生物信息学

蛋白质结构比较

吴凌云

中国科学院数学与系统科学研究院

目录

- **结构比较**
- 结构相似性
- 结构比对方法
- 结构比较方法

为什么做结构比较？

- 蛋白的结构决定了功能
- 进化上蛋白结构比序列要保守得多
- 结构比较在建立远距离的进化关系时非常有用，因为这些关系是无法通过序列方法发现的

结构比较的用途

- 度量结构的相似性
- 发现远距离进化关系
- 预测蛋白功能
- 分析蛋白家族中的结构差异
- 分析ligand结合时蛋白构象的变化
- 识别蛋白功能位点（结构保守区域）
- 识别常见结构motif
- ……

构象变化

- 当蛋白与ligand或者substrate结合时，会产生结构上的变化
  - 使得结合位点上的催化反应变得容易
  - 促进其他位点上的substrates的结合
- 通过比较结合了ligand的结构和未结合时的结构
  - 可以帮助我们了解ligand结合的过程
  - 辅助药物设计

结构保守性

- 在某些蛋白家族中，结构变化对蛋白功能的影响较小
- 而在另一些家族中，结构变化对功能的影响非常明显
- 通过对同一家族蛋白的多结构比对可以帮助我们识别结构中保守（重要）的部分
- 这些结构保守区域与功能的关系密切

结构motif

- 进化上没有关系的蛋白也可能有结构相似
- 没有明显进化关系，但结构相似的蛋白，我们称为类似（analogs）
- 这种结构相似可能是二级结构折叠的限制
- 二级结构的最优折叠方式是有限的
- 超二级结构（supersecondary structures）

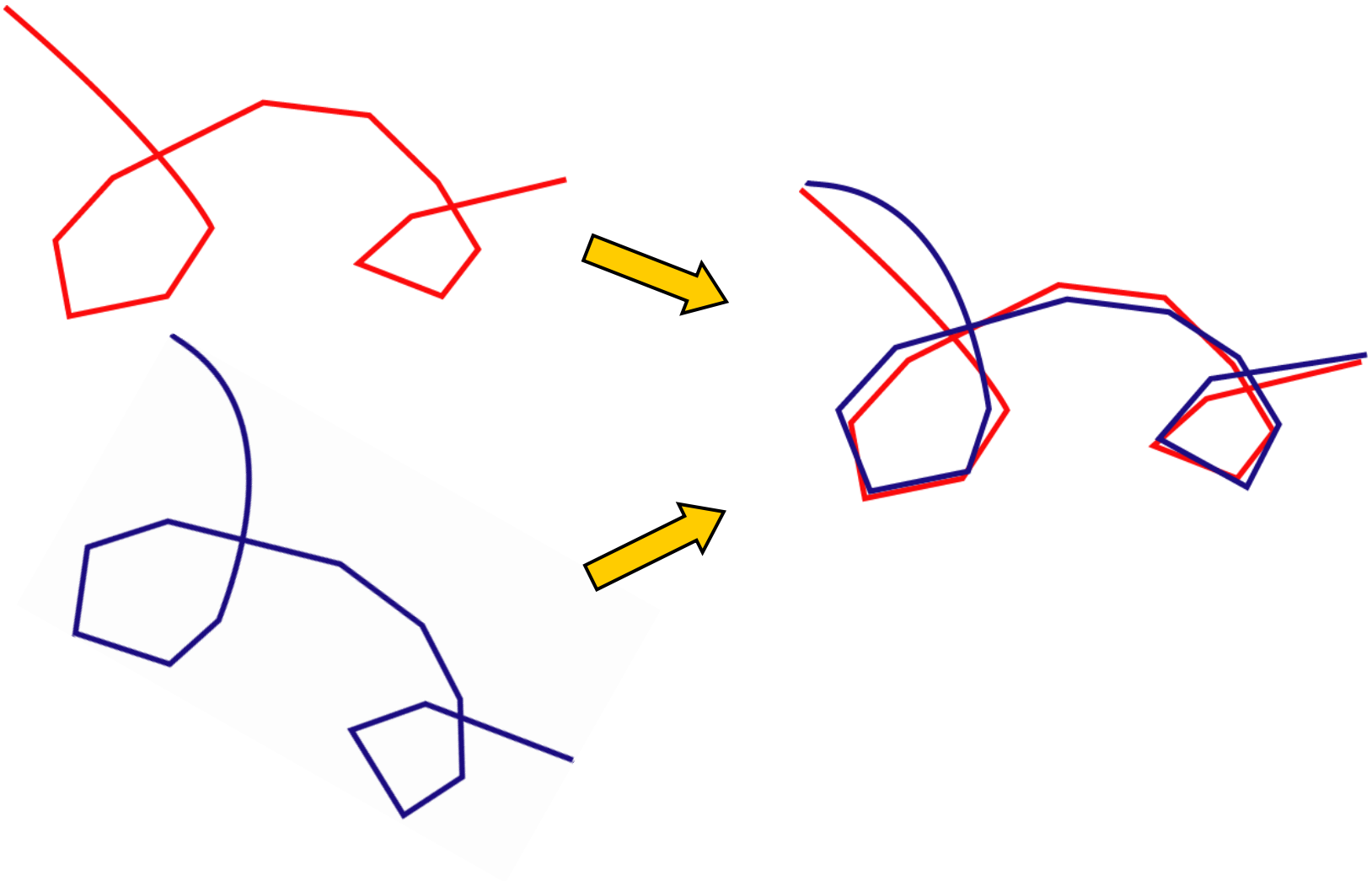
Divergent 和 Convergent

- Divergent Evolution （趋异进化）
  - 两种以上型态或分子具有共同起源，但在进化过程中逐渐分化，具有不同的功能或构造
- Convergent Evolution （趋同进化）
  - 两种以上型态或分子具有相同功能或构造，但却源自不同起源

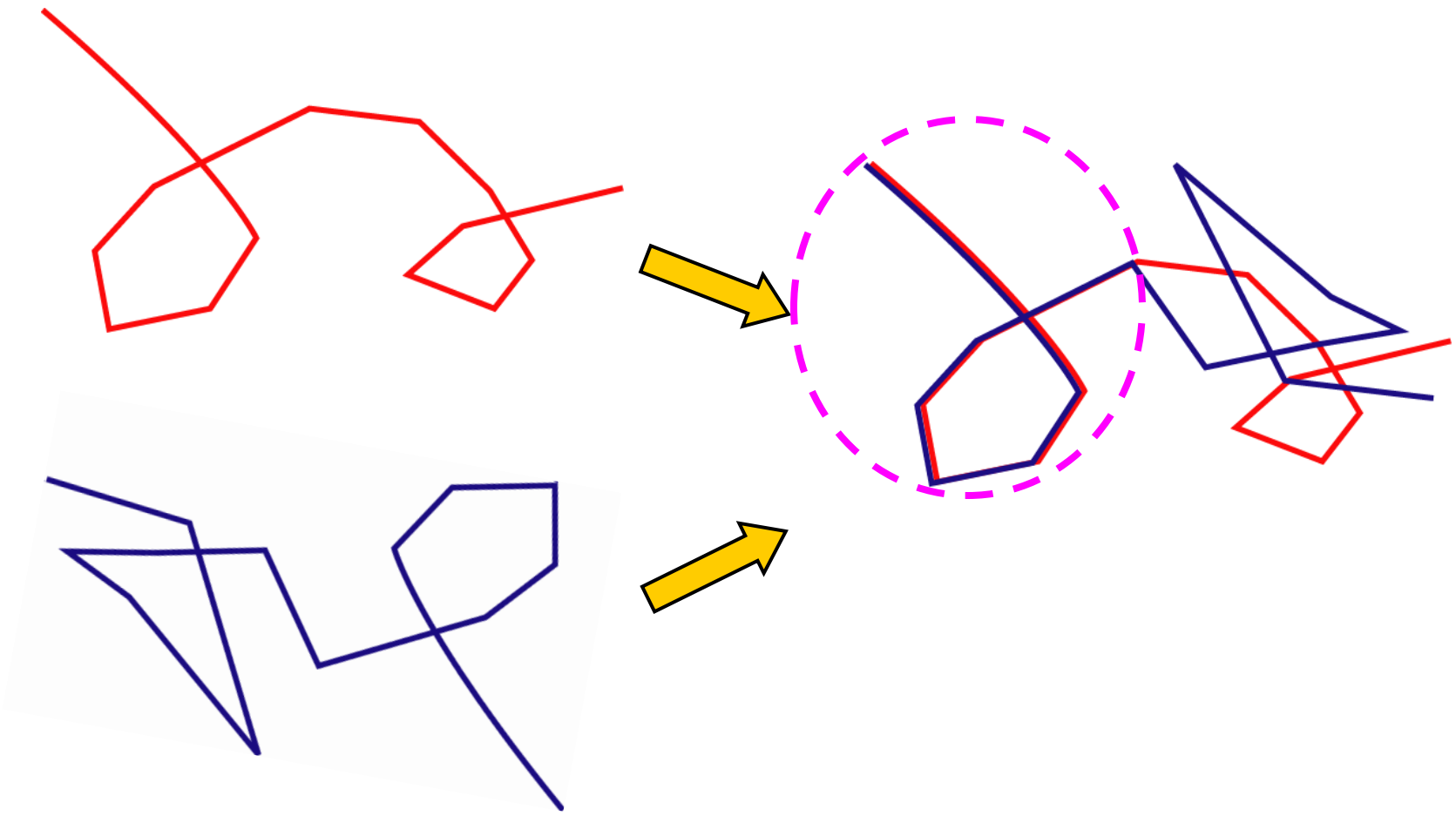
结构比对和序列比对

- 蛋白质序列是由20个字母组成的字符串
  - 给定一个评分系统，可以通过动态规划(DP)得到序列之间最佳的匹配方式
- 蛋白质结构是一个三维的形状
  - 希望找到类似DP的算法，得到两个三维形状之间的最佳匹配方式

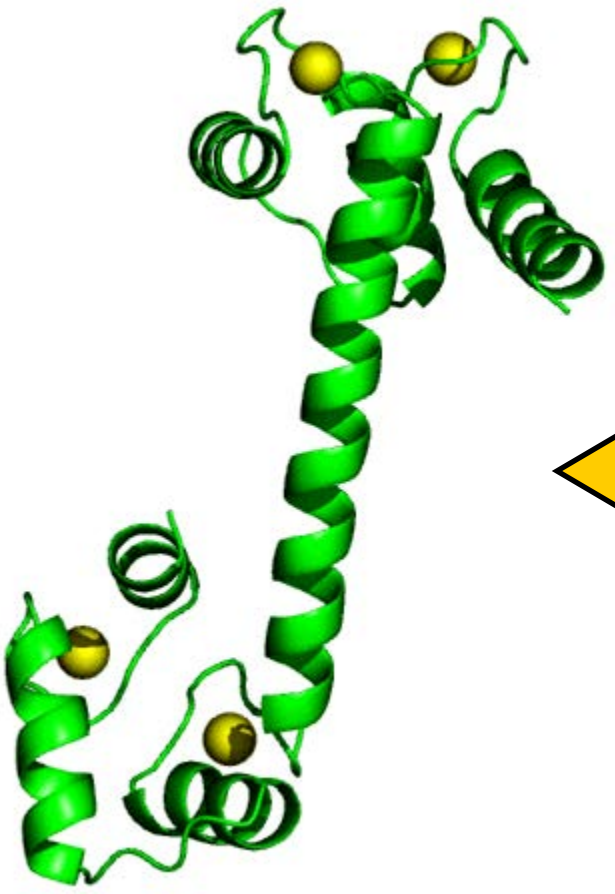
全局比对



局部比对



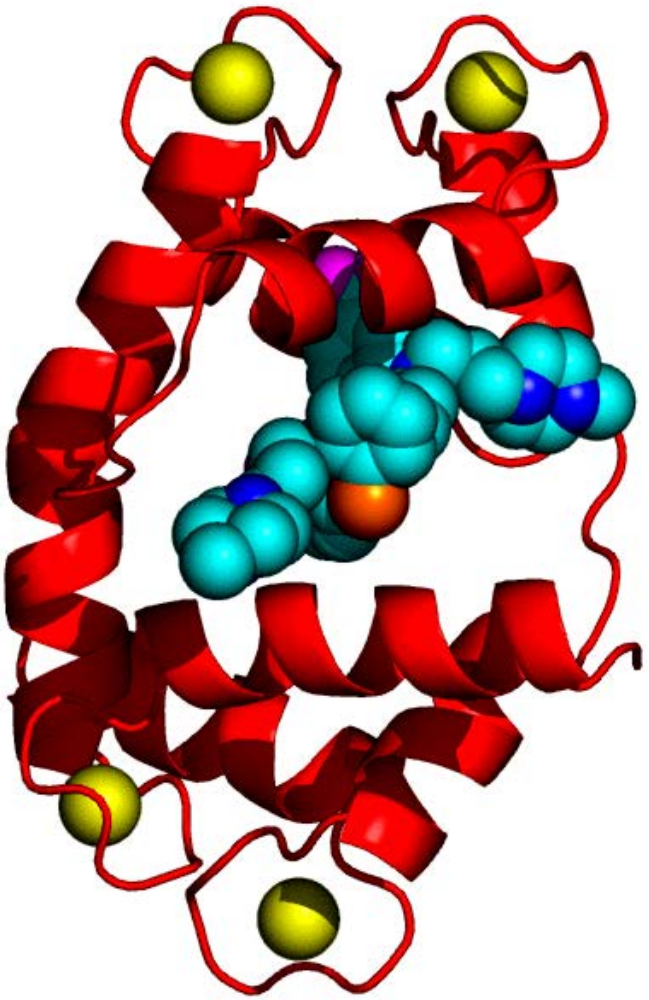
Example: Calmodulin



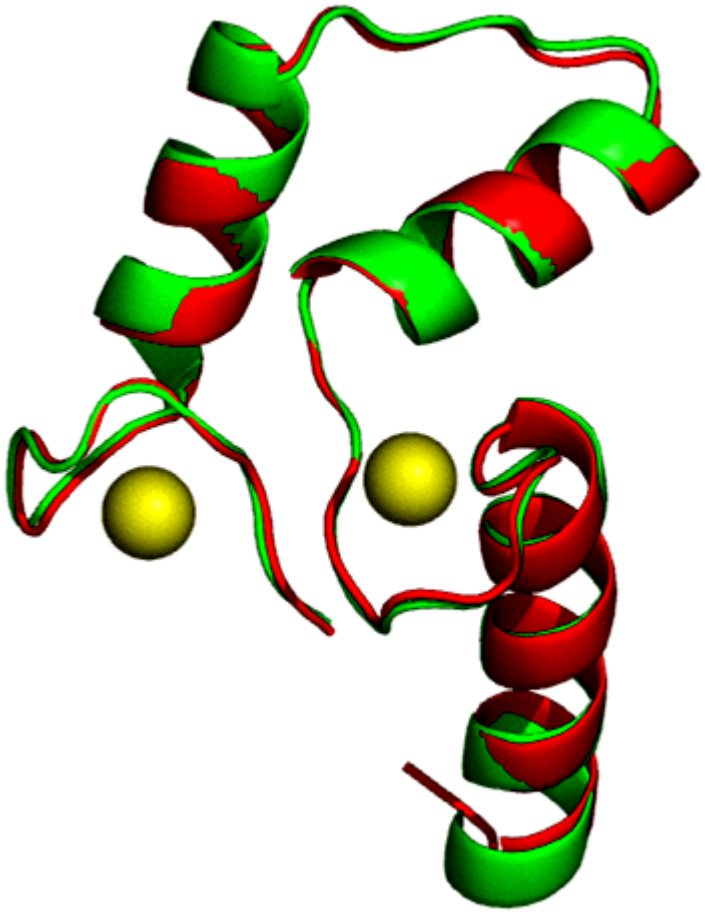
Two forms of calcium-bound Calmodulin (钙调蛋白):

Ligand free

Complexed with Trifluoperazine (三氟拉嗪)



Example: Calmodulin



Global alignment:

RMSD =15 Å /143 residues

Local alignment:

RMSD = 0.9 Å/ 62 residues

目录

- 结构比较
- **结构相似性**
- 结构比对方法
- 结构比较方法

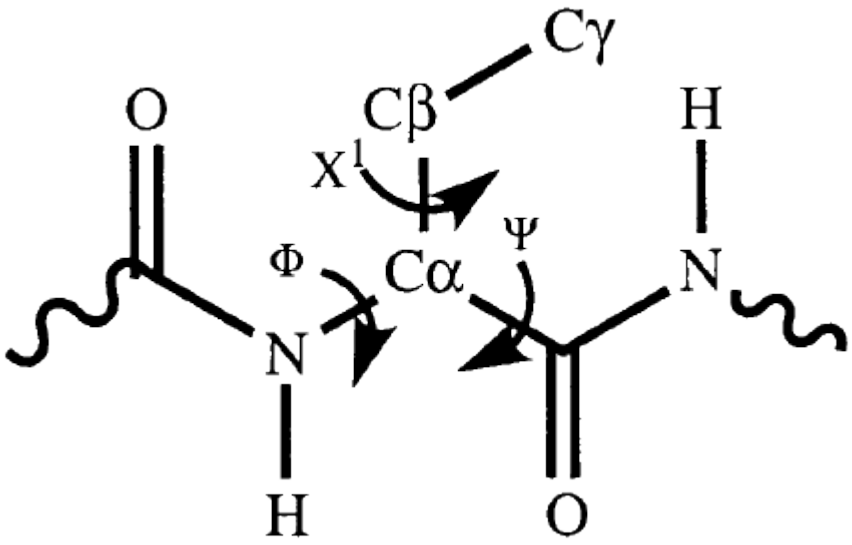
如何度量结构的相似性

- 视觉比较
- 二面角
- 距离矩阵
- RMSD (root mean square distance)

Is the resulting distance (similarity measure) D a metric?

$$D(A,B) \leq D(A,C) + D(C,B)$$

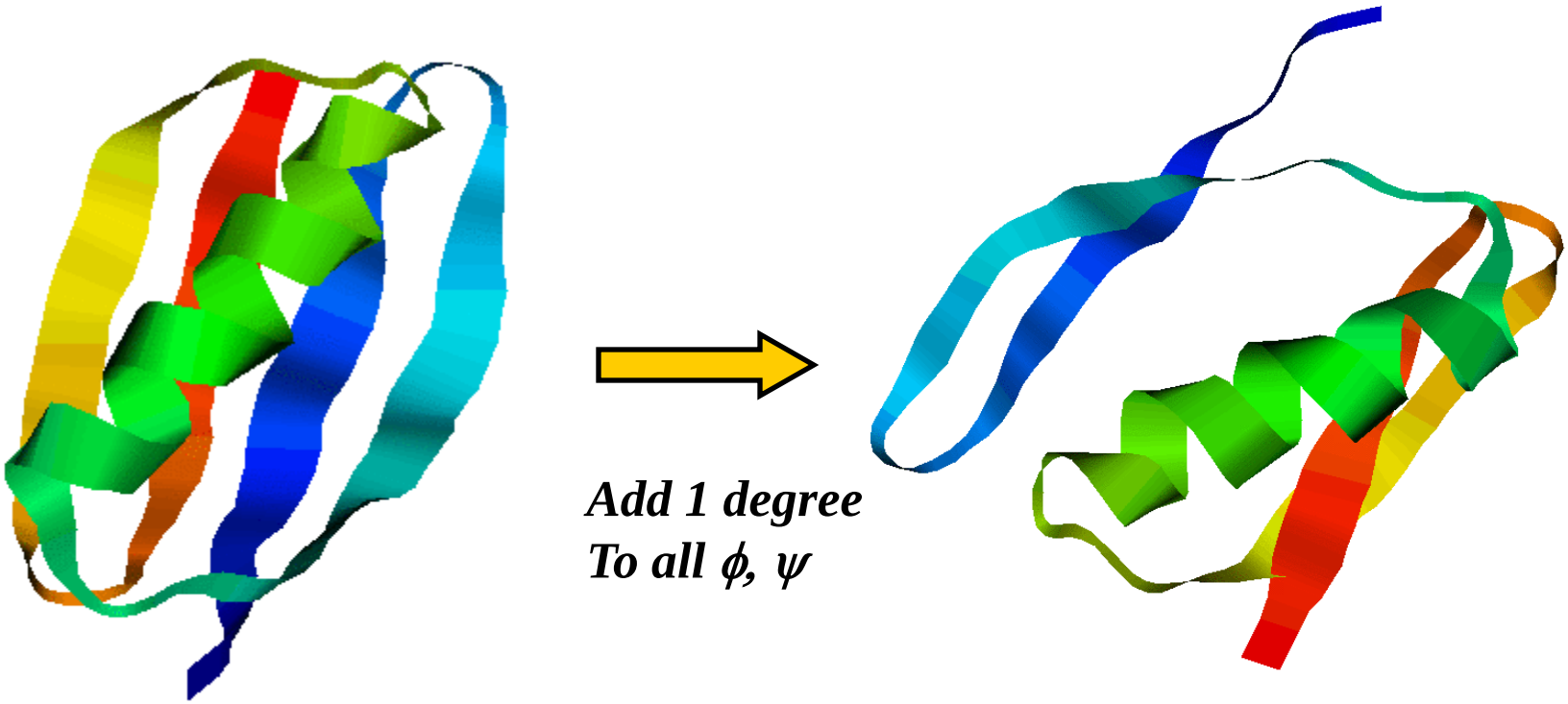
二面角



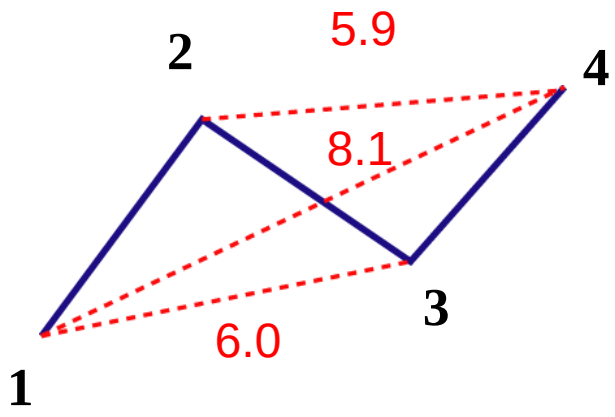
二面角——优点

- 局部的结构信息
- 对于蛋白质分子的旋转和平移保持不变
- 简洁（长度为n的蛋白质只有O(n)个二面角）

二面角——缺点



距离矩阵



	1	2	3	4
1	0	3.8	6.0	8.1
2	3.8	0	3.8	5.9
3	6.0	3.8	0	3.8
4	8.1	5.9	3.8	0

距离矩阵

- 优点
  - 对于旋转和平移保持不变
  - 可以用于比较蛋白质
- 缺点
  - 长度为n的蛋白质的距离矩阵大小为$O(n^2)$
  - 比较两个距离矩阵是一个困难的问题
  - 对手性(chirality)不敏感

RMSD

- 蛋白质的空间坐标表示
- 定义氨基酸（原子）对应关系：

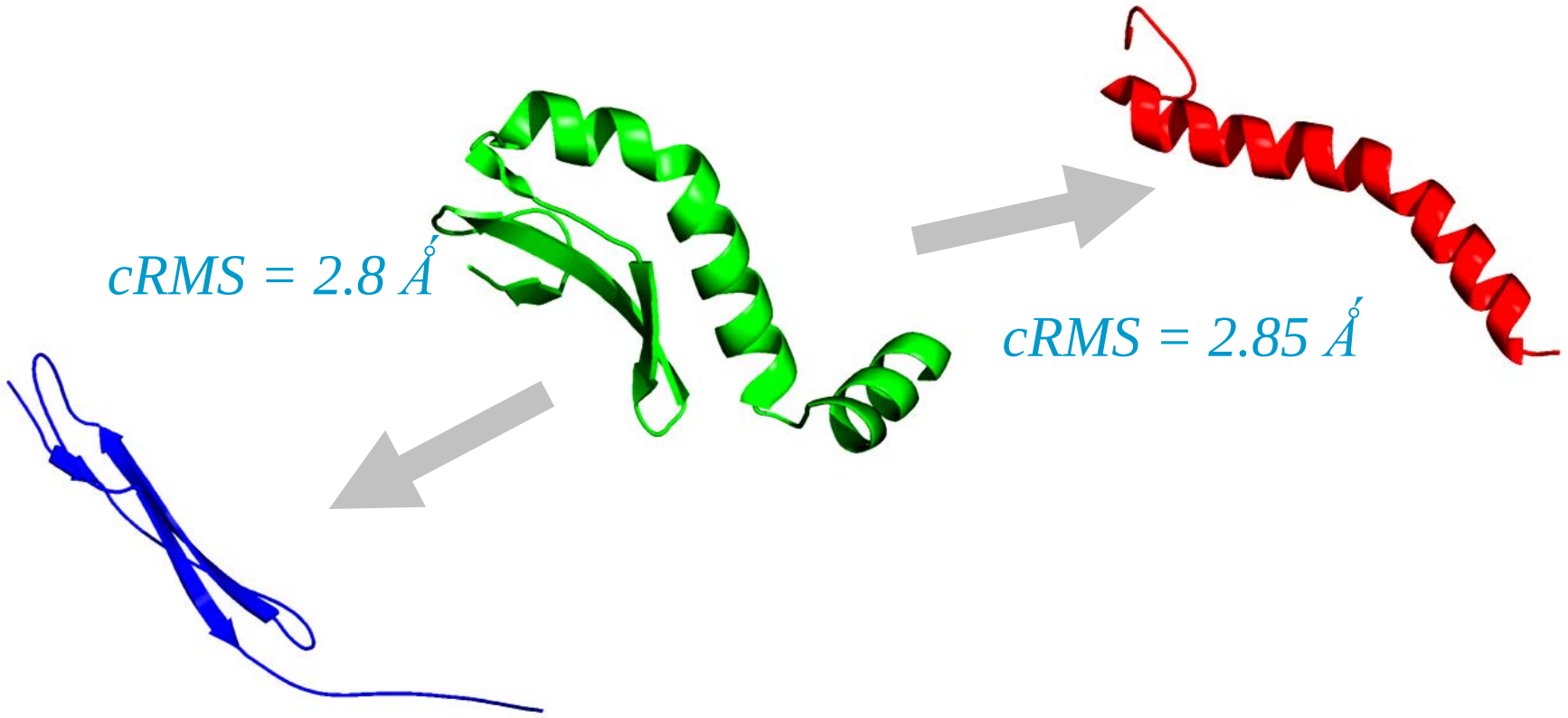
$$(A_{i_1}, A_{i_2}, \cdots, A_{i_K})$$

$$(B_{i_1}, B_{i_2}, \cdots, B_{i_K})$$

- 计算RMSD：

$$RMS(A,B) = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \| A_{i_k} - B_{j_k} \|^2}$$

RMSD is not a Metric



Superposition

- 简化问题：已知对应关系，求最佳的刚体变换T使得RMSD最小
- 目标函数：

$$\min_{T} \epsilon = \sum_{k=1}^{K} \|T(A_k) - B_k\|^2$$

- 老问题：Statistics, Robotics, Medical Image Analysis, …

Superposition

- 刚体变换 (rigid-body transformation) 由旋转 (rotation) **R**和平移 (translation) **t**组成：
- 求解优化问题：

$$T(x) = \boldsymbol{R}x + \boldsymbol{t}$$

$$\min_{\boldsymbol{R},\boldsymbol{t}} \epsilon = \sum_{k=1}^{K} \|\boldsymbol{R}A_k - B_k + \boldsymbol{t}\|^2$$

平移

$$\frac{\partial \epsilon}{\partial \boldsymbol{t}} = 2\sum_{k=1}^{K}(\boldsymbol{R}A_k - B_k + \boldsymbol{t}) = 0$$

$$\boldsymbol{t} = \frac{1}{K}\left(-\boldsymbol{R}\sum_{k=1}^{K}A_k + \sum_{k=1}^{K}B_k\right)$$

- 如果A和B都以原点为中心，则t=0

协方差矩阵

barycenters

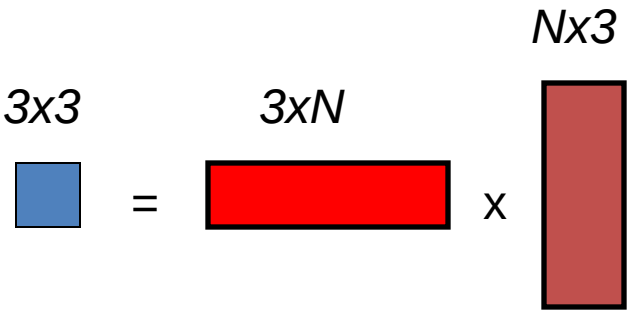
$$\mu_A = \frac{1}{K}\sum_{k=1}^{K} A_k, \quad \mu_B = \frac{1}{K}\sum_{k=1}^{K} B_k$$

$$X = (A_1 - \mu_A, A_2 - \mu_A, \cdots, A_k - \mu_A)$$

$$Y = (B_1 - \mu_B, B_2 - \mu_B, \cdots, B_k - \mu_B)$$

covariance matrix

$$C = XY^T$$



旋转

- Singular Value Decomposition

$$C = UDV^T$$

- 定义

$$S = \begin{cases} I, & \text{if } \det(C) > 0 \\ \text{diag}\{1, 1, -1\}, & \text{otherwise} \end{cases}$$

$$R = USV^T$$

Superposition算法

$$\mu_A = \frac{1}{K}\sum_{k=1}^{K} A_k, \quad \mu_B = \frac{1}{K}\sum_{k=1}^{K} B_k$$

$$X = (A_1 - \mu_A, A_2 - \mu_A, \cdots, A_k - \mu_A)$$
$$Y = (B_1 - \mu_B, B_2 - \mu_B, \cdots, B_k - \mu_B)$$

$$C = XY^T$$

$$C = UDV^T$$

$$S = \begin{cases} I, & \text{if } \det(C) > 0 \\ \text{diag}\{1, 1, -1\}, & \text{otherwise} \end{cases}$$

$$R = USV^T$$

$$t = \frac{1}{K}\left(-\boldsymbol{R}\sum_{k=1}^{K} A_k + \sum_{k=1}^{K} B_k\right)$$

O(N) in time!

目录

- 结构比较
- 结构相似性
- **结构比对方法**
- 结构比较方法

结构比对问题

- 3维空间中的两个点集 A=(A1, A2, …, An) 和 B=(B1, B2,…, Bm)
- 找到最优的子集 A(P) 和 B(Q) 满足 |A(P)|=|B(Q)|
- 同时找到最优的刚体变换 T_{opt} 使得 A(P) 和 B(Q) 之间的距离 D 最小
- A(P) 和 B(Q) 定义了一个 “比对”
- L = |A(P)|=|B(Q)| 称为比对长度

$$\min_{T} \quad D\left(T(A(P)), B(Q)\right)$$

关键子问题

- 找对应的子集和对应关系（alignment）
- 找最佳空间变换（superposition）

软件

- DALI (Holm and Sander, 1993)
- SSAP (Orengo and Taylor, 1989)
- STRUCTAL (Levitt et al, 1993)
- VAST [Gibrat et al., 1996]
- LOCK [Singh and Brutlag, 1996]
- CE [Shindyalov and Bourne, 1998]
- SSM [Krissinel and Henrik, 2004]
- …

结构比对方法分类

- Distance-matrix-based Methods
- Coordinate-based Methods
- Secondary-structure-based Methods

Distance-matrix-based Methods

- DALI [Holm and Sander, 1993]
- CE [Shindyalov and Bourne, 1998]
- FATCAT [Ye and Godzik, 2003]
- SSAP [Orengo and Taylor, 1989]
- …

Coordinate-based Methods

- STRUCTAL [Levitt et al, 1993]
- SAMO [Chen et al, 2006]
- TM-align [Zhang and Skolnick, 2005]
- ProSup [Lackner et al, 2000]
- …

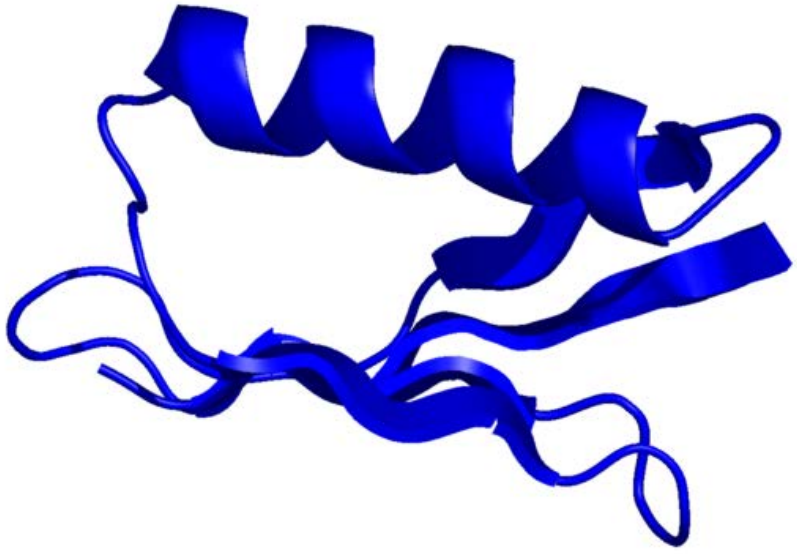
Secondary-structure-based Methods

- VAST [Gibrat et al., 1996]
- SSM [Krissinel and Henrik, 2004]
- LOCK [Singh and Brutlag, 1997]
- …

DALI

- **DALI (Distance ALIgnement)**
- Web Server:
  - http://www.ebi.ac.uk/dali/
  - http://www.ebi.ac.uk/DaliLite/
- Databases and resources:
  - http://ekhidna.biocenter.helsinki.fi/software#dali
- DaliLite (standalone version, pairewise comparison)
  - http://ekhidna.biocenter.helsinki.fi/dali_lite/downloads

距离矩阵



Structure

Euclidian distance

Distance geometry

0 3.8 …
0 3.8 …
…
0

Distance matrix

优缺点

- 优点
  - 不需要找最佳空间变换
- 缺点
  - 对手性不敏感
  - 比较两个距离矩阵是一个困难的问题

DALI算法

1. 找到所有相匹配的hexapeptites

$$D = \sum_{i=1}^{6}\sum_{j=1}^{6}(e - \frac{|d_{ij}^{A} - d_{ij}^{B}|}{d_{ij}^{*}}) / \exp(-(\frac{d_{ij}^{*}}{f})^{2})$$

$$d_{ij}^{*} = \frac{1}{2}(d_{ij}^{A} + d_{ij}^{B})$$

2. 使用模拟退火方法将匹配的hexapeptides拼接起来

STRUCTAL

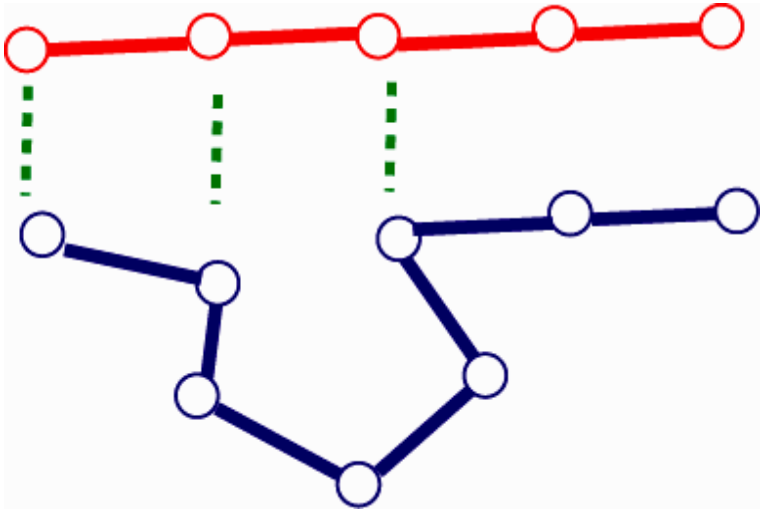
- 基于空间坐标
- 迭代循环算法
  - 给定一个初始的比对C
  - 计算最佳的空间变换T，并将其作用到蛋白质A上
  - 根据变换后的蛋白质A和B的坐标，更新比对C
  - 如果比对C更新了，则返回第二步，否则终止

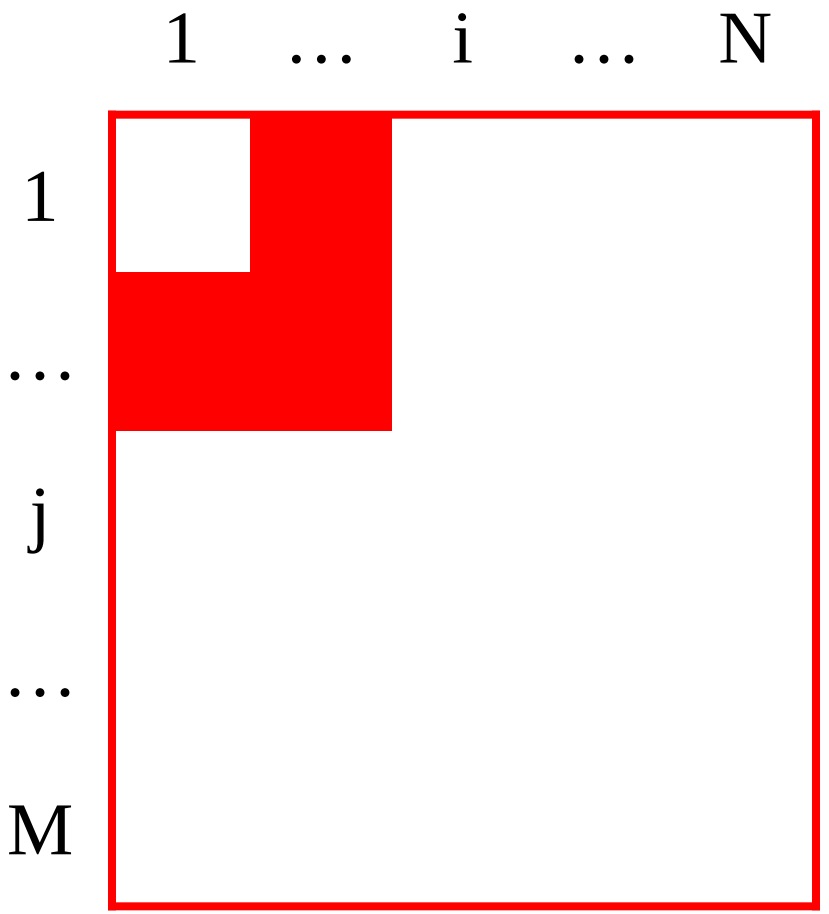
初始比对

- STRUCTAL使用了五种初始比对
  - 首端对齐，没有间隙
  - 尾端对齐，没有间隙
  - 中点对齐，没有间隙
  - 最佳序列比对结果（使用单位矩阵为分数矩阵）
  - 二面角相似性

更新比对

动态规划(DP)





分数矩阵

$$SA(i,j) = \frac{20}{1 + 5d(A_i, B_j)^2}$$

$$Penalty = 10$$

多项式时间近似算法(1)

- 如果
  - 给定一个空间变换，最优的比对可以在多项式时间内找到
  - 需要考虑的空间变换数目是多项式有界的
- 则存在蛋白质比对的多项式时间算法

多项式时间近似算法(2)

- R. Kolodny and N. Linial, *Proc. Natl. Acad. Sci. (USA)*, **101**, 12201-12206 (2004). Approximate Protein Structural Alignment in Polynomial Time.

$\forall \epsilon > 0$, $\exists$ a finite set $G(\epsilon)$ of transformation such that:

1. $|G(\epsilon)|$ is polynomial in N
2. $\forall T$, $\exists T_G \in G(\epsilon)$ such that $\|S(T) - S(T_G)\| < \epsilon$

复杂性： $O(N^{10}/\epsilon^6)$

SAMO

- Structure Alignment by Multi-objective Optimization

$$T(S, A, R) = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} s_{ij} |A + RX_i - Y_j|^2$$

$$m(S) = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} s_{ij}$$

BMC Structural Biology (2006), vol. 6, 18

SAMO

$$\text{v-minimize} \quad \big(T(S,A,R),\ -m(S)\big) \qquad \text{for } S,A,R$$

$$\text{subject to} \quad \sum_{i=1}^{n_x} s_{ij} \leq 1 \text{ for } j = 1,...,n_y$$

$$\sum_{j=1}^{n_y} s_{ij} \leq 1 \text{ for } i = 1,...,n_x$$

$$s_{ij} \in \{0,1\}$$

BMC Structural Biology (2006), vol. 6, 18

Epsilon Method

$$\text{minimize} \quad T(S,A,R) - \lambda^2 m(S) \qquad \text{for } S, A, R$$

$$\text{subject to} \quad \sum_{i=1}^{n_x} s_{ij} \leq 1 \text{ for } j = 1, ..., n_y$$

$$\sum_{j=1}^{n_y} s_{ij} \leq 1 \text{ for } i = 1, ..., n_x$$

$$s_{ij} \in \{0, 1\}$$

BMC Structural Biology (2006), vol. 6, 18

SAMO Algorithm

$$\underset{A,R}{\text{minimize}} \quad \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} s_{ij}|A + RX_i - Y_j|^2$$

$$\underset{S}{\text{minimize}} \quad \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} s_{ij}(|A + RX_i - Y_j|^2 - \lambda^2)$$

$$\text{subject to} \quad \sum_{i=1}^{n_x} s_{ij} \leq 1 \text{ for } j = 1, ..., n_y$$

$$\sum_{j=1}^{n_y} s_{ij} \leq 1 \text{ for } i = 1, ..., n_x$$

$$s_{ij} \in \{0, 1\}$$

BMC Structural Biology (2006), vol. 6, 18

SAMO Enhancement

- SAMO with Weights

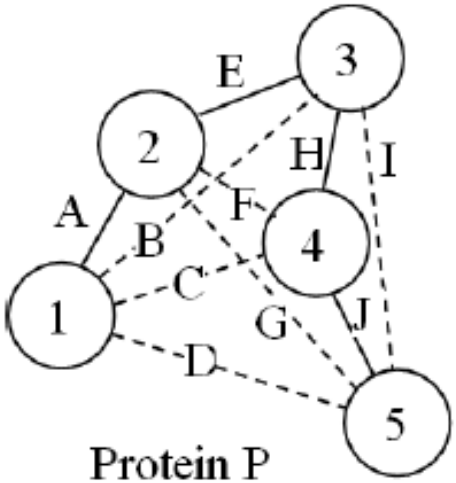
$$T(S,A,R)=\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}s_{ij}|A+RX_i-Y_j|^2w_{ij}$$

$$m(S)=\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}s_{ij}w_{ij}$$

SAMO Enhancement

- Possible weights
  - Amino acid similarity
  - Secondary structure similarity
  - Local structure similarity
- Problems
  - Influence to solution
  - Biological meaning

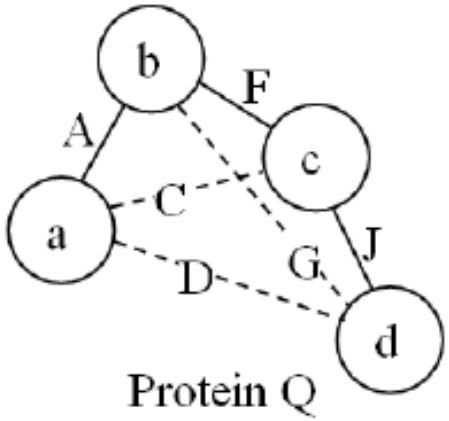
MatAlign



Protein P

	1	2	3	4	5
1	0	A	B	C	D
2	A	0	E	F	G
3	B	E	0	H	I
4	C	F	H	0	J
5	D	G	I	J	0

Distance matrix of P



Protein Q

	a	b	c	d
a	0	A	C	D
b	A	0	F	G
c	C	F	0	J
d	D	G	J	0

Distance matrix of Q

row-row matching scores

	a	b	c	d
1	4	1	1	1
2	1	4	1	1
3	1	1	1	1
4	1	1	4	1
5	1	1	1	4

P	1	2	4	5
Q	a	b	c	d

Alignment of P and Q

MatAlign

$$\mathrm{Match}(d1, d2) = \begin{cases} \alpha/(|d1 - d2| + \alpha) & \text{if } |d1 - d2| \leq T_{\mathrm{Match}} \\ 0 & \text{otherwise} \end{cases}$$

MatAlign

	1	2	3	4	5	6	7
1	0.00	11.00	1.00	2.00	3.00	4.00	16.00
2	11.00	0.00	12.00	13.00	14.00	15.00	17.00
3	1.00	12.00	0.00	5.00	6.00	7.00	18.00
4	2.00	13.00	5.00	0.00	8.00	9.00	19.00
5	3.00	14.00	6.00	8.00	0.00	10.00	20.00
6	4.00	15.00	7.00	9.00	10.00	0.00	21.00
7	16.00	17.00	18.00	19.00	20.00	21.00	0.00

Distance Matrix of Protein *A*

	1	2	3	4	5	6
1	0.00	1.05	2.10	3.15	11.05	4.20
2	1.05	0.00	5.05	6.10	12.10	7.15
3	2.10	5.05	0.00	8.20	13.15	9.05
4	3.15	6.10	8.20	0.00	14.20	10.10
5	11.05	12.10	13.15	14.20	0.00	15.05
6	4.20	7.15	9.05	10.10	15.05	0.00

Distance Matrix of Protein *B*

Row #1 from B's Distance Matrix ($\mathcal{DM}_B[1]$)

		1	2	3	4	5	6
		0.00	1.05	2.10	3.15	11.05	4.20
	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	1.00	1.00	1.00	1.00	1.00
2	11.00	0.00	1.00	1.09	1.10	1.11	1.95	1.95
3	1.00	0.00	1.00	1.95	1.95	1.95	1.95	2.19
4	2.00	0.00	1.00	1.95	2.86	2.86	2.86	2.86
5	3.00	0.00	1.00	1.95	2.86	3.73	3.73	3.73
6	4.00	0.00	1.00	1.95	2.86	3.73	3.86	4.56
7	16.00	0.00	1.00	1.95	2.86	3.73	3.90	4.56

Row #1 from A's Distance Matrix ($\mathcal{DM}_A[1]$)

$\mathcal{DM}_A[1]$	0.00	11.00	1.00	2.00	3.00	——	4.00	16.00
$\mathcal{DM}_B[1]$	0.00	——	1.05	2.10	3.15	11.05	4.20	——

MatAlign

		B					
		1	2	3	4	5	6
A	1	4.56	2.14	1.70	1.55	1.97	1.35
	2	1.61	2.15	2.39	2.27	4.52	1.78
	3	2.05	4.68	2.29	1.92	2.25	1.68
	4	1.80	2.23	4.65	2.56	2.27	2.25
	5	1.55	1.89	2.55	4.52	2.11	2.48
	6	1.26	1.61	1.96	2.75	1.91	4.56
	7	1.00	1.06	1.14	1.25	1.68	1.54

Row-Row Alignment Score Matrix (SM)

Ma

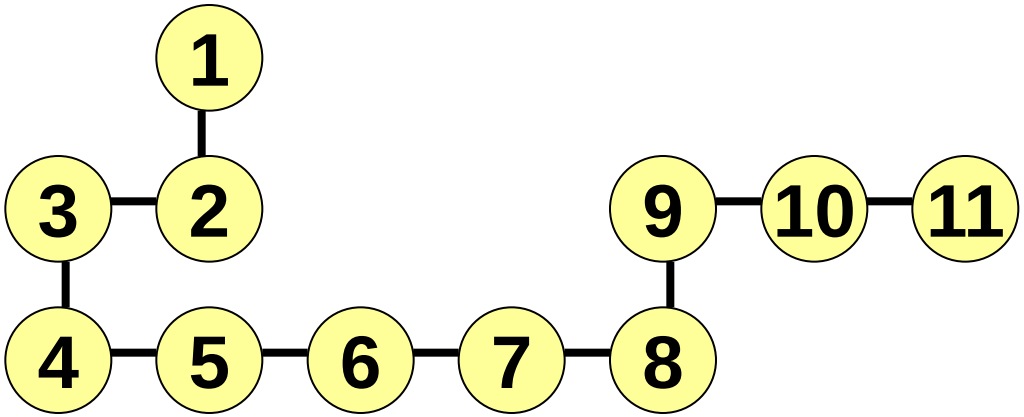
		B						
			1	2	3	4	5	6
		0.00	0.00	0.00	0.00	0.00	0.00	0.00
	1	0.00	4.56	4.56	4.56	4.56	4.56	4.56
	2	0.00	4.56	6.71	6.95	6.95	9.08	9.08
	3	0.00	4.56	9.25	9.25	9.25	9.25	10.76
A	4	0.00	4.56	9.25	13.89	13.89	13.89	13.89
	5	0.00	4.56	9.25	13.89	18.42	18.42	18.42
	6	0.00	4.56	9.25	13.89	18.42	20.33	22.98
	7	0.00	4.56	9.25	13.89	18.42	20.33	22.98

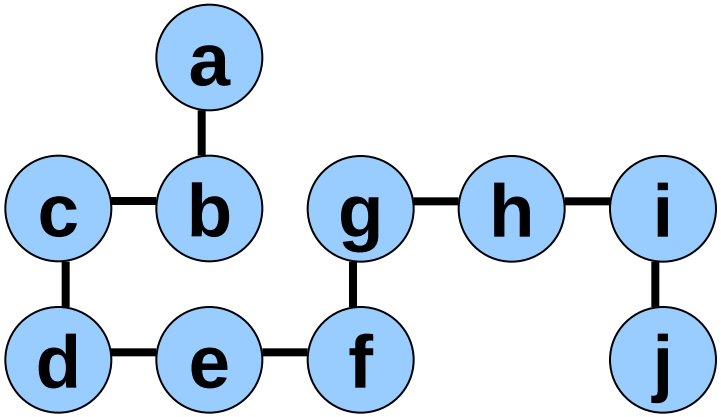
Dynamic Programming Matrix

A	1	2	3	4	5	——	6	7
B	1	——	2	3	4	5	6	——

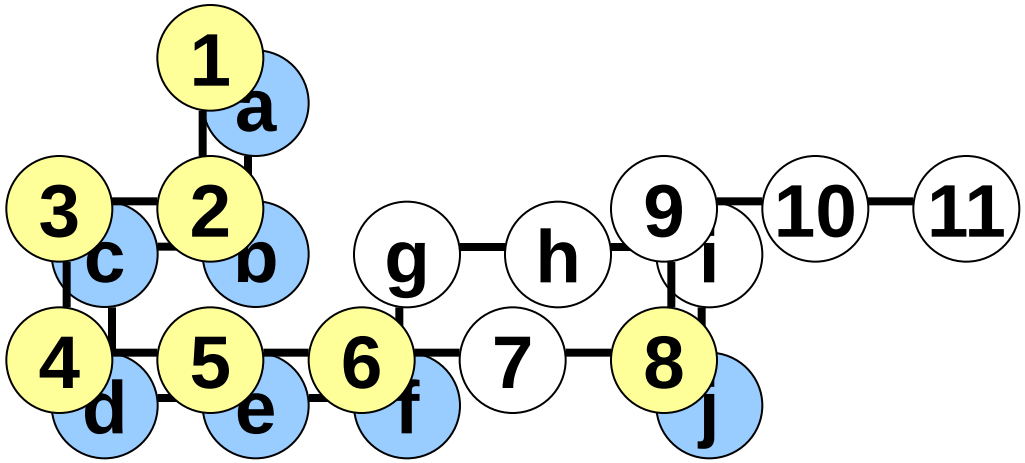
Resulted Aligned Pairs

Special Example



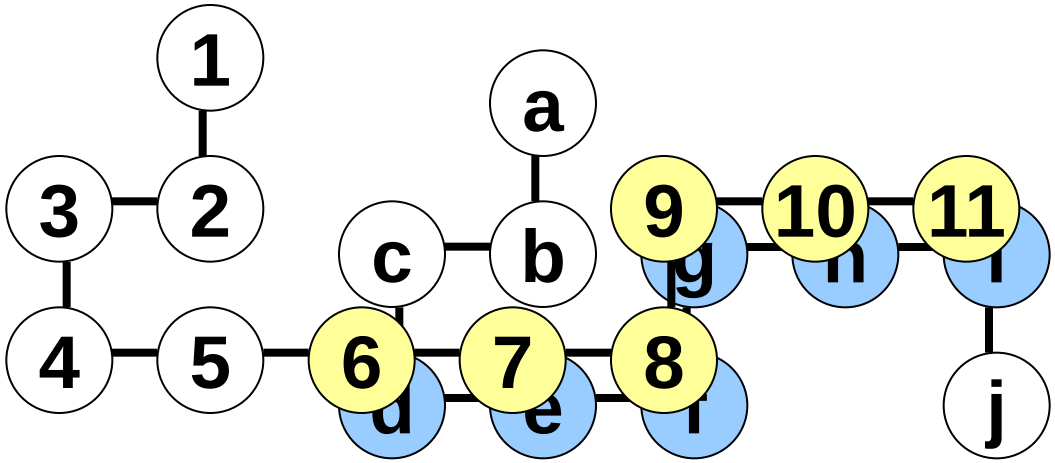


Special Example



1	2	3	4	5	6	7	-	-	-	8	9	10	11
a	b	c	d	e	f	-	g	h	i	j	-	-	-

Special Example



1	2	3	4	5	-	-	-	6	7	8	9	10	11	-
					a	b	c	d	e	f	g	h	i	j

Special Example

	a	b	c	d	e	f	g	h	i	j
1	7	5	6	6	5	5	4	3	2	2
2	4	7	7	6	7	5	4	4	3	2
3	4	4	7	8	7	6	5	5	3	3
4	2	5	5	7	4	4	6	5	4	4
5	3	7	5	7	7	4	6	5	4	4
6	5	5	4	3	6	8	7	5	5	4
7	2	3	2	2	3	7	5	5	5	3
8	1	1	1	3	2	1	4	5	5	7
9	2	2	3	2	2	2	1	7	9	6
10	2	1	2	2	1	1	1	1	7	6
11	1	1	1	1	1	2	1	1	1	3

Special Example

1 - a

1	2	3	4	5	6	7	-	-	-	8	9	10	11
a	b	c	d	e	f	-	g	h	i	j	-	-	-

7 - g

1	-	-	-	2	3	4	5	-	6	7	8	9	-	10	11
-	a	b	c	d	-	-	-	e	f	g	h	-	i	j	-

8 - h

1	2	3	4	-	-	5	6	-	-	-	7	8	9	10	11
-	-	-	-	a	b	c	-	d	e	f	g	h	i	j	-

9 - i

1	2	3	4	5	6	-	7	8	9	10	11
a	b	c	d	e	f	g	-	h	i	j	-

Special Example

	a	b	c	d	e	f	g	h	i	j
1	7	5	6	6	5	5	4	3	2	2
2	4	7	7	6	7	5	4	4	3	2
3	4	4	7	8	7	6	5	5	3	3
4	2	5	5	7	4	4	6	5	4	4
5	3	7	5	7	7	4	6	5	4	4
6	5	5	4	3	6	8	7	5	5	4
7	2	3	2	2	3	7	5	5	5	3
8	1	1	1	3	2	1	4	5	5	7
9	2	2	3	2	2	2	1	7	9	6
10	2	1	2	2	1	1	1	1	7	6
11	1	1	1	1	1	2	1	1	1	3

Special Example

1 - a

1	2	3	4	5	6	7	-	-	-	8	9	10	11
a	b	c	d	e	f	-	g	h	i	j	-	-	-

2 - b

1	2	3	4	5	6	7	-	-	-	8	9	10	11
a	b	c	d	e	f	-	g	h	i	j	-	-	-

6 - f

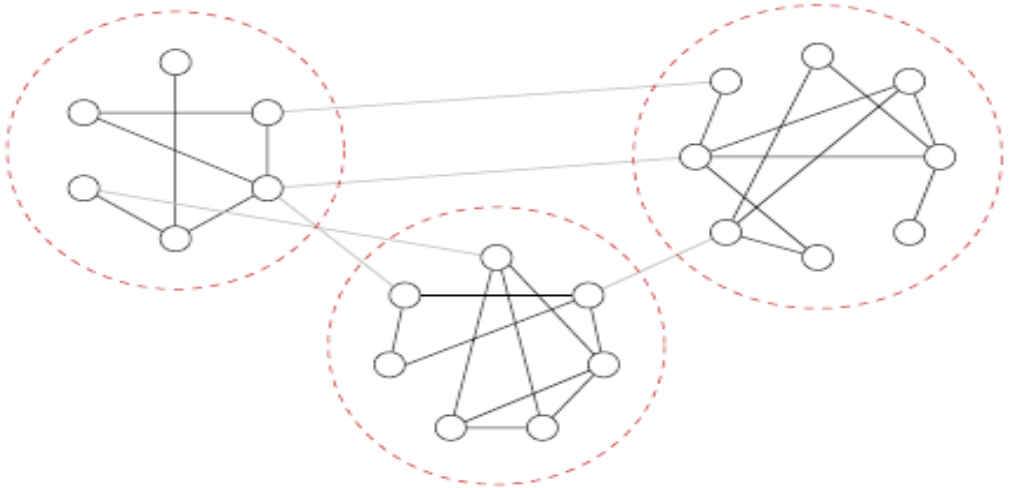
1	2	3	4	5	6	7	-	-	8	9	10	11
a	b	c	d	e	f	g	h	i	j	-	-	-

8 - j

1	2	3	4	5	6	7	-	-	-	8	9	10	11
a	b	c	d	e	f	-	g	h	i	j	-	-	-

SANA

- Structure Alignment by Neighborhood Alignment
  - Construct graph
  - Detect clusters

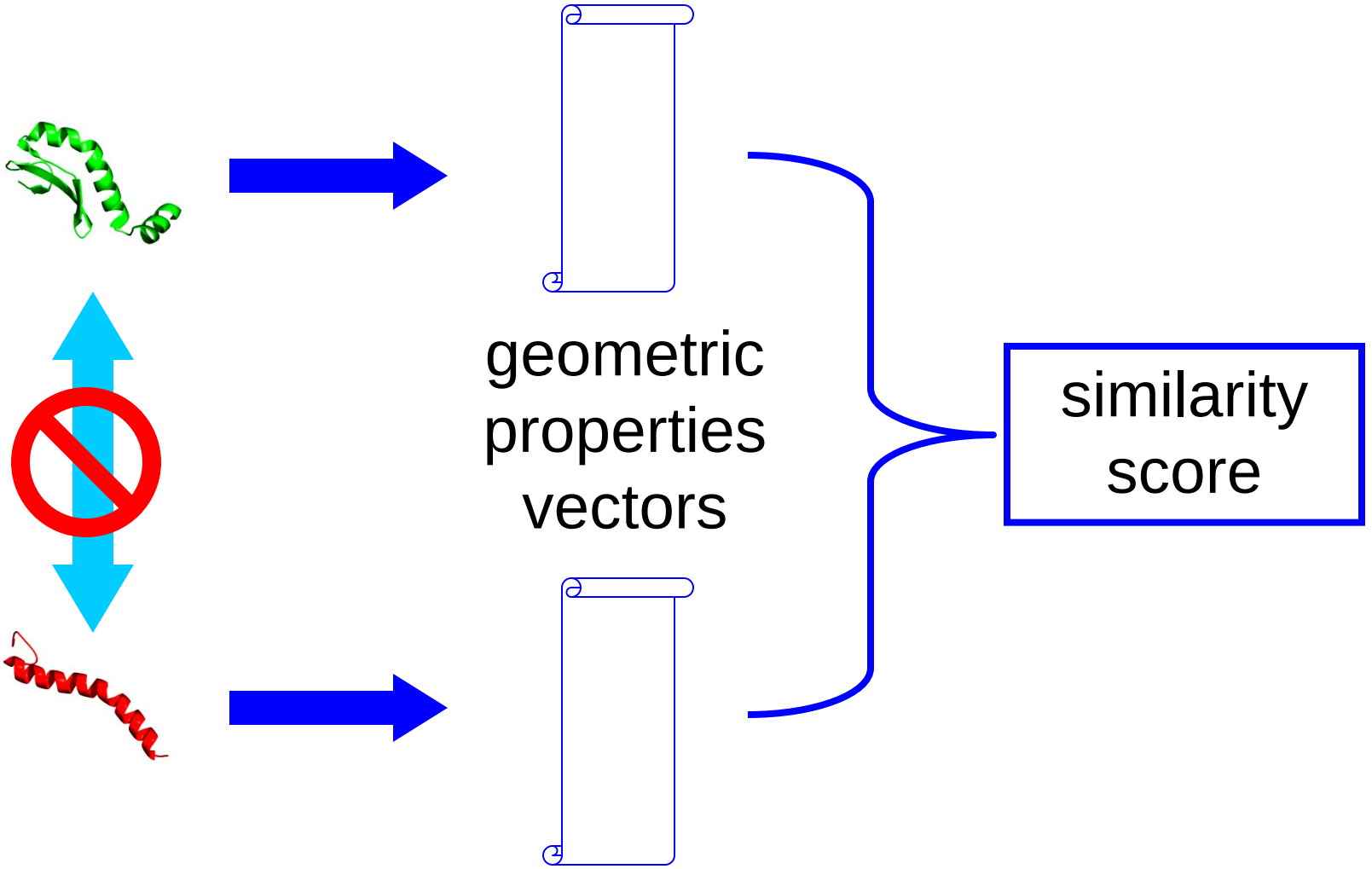


Amino Acids, 2010.

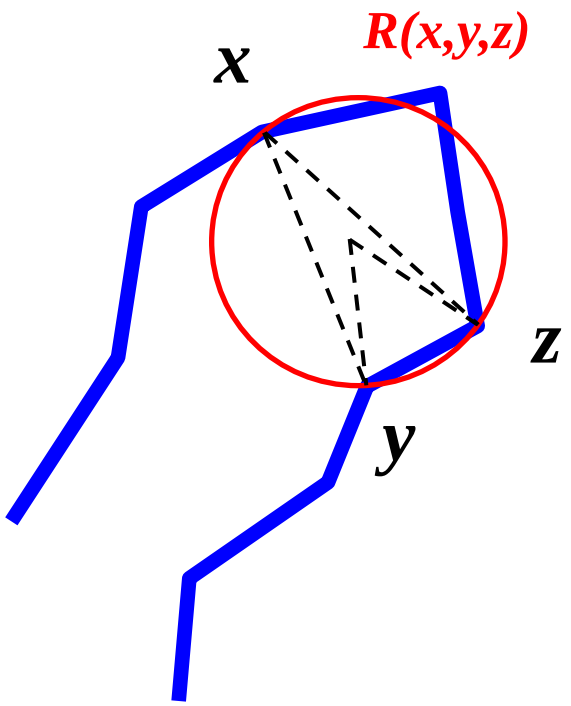
目录

- 结构比较
- 结构相似性
- 结构比对方法
- **结构比较方法**

蛋白质结构比较



曲率(Curvature)



Radius of Curvature

$$R(x,y,z)=\frac{d(x,y)}{2\left|\sin(\hat{z})\right|}$$

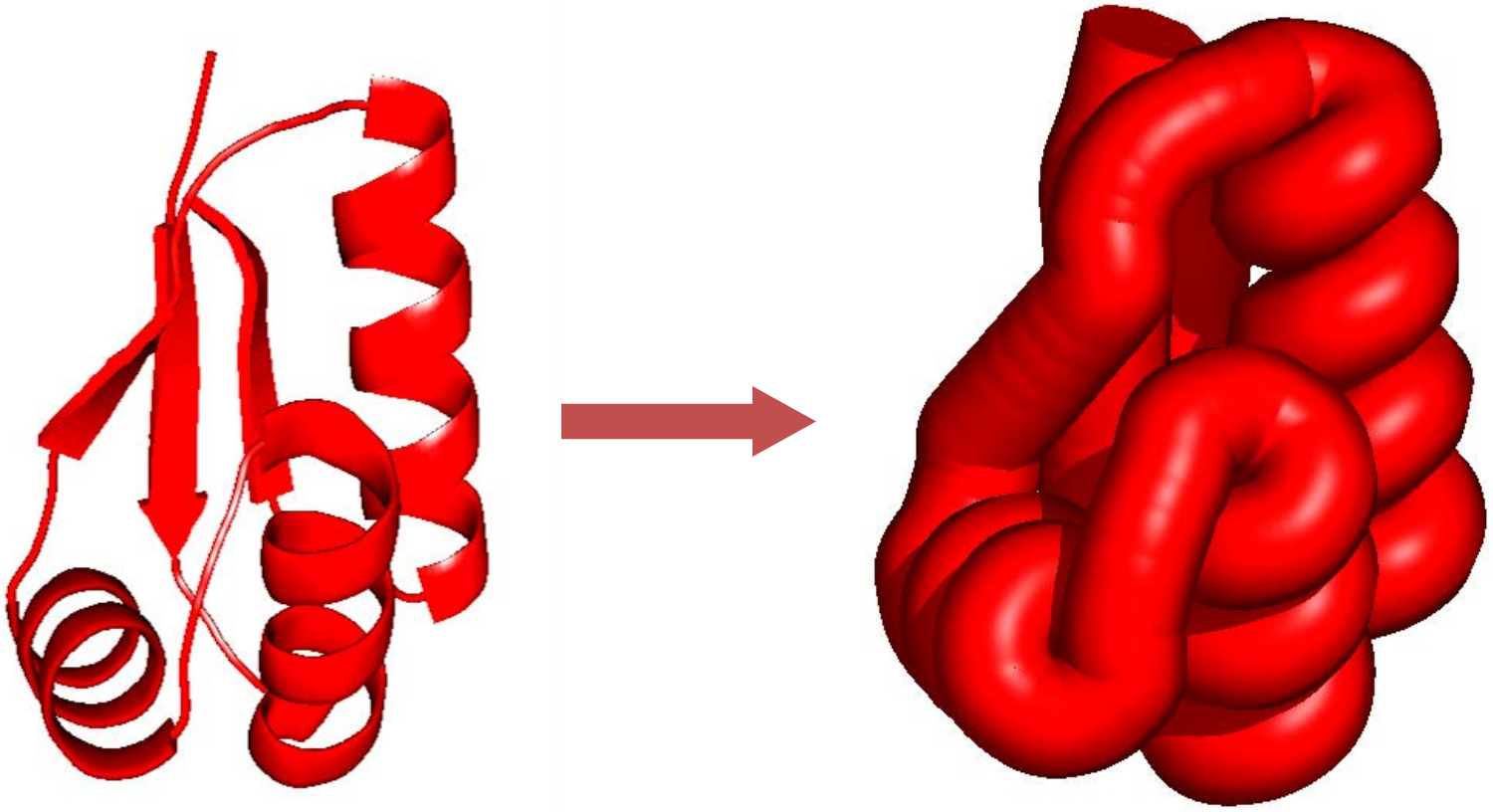
Global radius of curvature:

$$\rho(x)=\min_{(y,z)}\{R(x,y,z)\}$$

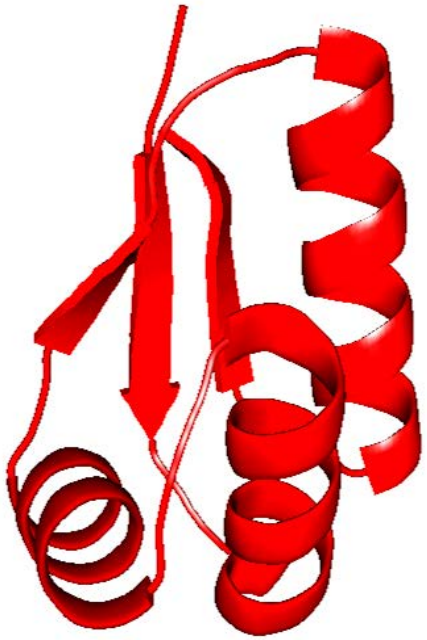
Thickness:

$$\Delta=\min_{x}\{\rho(x)\}$$

厚度(thickness)



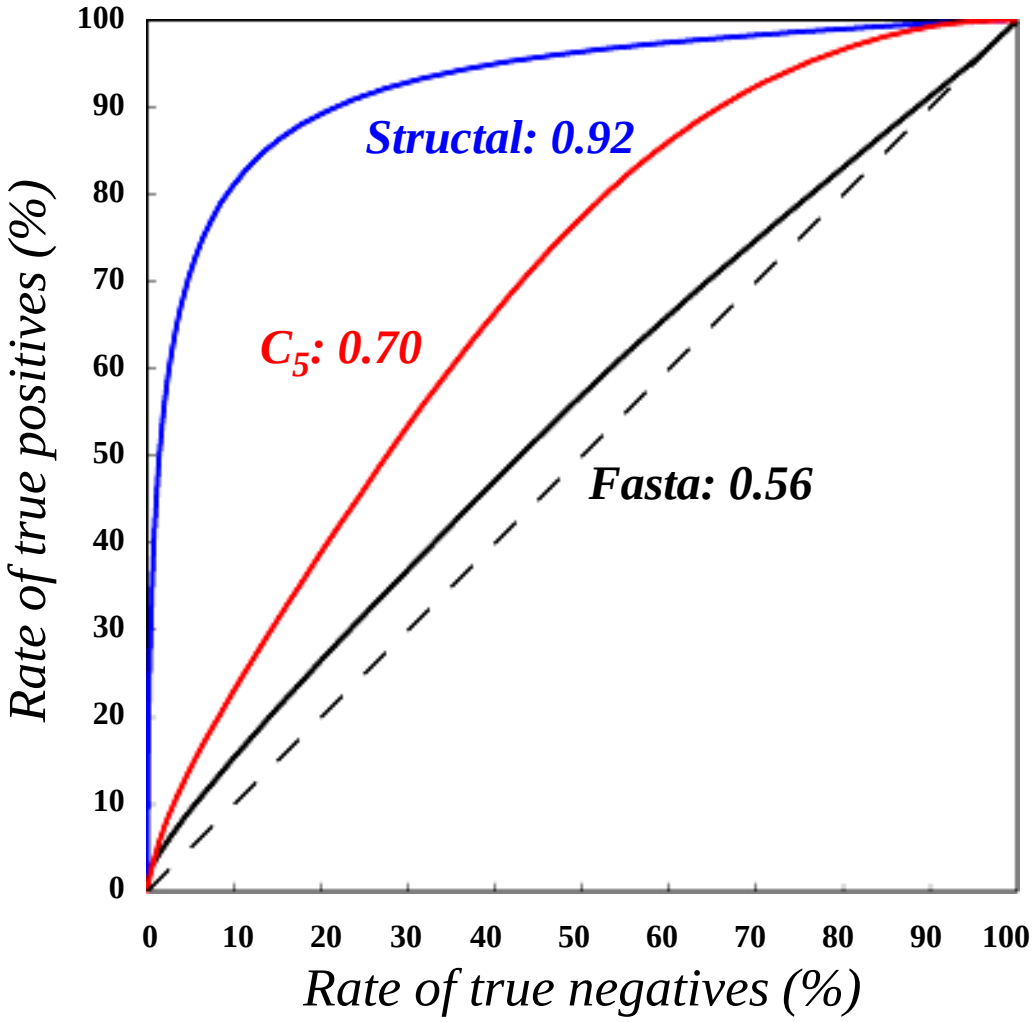
曲率特征向量



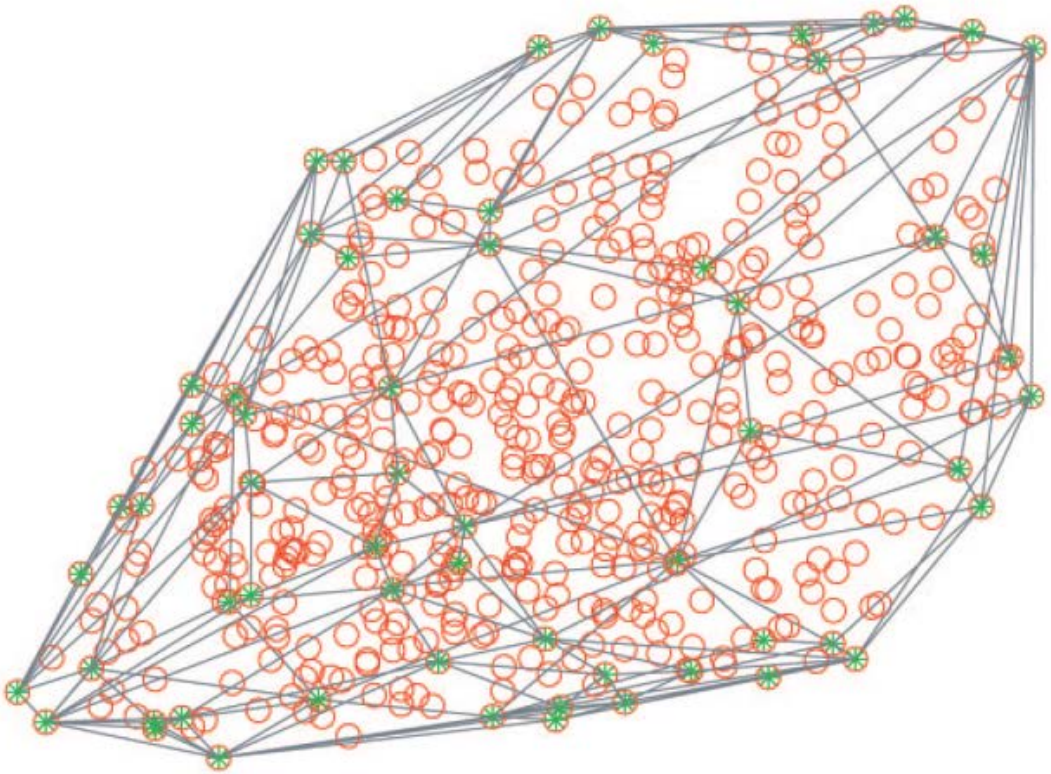
$$U_p = \left(\iiint \frac{1}{R(x,y,z)^p} dC_x dC_y dC_z \right)^{1/p}$$

$$C_5 = [U_1 \quad U_2 \quad U_3 \quad U_4 \quad U_5]$$

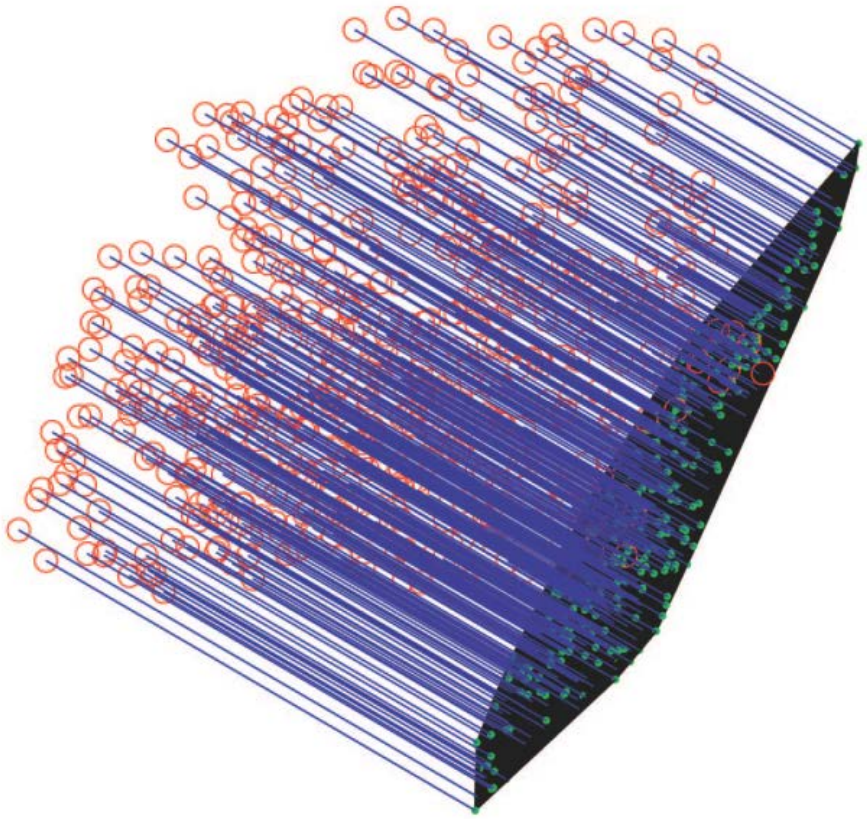
结果比较


Structal: 0.92
C_5: 0.70
Fasta: 0.56
Rate of true positives (%)
Rate of true negatives (%)
0
10
20
30
40
50
60
70
80
90
100

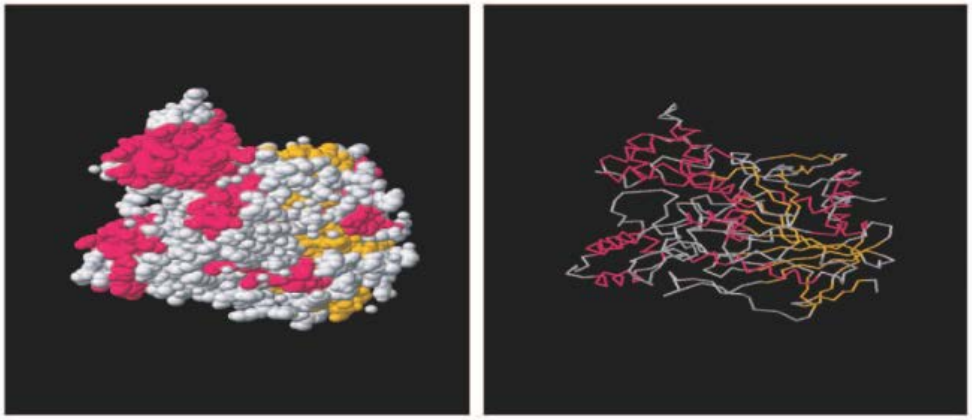
蛋白质的凸包（Convex Hull）



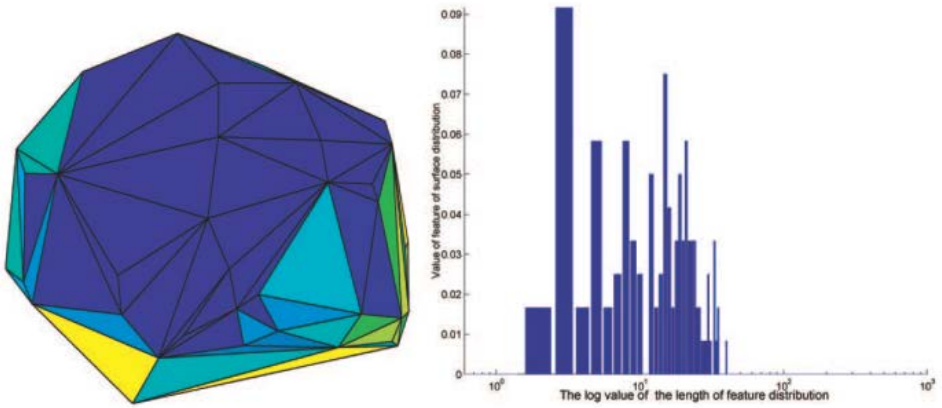
支撑面



Feature Sequence of Surface

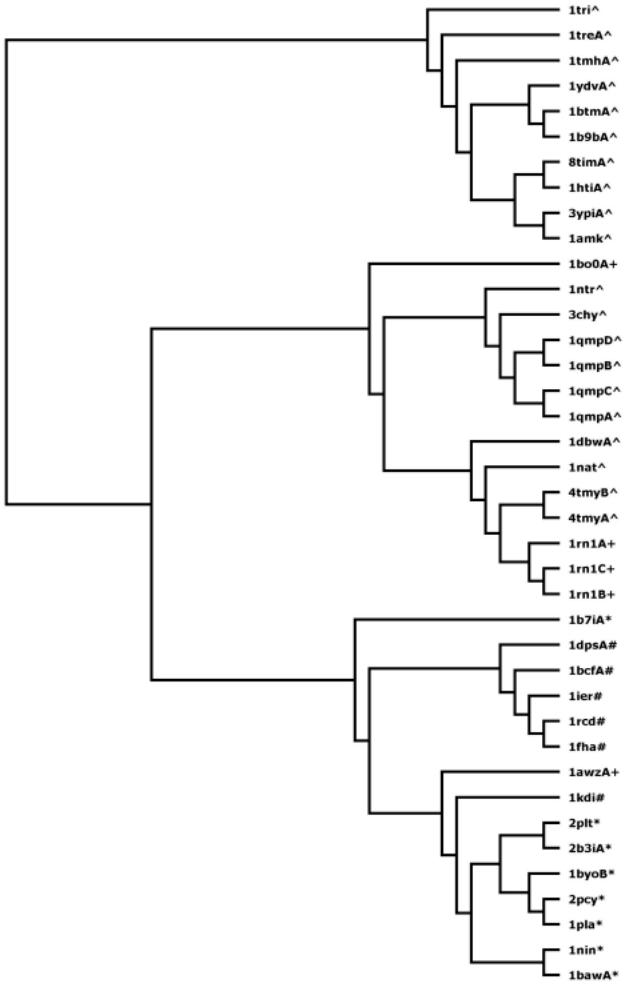


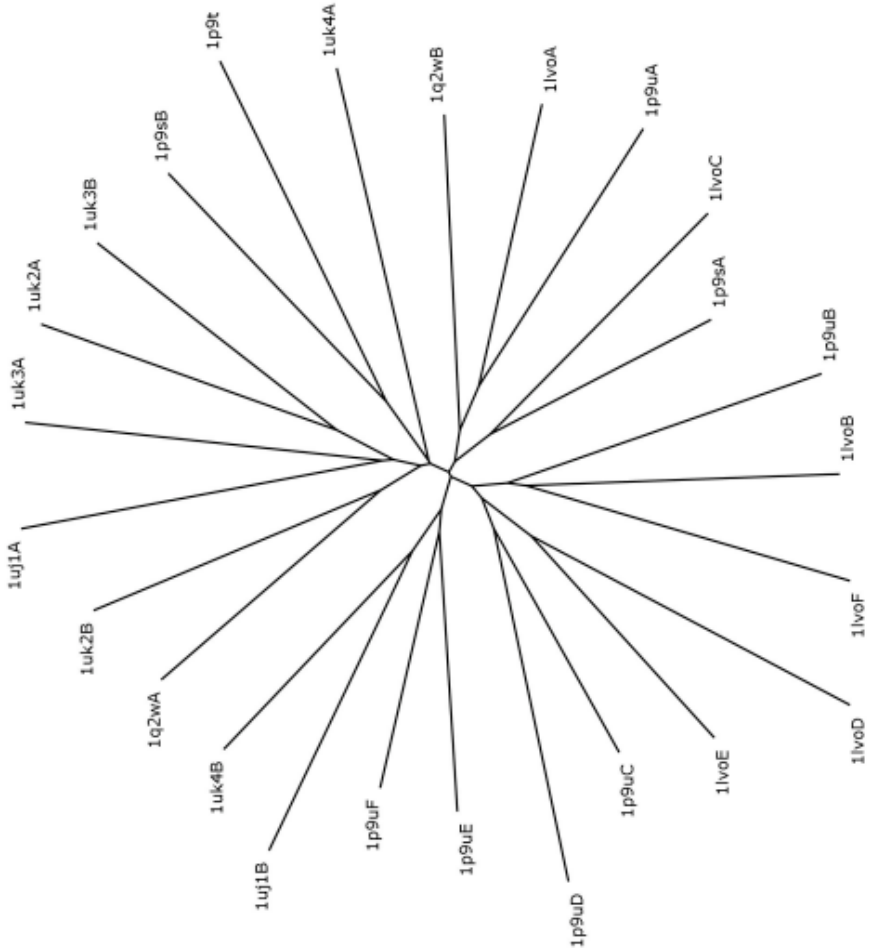
(a) (b)



(c) (d)

应用：结构分类


1tri^
1treA^
1tmhA^
1ydvA^
1btmA^
1b9bA^
8timA^
1htiA^
3ypiA^
1amk^
1bo0A+
1ntr^
3chy^
1qmpD^
1qmpB^
1qmpC^
1qmpA^
1dbwA^
1nat^
4tmyB^
4tmyA^
1rn1A+
1rn1C+
1rn1B+
1b7iA*
1dpsA#
1bcfA#
1ier#
1rcd#
1fha#
1awzA+
1kdi#
2plt*
2b3iA*
1byoB*
2pcy*
1pla*
1nin*
1bawA*


1p9t
1uk4A
1q2wB
1lvoA
1p9uA
1lvoC
1p9sA
1p9uB
1lvoB
1lvoF
1lvoD
1lvoE
1p9uC
1p9uD
1p9uE
1p9uF
1uj1B
1uk4B
1q2wA
1uk2B
1uj1A
1uk3A
1uk2A
1uk3B
1p9sB