Protein Structure Comparison by Alignment of Distance Matrices

李懿　1810043
王桂月1710077
邬晓彤1710174

1.引言（Introduction）

2.方法（Methods）

3.结果（Results）

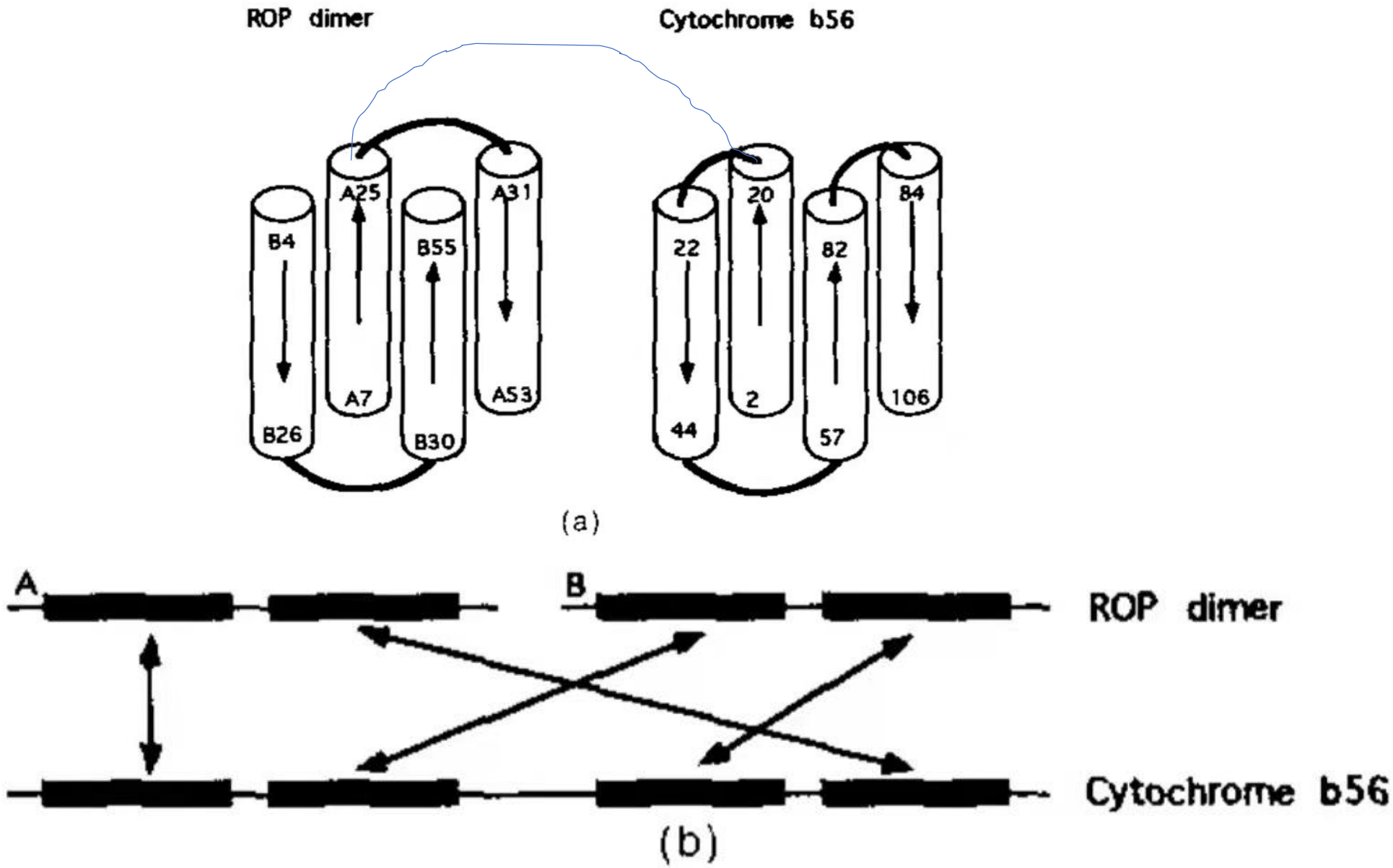
4.讨论（Discussion）

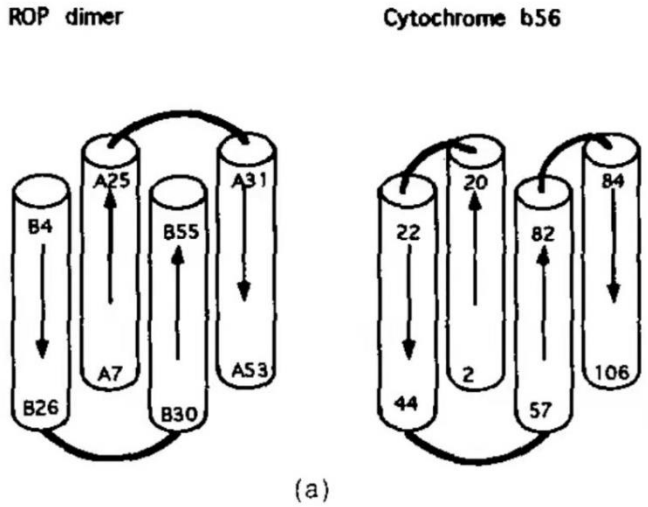
一.引言

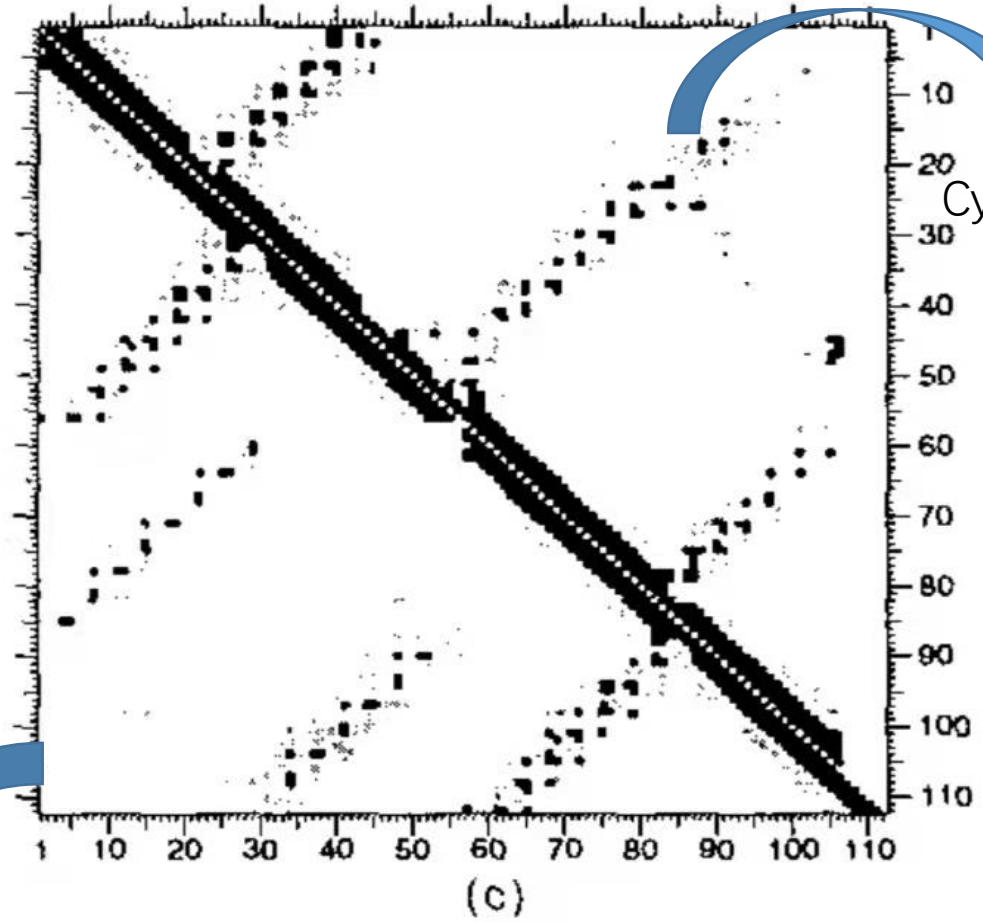
1.序列差异很大的蛋白质也可能有非常相似的结构

2.相似的3D结构有相似的残基间距离

3.通过距离矩阵比较蛋白质的3D结构


ROP dimer
Cytochrome b56
B4
A25
B55
A31
B26
A7
B30
A53
22
20
82
84
44
2
57
106
(a)
A
B
ROP dimer
Cytochrome b56
(b)





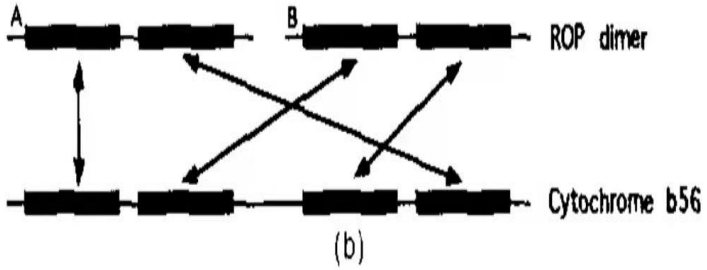
Cytochrome b56

ROP dimer

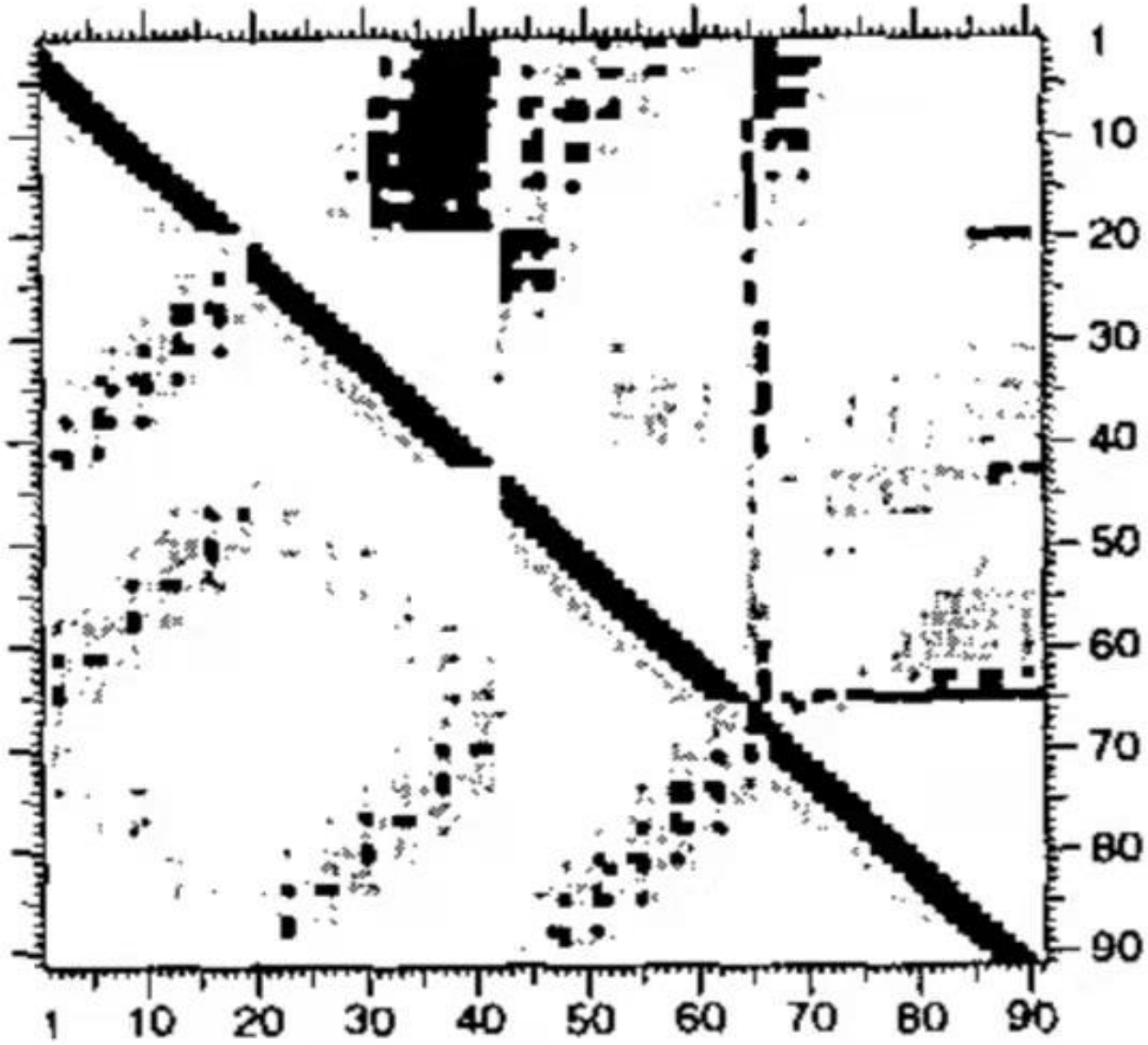
黑色:<8埃米
深灰色:8~12埃米
浅灰色:12~16埃米

1埃米=10^{-10}米

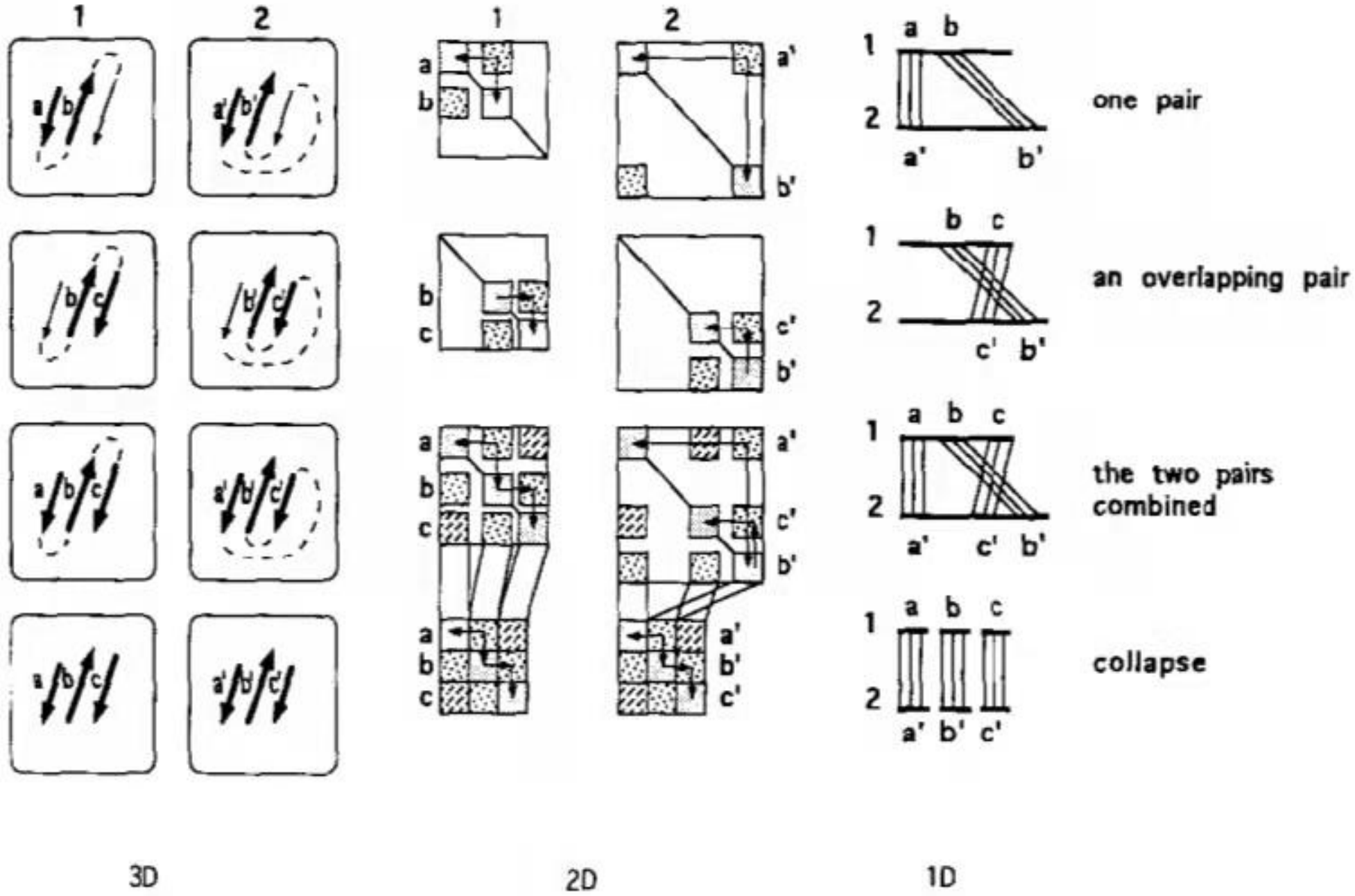
对齐后的距离差异矩阵（上三角）



白色:<1埃米
黑色:>4埃米



(d)


1
2
a
b
a'
b'
c
c'
one pair
an overlapping pair
the two pairs combined
collapse
3D
2D
1D

二.方法

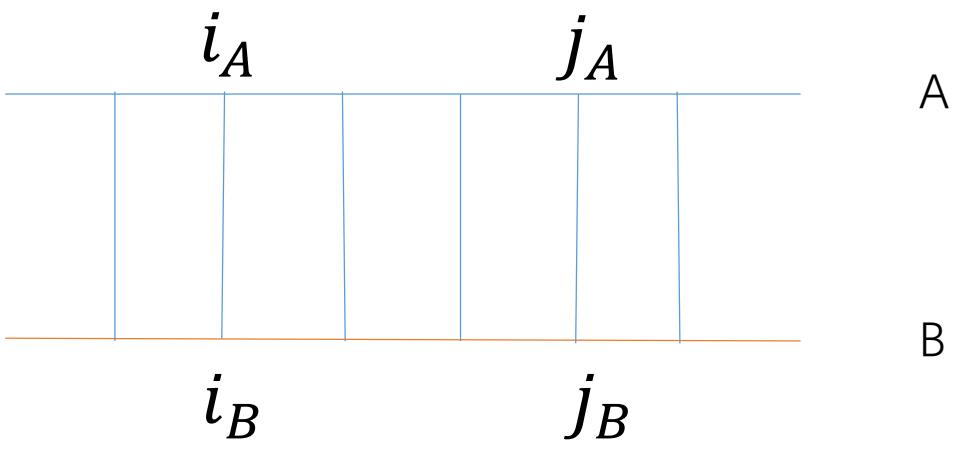
1.定义

$$S=\sum_{i=1}^{L}\sum_{j=1}^{L}\emptyset(i,j)$$

$$i=(i_A,\ i_B)$$

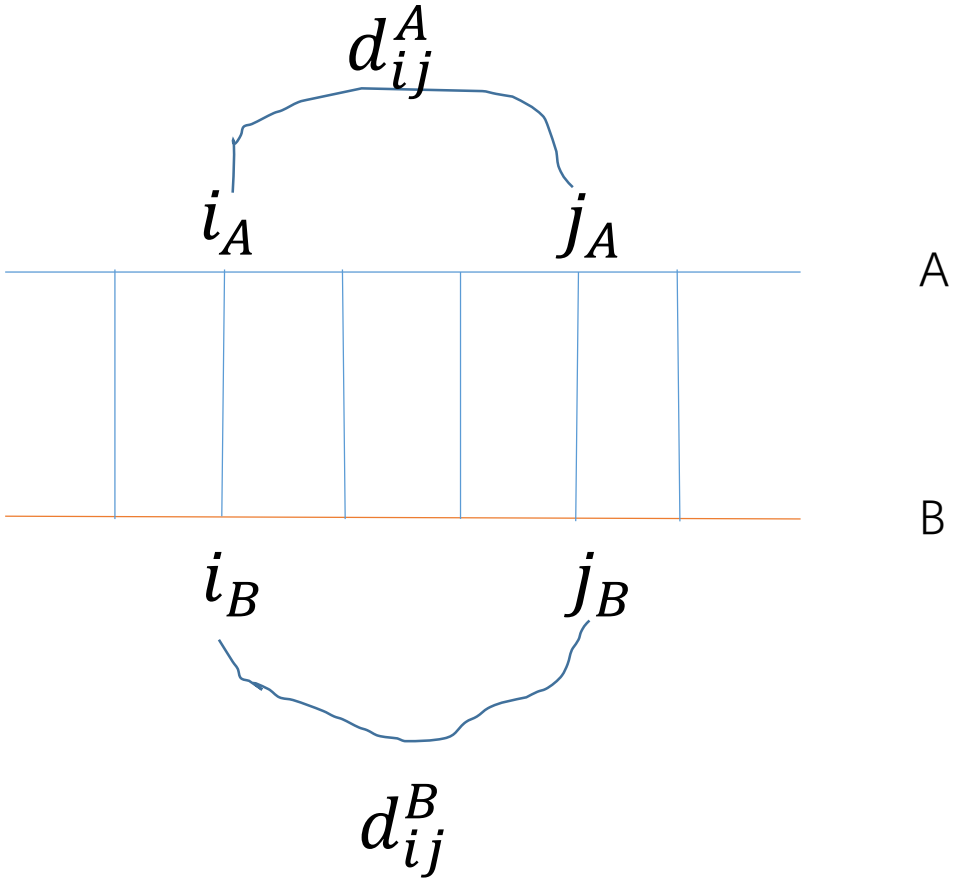
$$j=(j_A,\ j_B)$$

$\emptyset(i,j)$: 相似性度量



蛋白质A和B

刚性相似度得分（Rigid similarity score）



$$\emptyset^{R}(\mathrm{i},\mathrm{j}) = \theta^{R}\text{-}\left|d_{ij}^{A} - d_{ij}^{B}\right|$$

θ:零级相似度

θ^{R}=1.5埃米

弹性相似度得分(Elastic similarity score)

$$\emptyset^{E}(\mathrm{i},\mathrm{j})=\begin{cases}\left(\theta^{E}-\frac{\left|d_{ij}^{A}-d_{ij}^{B}\right|}{d_{ij}^{*}}\right)w\left(d_{ij}^{*}\right), & i\neq j\\ \theta^{E} & \end{cases}$$

d_{ij}^{*}是d_{ij}^{A}和d_{ij}^{B}的平均值

θ^{E}=0.2

W(r)=exp(-r^{2}/α^{2}),
$\alpha=20$埃米

2.Building/Refining Data Representation

Reduce within distance matrices

Reduce in pair list until:

- 1.The mean intra-pattern distance reaches 25埃米
- 2.80000 contact pairs with a positive similarity score are added

Table 1

Simplifying combinatorial complexity in the comparison of hen egg-white lysozyme (1lyz) with T4 lysozyme (2lzm)

A. Distance matrices	
1lyz	
No. of overlapping hexapeptides	124
Total no. of contact patterns	7626
No. of contact patterns in reduced distance matrix	5332
2lzm	
No. of overlapping hexapeptides	159
Total no. of contact patterns	12,561
No. of contact patterns in reduced distance matrix	4709
B. *Pair list*	
Total no. of pairs of contact patterns	96×10^6
Total no. of pairs of contact patterns after reduction	71×10^6
No. of checks by filters on row/column sums†	9×10^6
No. of residue-by-residue similarity score calculations	2×10^5
No. of kept pairs of contact patterns after ranking by score	4×10^4

3.Assembly of aligment

$$p = e^{(\beta*(s'-s))}$$

S:old score

s':new score

β:parameter

Results

1 算法鲁棒性的验证（robustness 坚固性，不易改变）

2 对齐的质量

3 蛋白质结构的所有对齐

算法鲁棒性的验证

Table 1
Simplifying combinatorial complexity in the comparison of hen egg-white lysozyme (1lyz) with T4 lysozyme (2lzm)

A. Distance matrices	
1lyz	
No. of overlapping hexapeptides	124
Total no. of contact patterns	7626
No. of contact patterns in reduced distance matrix	5332
2lzm	
No. of overlapping hexapeptides	159
Total no. of contact patterns	12,561
No. of contact patterns in reduced distance matrix	4709
B. Pair list	
Total no. of pairs of contact patterns	96×10^6
Total no. of pairs of contact patterns after reduction	71×10^6
No. of checks by filters on row/column sums†	9×10^6
No. of residue-by-residue similarity score calculations	2×10^5
No. of kept pairs of contact patterns after ranking by score	4×10^4
C. Monte Carlo optimization	
Screening	
No. of parallel trajectories	80
No. of expansion/trimming cycles‡	1
No. of kept alignments after ranking by score	10
Optimization of divergent alignments	
No. of parallel trajectories	10
No. of expansion/trimming cycles‡	80
No. of kept alignments after ranking by score	1
Refinement of best alignment	
No. of parallel trajectories	10
No. of expansion/trimming cycles‡	40
No. of kept alignments after ranking by score	1

At each step of the algorithm, the search tree is heavily pruned

Table 2
Seed test

	Correct alignment (no. of runs)	Incorrect alignment (no. of runs)
T4 lysozyme (2lzm) – hen egg-white lysozyme (1lyz)		
Correct seed	4	0
Incorrect seed	2	74
Colicin A (1colA) – ark hemoglobin (1sdhA)		
Correct seed	6	0
Incorrect seed	16	86

As a test of the radius of convergence of the algorithm, the build-up was followed from the initial seed alignment to the final optimized and refined alignment. The alignment with the highest score (and close variants) was classified as correct, and seeds were classified as correct if they overlapped with the correct full alignment. Optimization of the similarity score can lead to the correct alignment even though the alignment is initialized from an incorrect seed.

Table 3
Internal symmetry of a $(\beta\alpha)_8$ barrel

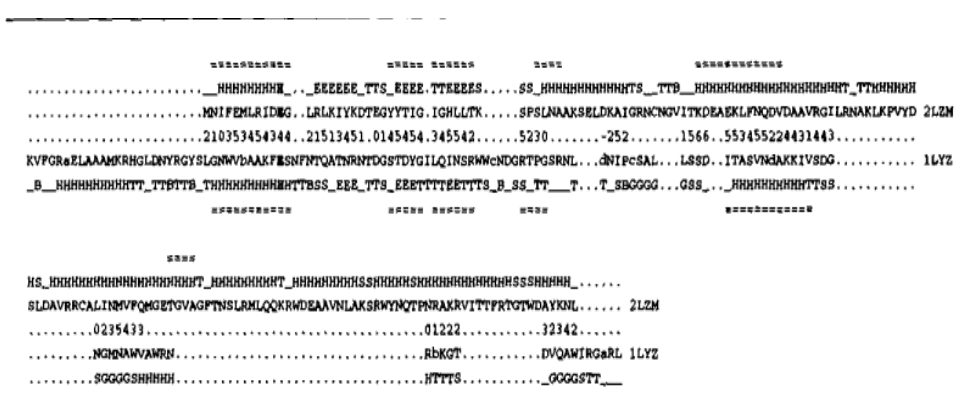
$(\beta\alpha)$ units 1-2-3-4-5-6-7-8 aligned with	Similarity score	No. of equivalenced residues	r.m.s.d. (Å)
5-6-7-8-1-2-3-4	1194	171	2·9
7-8-1-2-3-4-5-6	1101	177	3·4
8-1-2-3-4-5-6-7	989	177	3·2
4-5-6-7-8-1-2-3	970	167	3·0
6-7-8-1-2-3-4-5	944	170	3·2
3-4-5-6-7-8-1-2	818	169	3·7
2-3-4-5-6-7-8-1	757	159	3·4

表1 为了检验蒙特卡罗验证法的重复性，使用不同的随机数重复100次。通过对T4溶菌酶与蛋清溶菌酶的比较，发现总共有2%的测试运行找到了全局最优值和 94%的运行进入第二优最优值。

表2 为了测试算法的收敛半径，从所有测试组中生成了完全优化的对齐。绝大多数对不正确的最优值有反应。大多数对齐仍然停留在局部最优中，但在一条路径中，可以从不正确到达最终的对齐，这说明对起点并不敏感。

表3 检测具有结构意义的多重优化方案。将次优解方案打印出来，通过色氨酸合成酶与自身的对比，得到了与预期相同的7个循环排列序列。这就证明了算法的鲁棒性。

对齐的质量



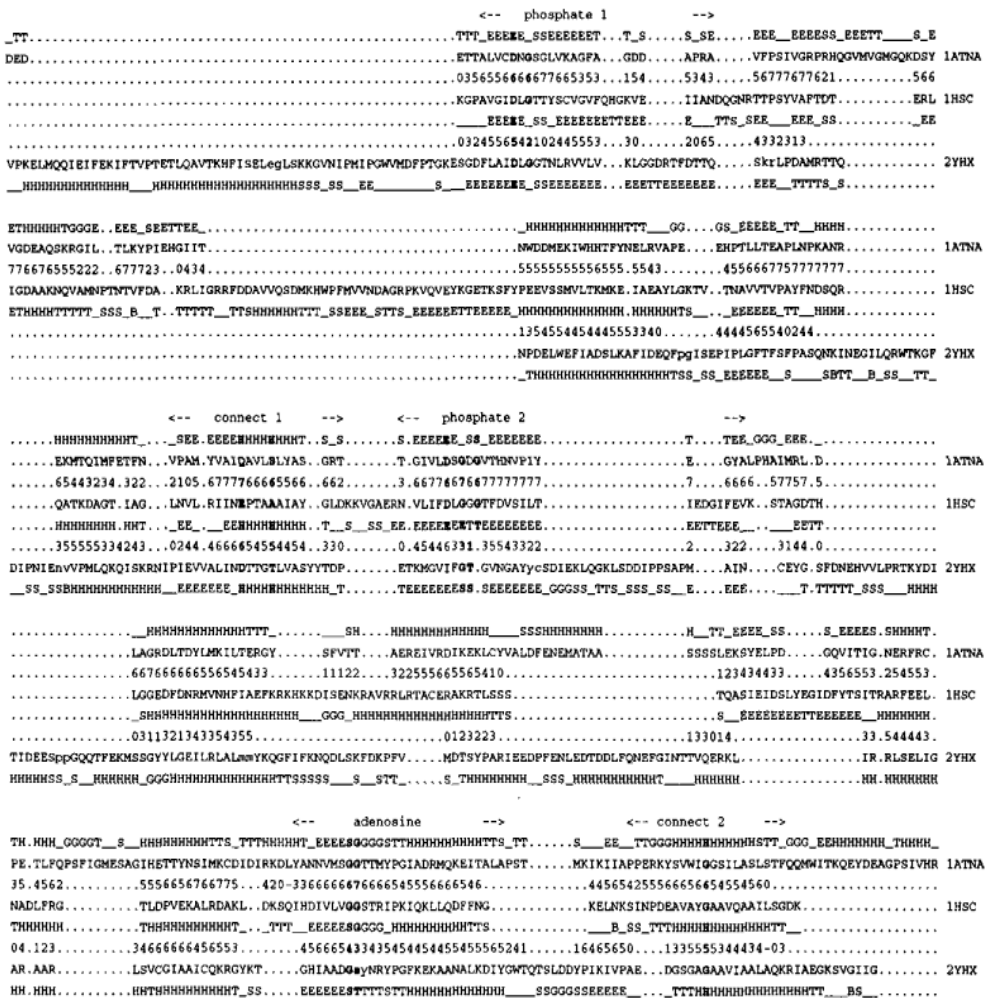
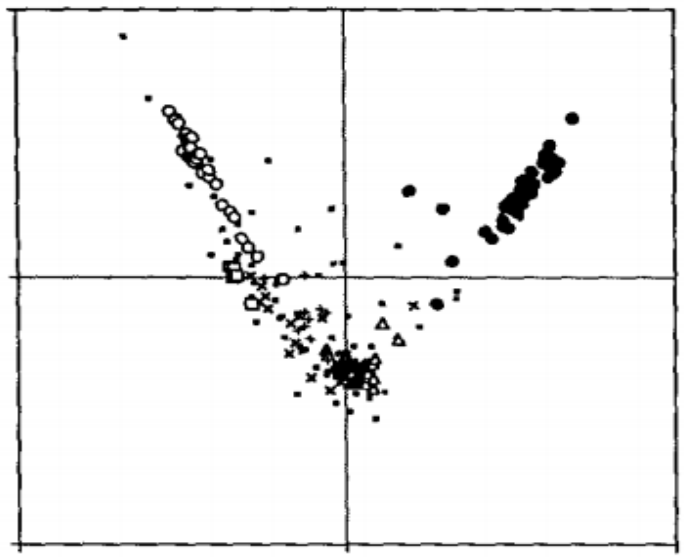


图1（左） 验证准确性。保守的功能残基提供了锚点，通过它可以验证不同蛋白家族成员之间的结构对齐的准确性。

图2 （右） 是域间运动的检测。肌动蛋白、热休克蛋白 hsp70 和己糖激酶是三种功能多样的蛋白，具有 一个共同的atp酶结构域。 atp结合位点位于两个子结构域之间的裂缝中，由螺旋-螺旋接触点形成的铰链连接。肌动蛋白和热休 克蛋白的晶体结构为“封闭”构象，而己糖激酶的晶体结构为“开放”构象。我们比较封闭形式和开放形式的每个域，比较链的长度，我们可以区分这三种蛋白质。然后通过减少相似性阈值，我们可以将重点转移到核心方面，从而提高对齐的质量。

蛋白质结构的所有对齐



α-helical proteins:

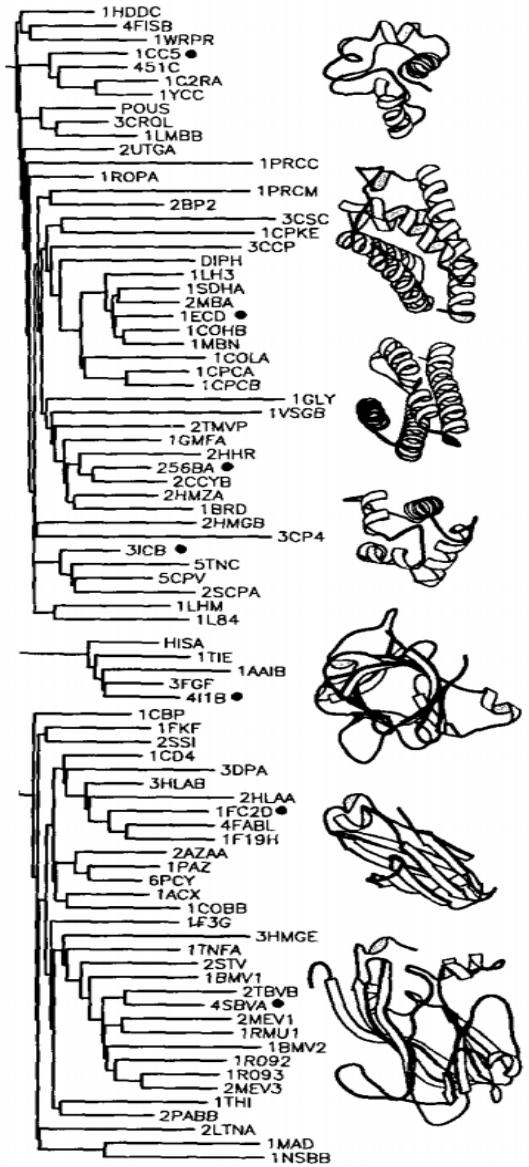
1HDDC	engrailed homeodomain	1COHB	hemoglobin
4FISB	factor for inversion stimulation	1MBN	myoglobin
1WRPR	Trp repressor	1COLA	colicin A
1CC5	cytochrome c5	1CPCA	C-phycocyanin
451C	cytochrome c551	1CPCB	C-phycocyanin
1C2RA	cytochrome c2	1GLY	glucoamylase
1YCC	cytochrome c	1VSGB	variant surface glycoprotein
POUS	POU-specific domain	2TMVP	tobacco mosaic virus
3CROL	434cro	1GMFA	growth factor
1LMBB	lambda repressor	2HHRA	human growth hormone
2UTGA	uteroglobin	256BA	cytochrome b562
1PRCC	photosynthetic reaction centre	2CCYB	cytochrome c'
1ROPA	ROP (repressor of primer) protein	2HMZA	hemerythrin
1PRCM	photosynthetic reaction centre	1BRD	bacteriorhodopsin
2BP2	phospholipase	2HMGB	hemagglutinin
3CSC	citrate synthase	3CP4	cytochrome P450 CAM
1CPKE	cAMP-dependent protein kinase	3ICB	intestinal calcium-binding protein
3CCP	cytochrome c peroxidase	5TNC	troponin C
DIPH	diphtheria toxin	5CPV	parvalbumin B
1LH3	leghemoglobin	2SCPA	sarcoplasmic calcium-binding protein
1SDHA	hemoglobin	2LHM	human lysozyme
2MBA	myoglobin	1L84	T4 lysozyme
1ECD	erythrocruorin		

β-trefoils:

HISA	hisactophilin	3FGF	fibroblast growth factor
1TIE	trypsin inhibitor	4I1B	interleukin 1-beta
1AAIB	ricin B chain		

antiparallel β-barrels:

1CBP	cucumber basic protein	1TNFA	tumor necrosis factor
1FKF	FK506 binding protein	2STV	coat protein of satellite tobacco necrosis virus
2SSI	subtilisin inhibitor	1BMV1	coat protein of bean pod mottle virus
1CD4	T-cell surface glycoprotein	2TBVB	coat protein of tomato bushy stunt virus
3DPA	papD protein	4SBVA	coat protein of southern bean mosaic virus
3HLAB	class I histocompatibility antigen	2MEV1	coat protein of mengovirus
2HLAA	class I histocompatibility antigen	1RMU1	coat protein of rhinovirus
1FC2D	immunoglobulin	1BMV2	coat protein of bean pod mottle virus
4FABL	immunoglobulin	1R092	coat protein of rhinovirus 14
1F19H	immunoglobulin	1R093	coat protein of rhinovirus 14
2AZAA	azurin	2MEV3	coat protein of mengovirus
1PAZ	pseudoazurin	1THI	thaumatin
6PCY	plastocyanin	2PABB	prealbumin
1ACX	actinoxanthin	2LTNA	lectin
1COBB	superoxide dismutase	1MAD	methylamine dehydrogenase
1F3G	phosphocarrier III	1NSBB	neuraminidase sialidase
3HMGE	hemagglutinin		



对225种具有代表性的蛋白质进行了全结构的比较，总共25200对比较。我们从大量对比中得到了四种类型的结果：相似的三维折叠中的不同拓扑结构、 结构家族、新的结构相似性，以及对序列结构模式的观察。

蛋白质的拓扑结构有的相同，有的不同，但是二维的排列却是惊人的相似。全对全搜索的结果体现了蛋白质的相似性。，这些相似性 可以用来在高维空间中定位每种结构类型。蛋白质折叠的普遍分布可以分成两类：树状和集群。并且，未来可以通过程序将蛋白质的亚结构和折叠域自动分类。

左图为集群　　右图为树状

Thank you

小组分工：

王桂月1710077 ：负责introduction和methods

李懿1810043：负责results和discussion