,  $X_j = 0$

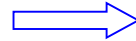
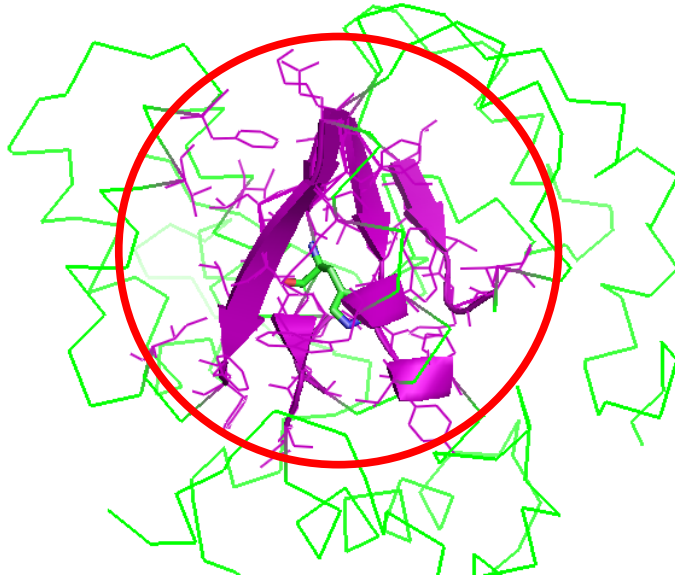
R: radius

L: protein length

$a_i$ : a specific residue type

r: distance between a  
residue and the center  
residue

# Structural neighbor profile



The variant site is H128,  
radius is 10 Angstroms.

Neighbors are:

42-47: LLICTY

50-52: AGT

55: I

59: V

106-110: LKTHL

112: T

125-127: KFL

129-131: VAR

176-177: HV

180-181: WW

184: K

188-194: QILFLFY

197: I

208: V

211: F

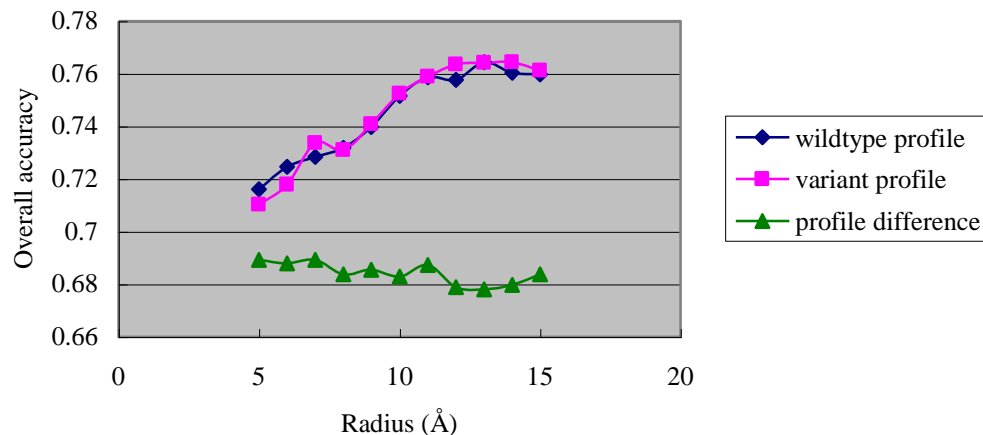
# Structural neighbor profile: vector

<b><u>a.a.</u></b>	A	C	D	E	F	G	H	I	K	L
<b><u>N</u></b>	2	1	0	0	4	1	2	4	3	7

<b><u>a.a.</u></b>	M	N	P	Q	R	S	T	V	W	Y
<b><u>N</u></b>	0	0	0	1	1	0	4	4	2	2

# Structural neighbor profile

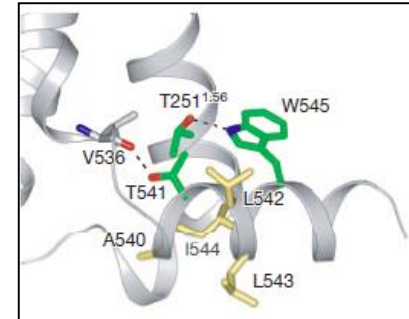
Predictive power of different structural neighbor profile



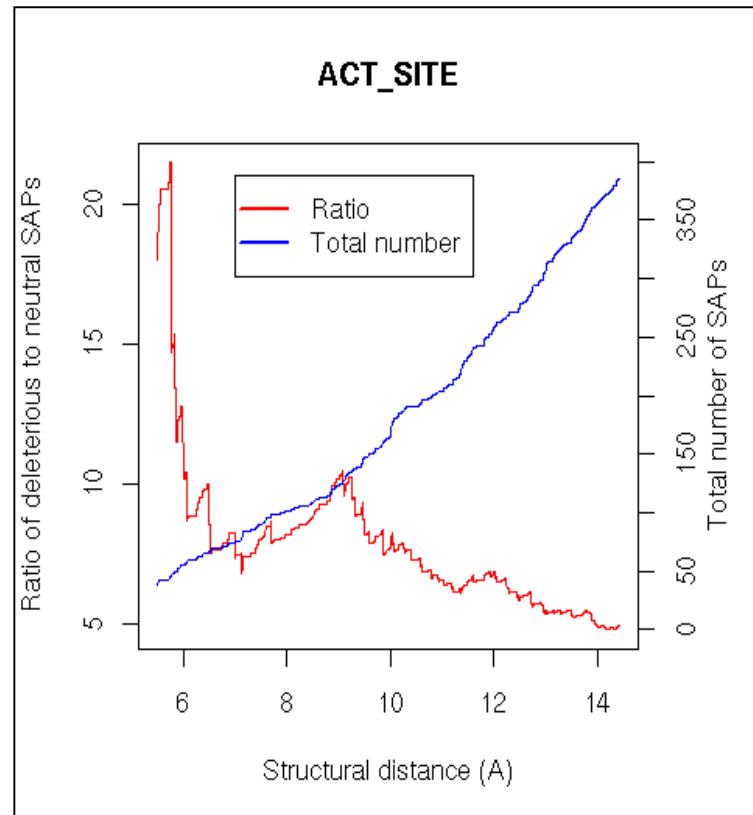
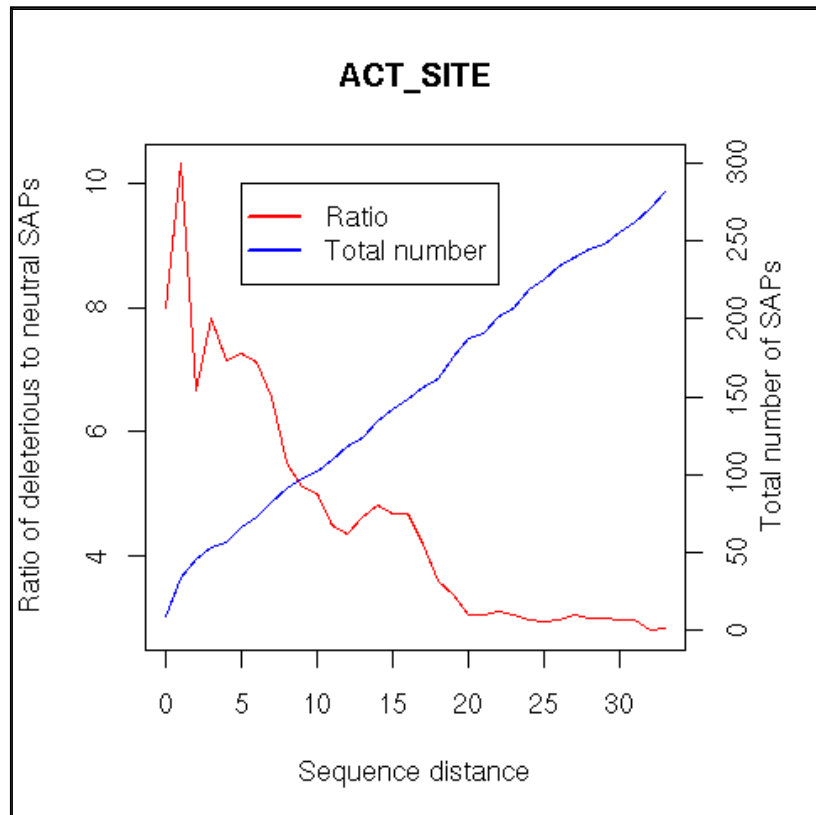
- Different radius had different predictive power.
- We selected 13 Angstroms as the optimal radius.

## Nearby functional sites

- Amino acid variations located exactly on functional, active, and binding sites tend to have large effect on protein function.
  - But coverage is low.
- We considered that variations **in the vicinity of important sites** could also affect protein function.
  - Significantly increased coverage.

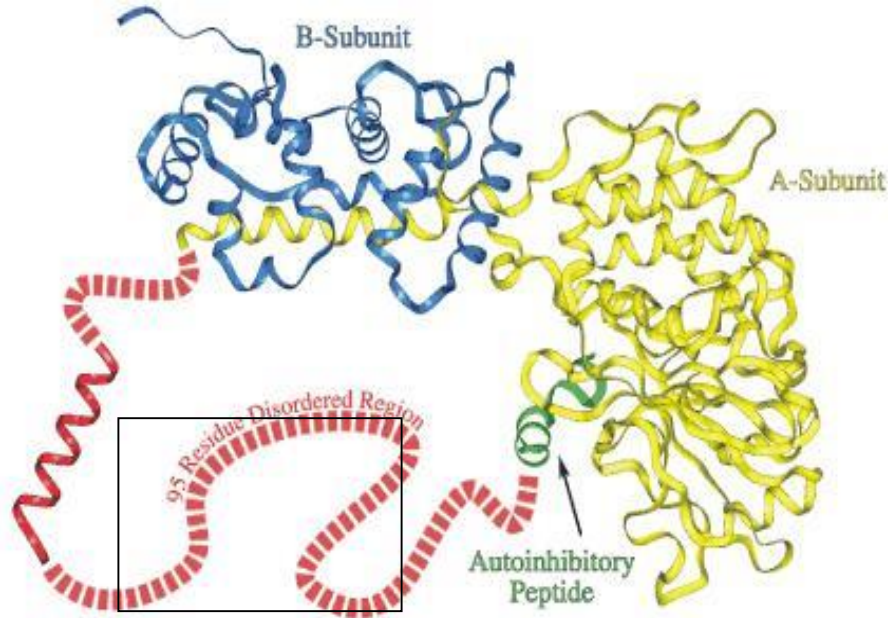


# Nearby functional sites





# Disordered Region

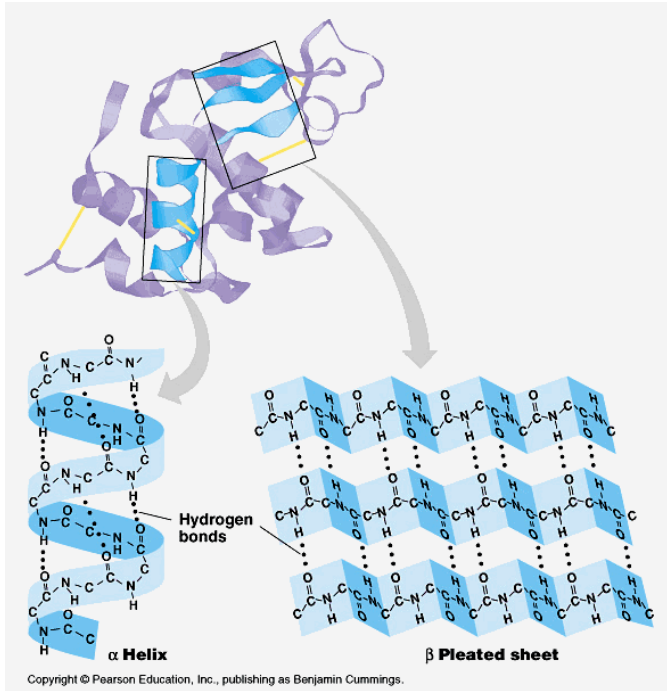


122 SAPs in disordered regions,  
114 (93%) are disease-associated.

(Image adapted from: Kissinger CR, et al. 1995. "Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex." Nature 378:641-4.)

From: <http://ist.temple.edu/disprot/index.php>

# Hydrogen bond change



Changed Hydrogen bond	Disease	Polymorphism	ratio
-6	1	0	1/0
-5	12	1	12
-4	44	2	22
-3	114	16	7.25
-2	230	55	4.18
-1	403	213	1.89
0	1142	716	1.59
1	224	142	1.58
2	68	36	1.89
3	11	4	2.75
4	0	2	0
5	0	2	0

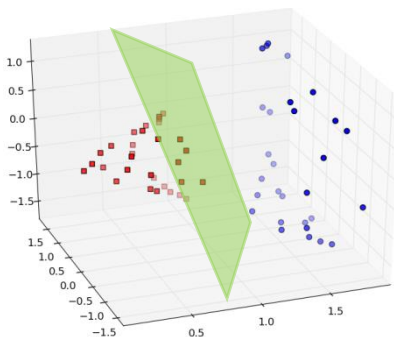
## Other attributes

- 52 variants in **transmembrane** regions, 49 (94%) are disease-associated
- 194 variants altered  $\beta$ -aggregation properties, 169 (87%) are disease-associated
- 435 variants from HLA families, all except one are “polymorphism”.

# SVM classifier

- **SVM – support vector machine**

Separate transformed data with a hyper plane in a high-dimensional space



- **Kernel function – Radial Basis Function(RBF)**

- **Grid-search to select proper values of parameter**

# Five-fold cross-validation

Part	Total proteins	Total SAP	Deleterious SAP	Neutral SAP
1	105	686	449	237
2	104	688	450	238
3	105	688	450	238
4	105	688	450	238
5	103	688	450	238
Total	522	3438	2249	1189

# Accuracy: ACC and MCC

SAP status	Predicted as disease-association (+)	Predicted as polymorphism (-)
Disease-association (+)	TP	FN
Polymorphism (-)	FP	TN

Overall accuracy:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Matthew correlation coefficient:

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TN + FN)(TN + FP)(TP + FN)(TP + FP)}}$$




# Predictive power





Attribute groups	ACC (%)	MCC
Residue frequencies	77.5	0.489
13 Å structural neighbor profile	76.4	0.467
Conservation scores	74.8	0.425
Nearby functional sites	69.3	0.246
Solvent accessibilities	68.0	0.232
C <sub>β</sub> density	67.0	0.190
Final attribute set	82.6	0.604

# SAPRED web server

<http://sapred.cbi.pku.edu.cn/>



## SAP Disease-Association Predictor

[Home](#)[Supp](#)[Document](#)[Contact](#)

---

**Program**  
[Run sapred](#)  
[Run sapred\\_seq](#)  
**Management**  
[My Data](#)  
[History](#)  
**Users**  
[Login in](#)  
[Try Out](#)  
[Free Registration](#)

Single Amino acid Polymorphisms (**SAPs**), also known as non-synonymous Single Nucleotide Polymorphisms (**nsSNPs**), account for about 50% of the gene lesions known to be related to inherited diseases. Through large-scale efforts such as HapMap project and The Cancer Genome Atlas (TCGA), available SAP data is accumulating rapidly in databases such as dbSNP, HGVBBase, Swiss-Prot variant page and many allele-specific databases. This provides us the opportunities and needs to understand and predict their disease-association.

**SAPRED**, the SAP disease-association predictor, offers the researchers an automatic pipeline to predict the disease-association of SAPs. Compared with other similar tools, **SAPRED** utilizes several novel attributes such as Structural Neighbor Profile and Nearby Functional Sites, in addition to incorporating other well-known attributes such as Residue Frequency and Conservation. By feeding these attributes to the internal trained SVM classifier, **SAPRED** outputs the final prediction result as well as the corresponding likelihood. The attributes themselves are also presented due to their potential biological significance.

Currently **SAPRED** affords two types of predictions. One is based on both the structural and sequence information, the other relies on the sequence information only. The former aims at higher prediction accuracy and more attributes with putative biological insights, while the latter covers much more inputs whose structural models are not available at present.

# Run SAPRED



## SAP Disease-Association Predictor

Currently SAPRED supports only one substitution each time. The user should also supply the PDB files of the wildtype and variant protein with enough quality for extracting structural information. Make sure the PDB file contains only one chain and its residue numbering system is consistent with the FASTA sequence. Such PDB file can be prepared using Swiss-Model or Modeller. If the structure model is unavailable, the user can switch to sapred\_seq instead, whose accuracy is a little lower. The users can try these demos first: [demo1](#) [demo2](#) [demo3](#)

### Input

<input type="checkbox"/> * input fasta file:	<input type="text" value="-----"/>	
* mutation name:	<input type="text"/>	
<input type="checkbox"/> * wildtype pdb file:	<input type="text" value="-----"/>	
<input type="checkbox"/> * variant pdb file:	<input type="text" value="-----"/>	

### Output

* save result in directory:	<input type="text" value="Work Directory"/>	
* prediction result: ( bio:sapred:sapred )	<input type="text" value="Untitled.sapred"/>	

Run

# Results

## Prediction Result

Prediction	Disease Likelihood	Neutral Likelihood
Disease	0.88069	0.11931

## Explanation of Results: Structural features

## Structure-derived attributes

13A structural neighbor profile	A	0	G	1	M	1	S	3
	C	4	H	1	N	2	T	2
	D	2	I	3	P	3	V	0
	E	0	K	8	Q	2	W	0
	F	1	L	5	R	1	Y	0
Structurally Nearby Functional Sites	ACT_SITE							14
	BINDING							14
	METAL							14
	MOD_RES							14
	DISULFID							14
Solvent Accessibilities	Side-chain absolute							88.44
	Side-chain relative							48.5
C-beta density	16							
Secondary structures	Secondary structure							C
	No alteration of secondary structure							Y
	No alteration of 3-mer secondary structure							Y
Dihedrals	Phi Difference						1.350	
	Psi Difference						1.530	
	Chi1 Difference						0.914	
Changed Hydrogen Bonds	0							
Changed Disulfide Bonds	0							
RMSD	0.07358							
Difference between energy	2.1752							

# Explanation of Results: sequence features

## Sequence-derived attributes

Residue frequencies	Frequency of wildtype residue	0.8591
	Frequency of variant residue	0.0052
	Difference residue frequency	-0.8539
Conservation Scores	neibor3L	0.7588
	neibor2L	1.0534
	neibor1L	0.7201
	conserv	1.1102
	neibor1R	0.1574
	neibor2R	0.1906
	neibor3R	1.0747
Sequentially Nearby Functional Sites	ACT_SITE	30
	BINDING	30
	METAL	30
	MOD_RES	30
Aggregation Properties	tango_wt	0.00
	diff_tango	0
	frag_equal	Y
BLOSUM Score	-3	
GRANTHAM Score	116	
In TRANSMEM Region	N	
In Disordered Region	N	
In HLA Family	N	



# Results using SAPRED\_Seq

**ACC=81.5%**

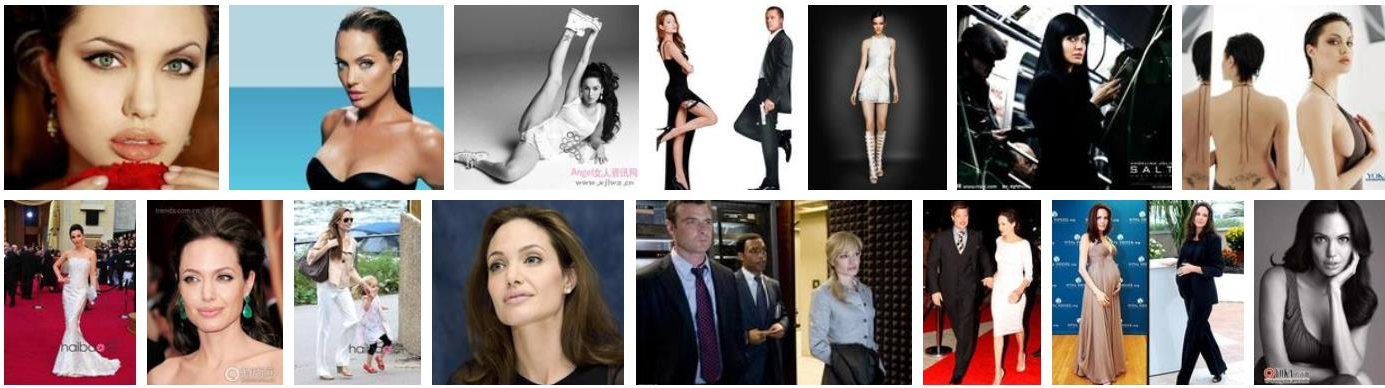
**MCC=0.577**

## Prediction Result

Prediction	Disease Likelihood	Neutral Likelihood
Disease	0.871749	0.128251

## Sequence-derived attributes

Residue frequencies	Frequency of wildtype residue		0.8591
	Frequency of variant residue		0.0052
	Difference residue frequency		-0.8539
Conservation Scores	neibor3L	0.7586	
	neibor2L	1.0534	
	neibor1L	0.7201	
	conserv	1.1102	
	neibor1R	0.1574	
	neibor2R	0.1906	
	neibor3R	1.0747	
Sequentially Nearby Functional Sites	ACT_SITE		30
	BINDING		30
	METAL		30
	MOD_RES		30
Aggregation Properties	tango_wt	0.00	
	diff_tango	0	
	frag_equal	Y	
BLOSUM Score	-3		
GRANTHAM Score	116		
In TRANSMEM Region	N		
In Disordered Region	N		
In HLA Family	N		



The New York Times

May 14, 2013

## My Medical Choice

By ANGELINA JOLIE  
LOS ANGELES

MY MOTHER fought cancer for almost a decade and died at 56. She held out long enough to meet the first of her grandchildren and to hold them in her arms. But my other children will never have the chance to know her and experience how loving and gracious she was.

# Angelina Jolie has a genetic variation in *BRCA1*

## Do you think she had made the right decision to remove her breasts?

# Lots of factors to consider in making this complicated decision

- Given her genetic mutation, what is the likelihood of her getting breast cancer?
  - $P(\text{cancer} | \text{mutation})$
  - $P(\text{cancer-free} | \text{mutation})$
- Even after she got mastectomy, what is the likelihood of her getting breast cancer?
  - $P(\text{cancer} | \text{mutation, mastectomy})$
- If she didn't get mastectomy, what could her outcome be?
  - $P(\text{early detection} | \text{mutation, cancer})$
  - $P(\text{cure} | \text{mutation, cancer, early detection})$
  - $P(\text{new treatment before cancer develops})$
- What is the risk of death from the procedure of mastectomy?
  - $P(\text{death from surgery} | \text{mastectomy})$
- Age of onset of cancer
- Emotional stress: fear of cancer vs. distress over loss of breasts
- Cost of mastectomy

# Using family history to increase prediction power

- Strong family history
  - Jolie's mother, Marcheline Bertrand, died from ovarian cancer at 56 after a 10 year battle.
  - Her aunt, 61-year-old Debbie Martin, is dying of breast cancer after a 9 year battle.
  - Her grandmother, Lois Bertrand, died of cancer at 45.
  - Her great-grandmother, Virginia Gouwens, died of ovarian cancer at 53.
  - Her uncle, Raleigh, also died of cancer in 2009.
- Strong family history, early-onset, poor prognosis
- Has the causal mutation been identified in her affected relatives and does it co-segregate with cancer in her family?
- Does Jolie carry the same mutation?

# **Lots of remaining challenges**

- Prediction accuracy
- Integration of multiple sources of evidence
- Noncoding variants
- More training data
- Ethical issues

**Wherever there are challenges, there are opportunities.**

# 生物信息学：导论与方法

## Bioinformatics: Introduction and Methods

Ge Gao 高歌 & Liping Wei 魏丽萍

Center for Bioinformatics, Peking University



<https://www.coursera.org/course/pkubioinfo>