生物信息学

第十三章 结构生物信息学

蛋白质功能多样性

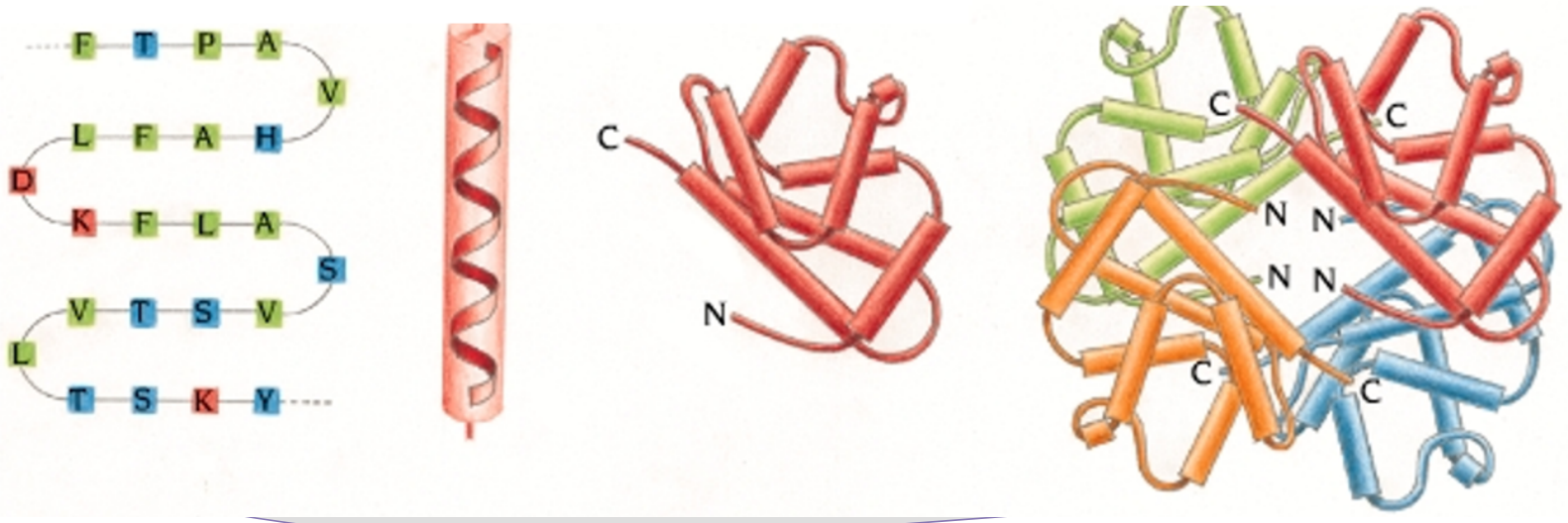
- *蛋白质的结构* – 主要由一级序列所决定
- *蛋白质的功能* – 主要由三级结构所决定

- 蛋白质功能多样性
  - 由分子结构的多样性和复杂性决定
  - 蛋白质的高级结构由蛋白质一级结构决定
  - 受到所处溶液环境影响
  - 蛋白质的功能主要由蛋白质的三级、四级结构所决定

蛋白质结构类型

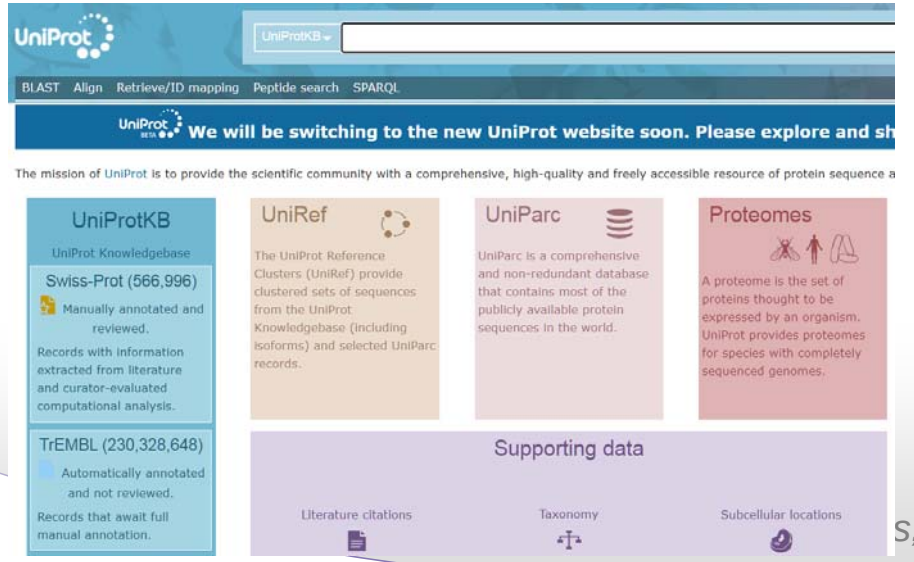
- 1952年，丹麦Kaj Ulrik Linderstrøm-Lang提出
  - 一级结构（primary structure）
  - 二级结构（secondary structure）
  - 超二级结构（super secondary structure）、模体（motif）、结构域（domain）
  - 三级结构（tertiary structure）
  - 四级结构（quaternary structure）：蛋白质复合物

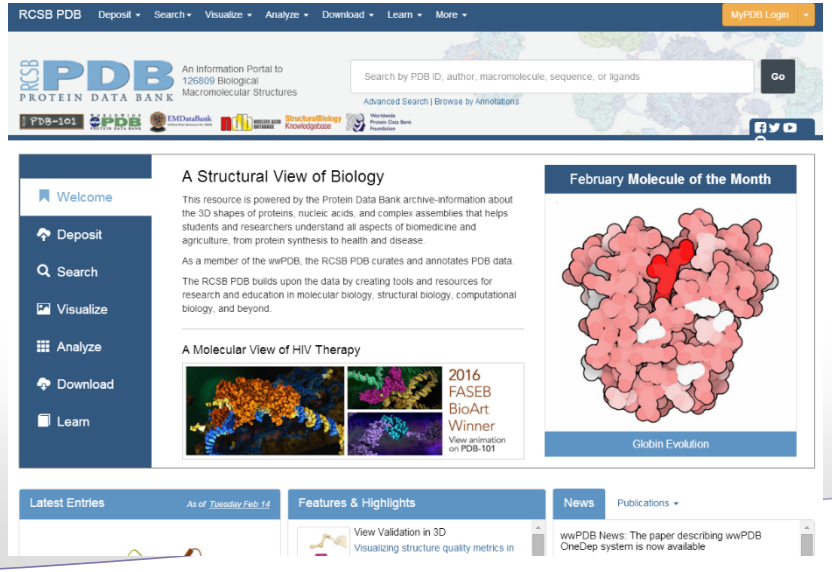
一级　　二级　　三级　　四级



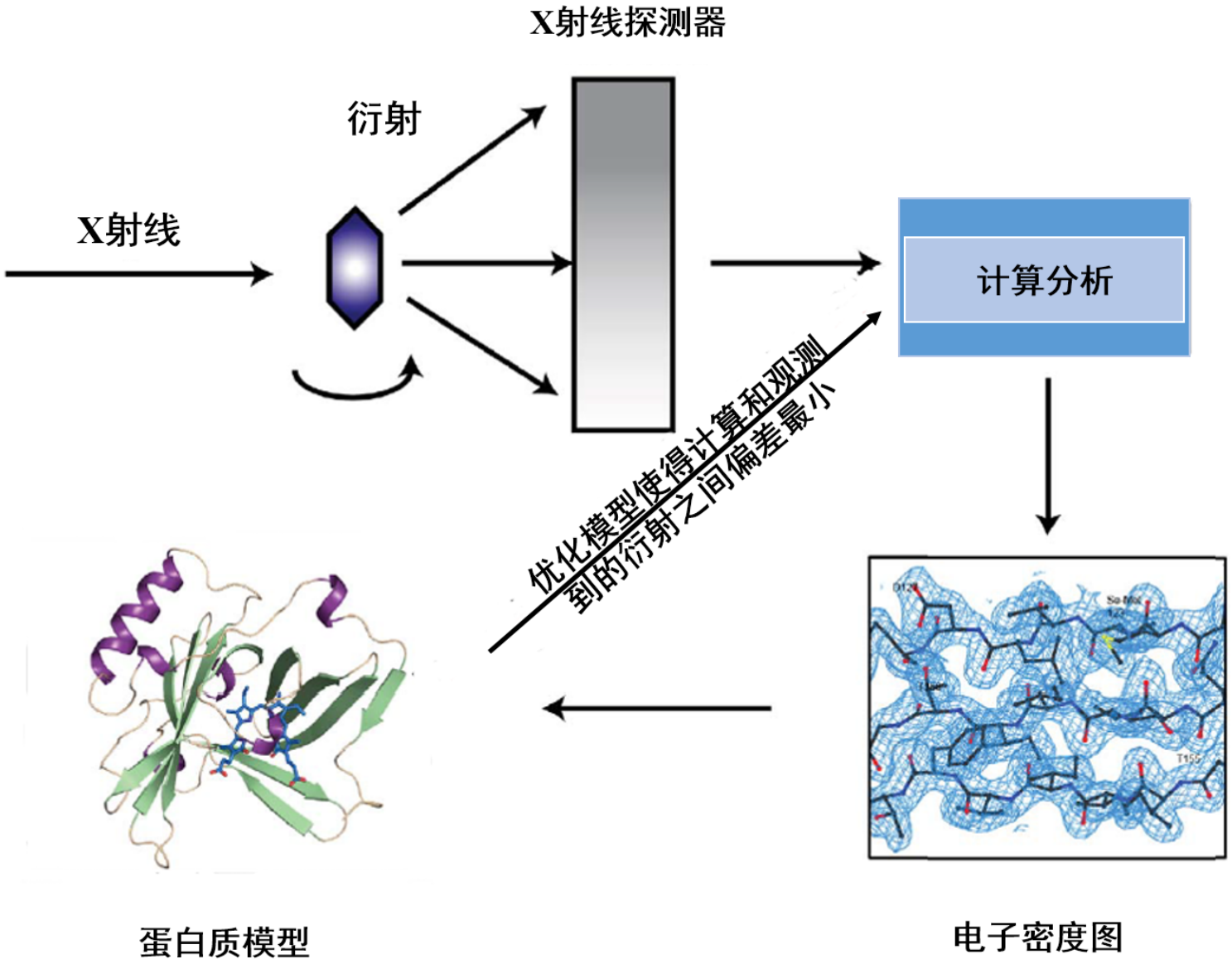
蛋白质序列与结构

- 蛋白质序列与结构的数据量比较（2022-05）
- UniProt：230,895,644条蛋白质序列
  - Swiss-Prot子库：566,996条手工注释的蛋白质序列
  - TrEMBL子库：230,328,648条计算机自动注释的蛋白质序列
- PDB：190,639个生物大分子结构

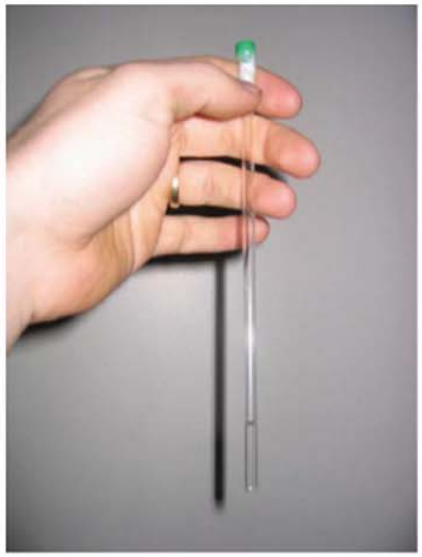


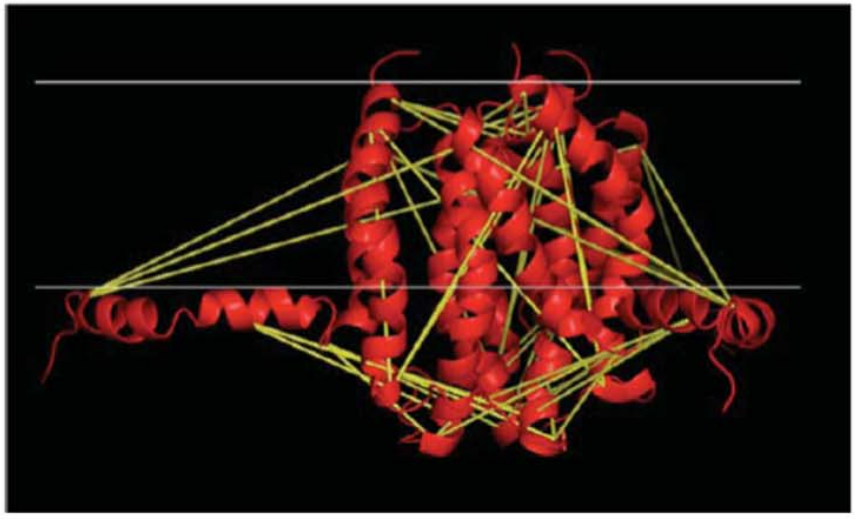


实验解析方法：X射线衍射



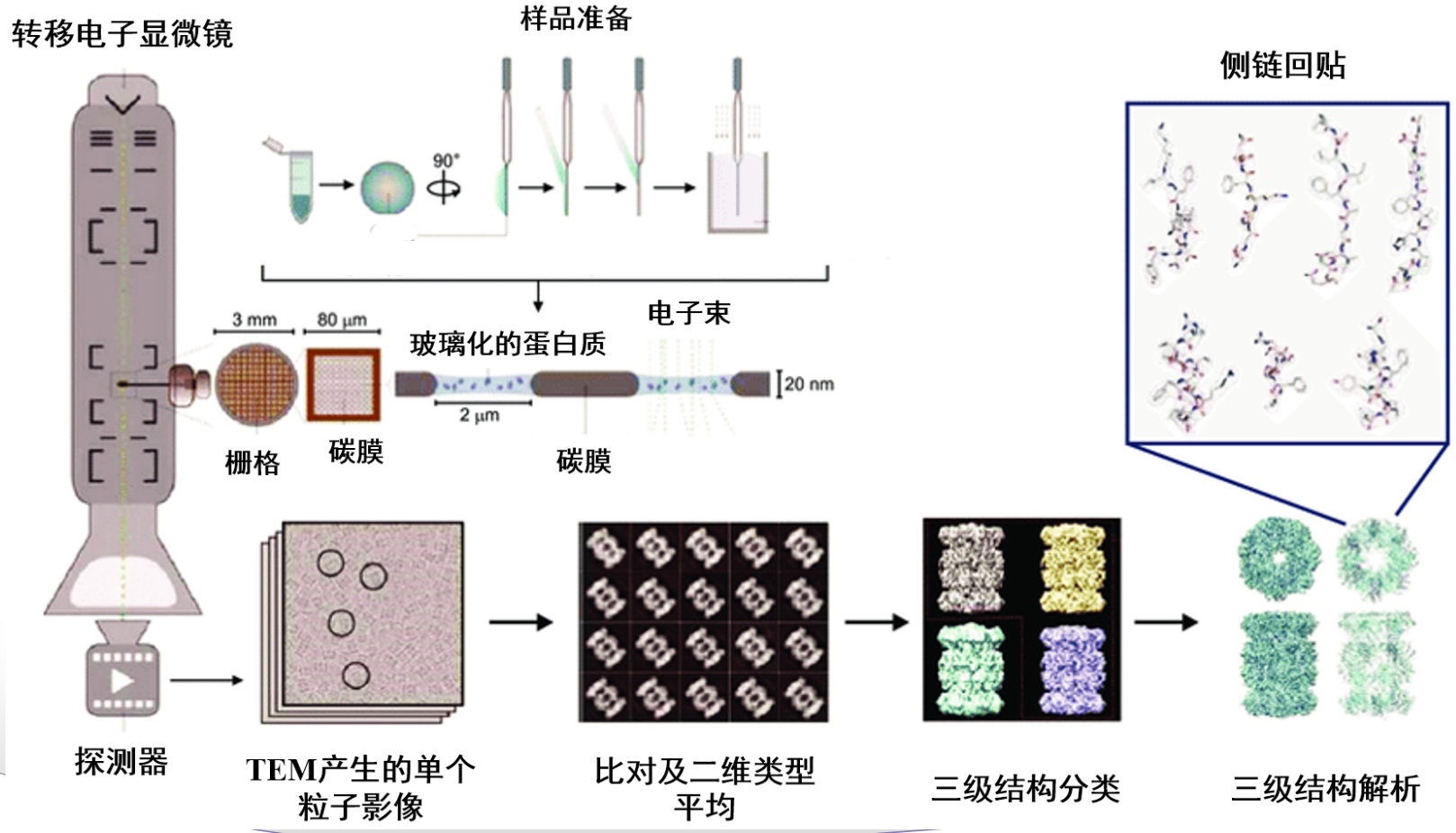
核磁共振





冷冻电子显微镜

- 能够很好地分辨蛋白质结构的表面
- 确定蛋白质内部的精细结构较为困难



蛋白质结构预测

- 蛋白质结构预测的必要性
  - 实验测定的速度远远赶不上蛋白质序列解析的速度
  - 结构测定的实验技术很难达到较高的通量
  - 减少两者之间的差距：发展生物信息学算法，从序列出发预测蛋白质结构
- 蛋白质结构预测的基本原理
  - Anfinsen原理（1961年）
  - 蛋白质分子的一级序列决定其空间结构
  - 蛋白质天然构象是能量最低的构象

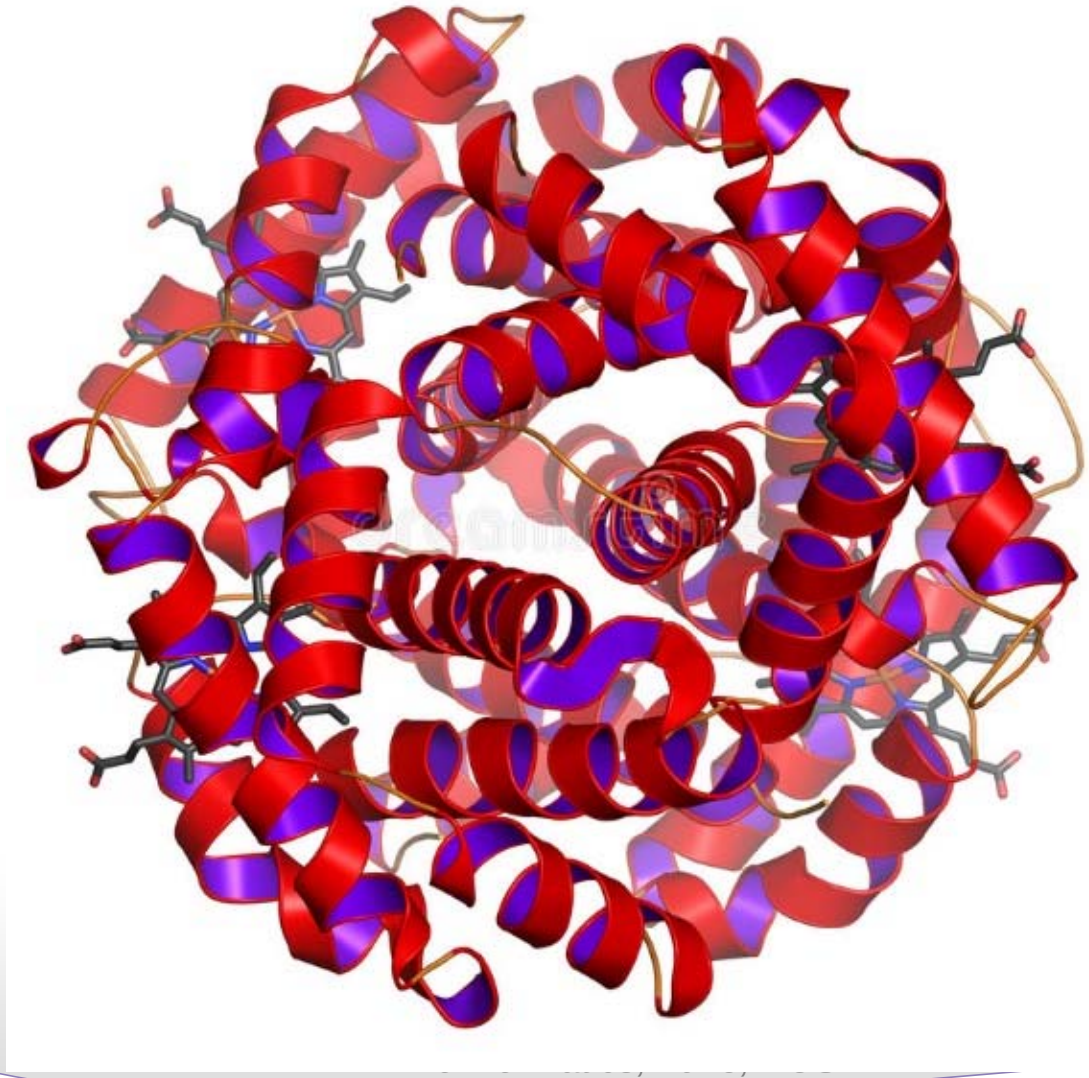
蛋白质结构预测的应用场景

- 点突变作用分析
- 酶促反应机制研究
- 蛋白质复合物和活性位点的界面相互作用分析
- 晶体衍射数据的定相（phasing）
- 相近家族的归类
- 新药设计：配体类药化合物的设计和虚拟筛选

蛋白质总体分类

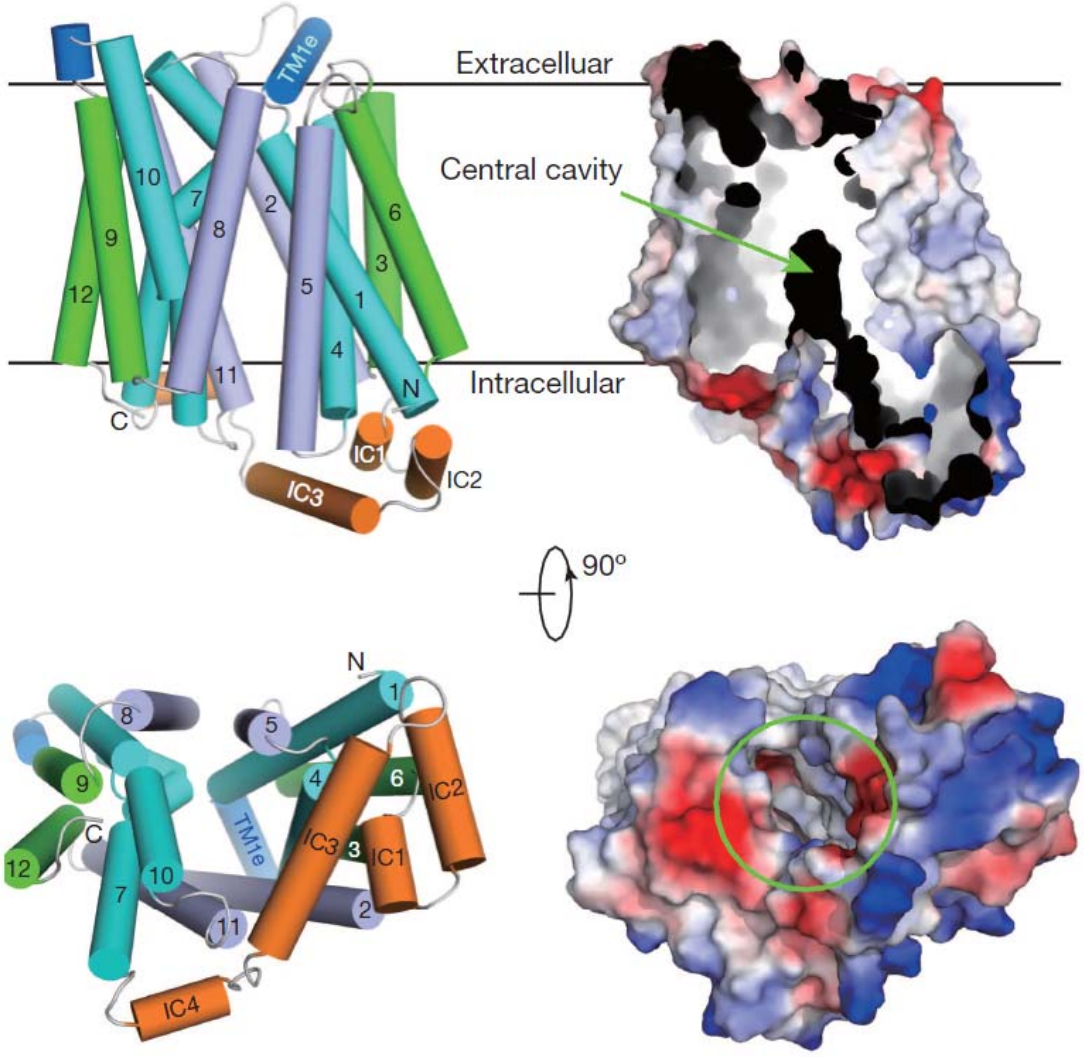
- 球蛋白（globular protein）：具有疏水内核和亲水表面
- 膜蛋白（membrane protein）：具有特定的疏水表面
- 蛋白质的临界稳定（marginally stable）状态
  - 在细胞内，蛋白质折叠之后并不总是处在能量最低的构象
  - 随着蛋白质功能的动态变化其构象也发生相应的变化
- 蛋白质无序区（intrinsically disordered region，IDR）
  - 需要与其他蛋白质结合后才能够获得稳定的结构
  - 介导蛋白质-蛋白质相互作用
  - 促进蛋白质“液-液相分离”（liquid-liquid phase separation，LLPS）的发生

球蛋白



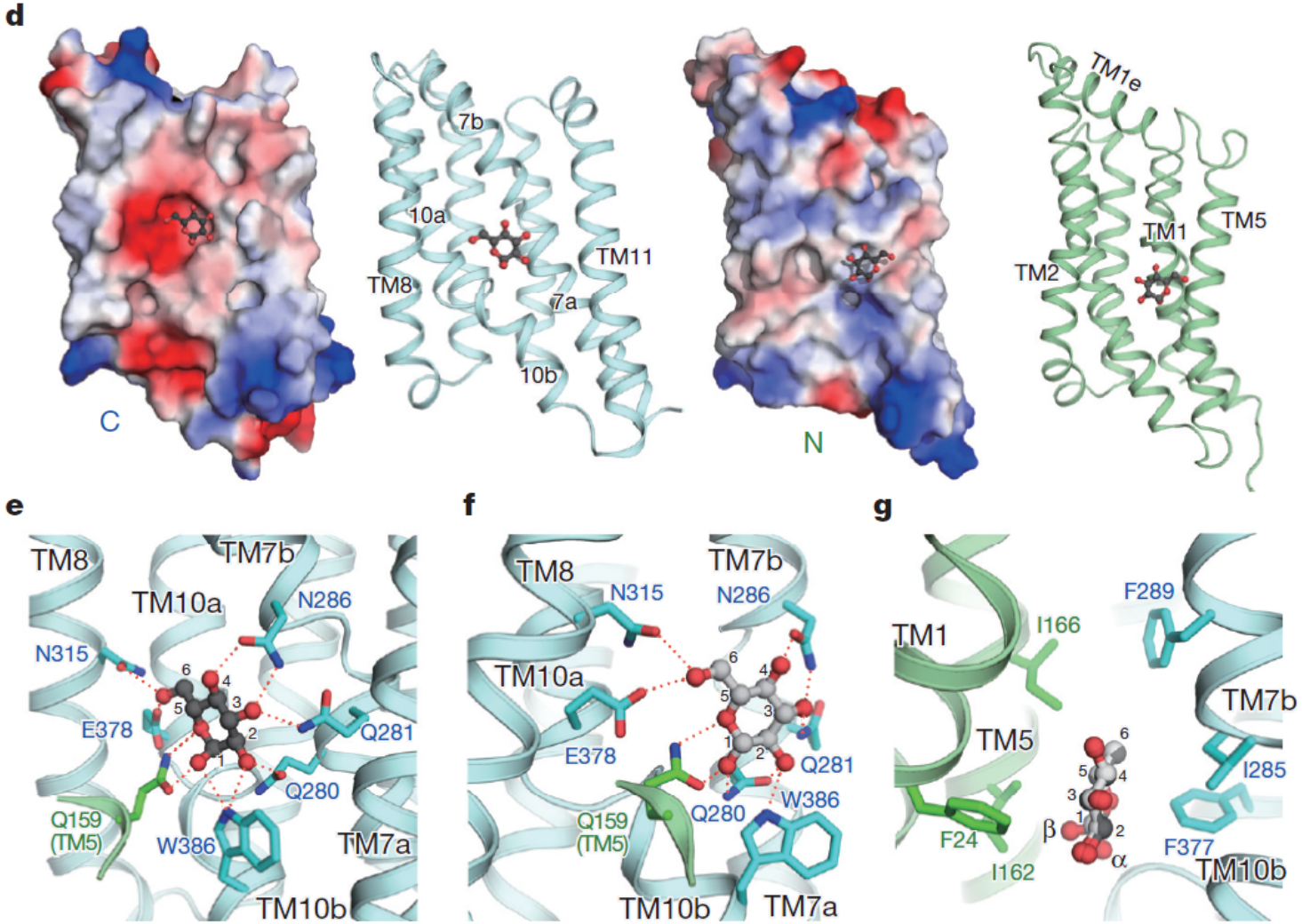
膜蛋白：GLUT1

- 2014，12次跨膜，N45T & E329Q突变体

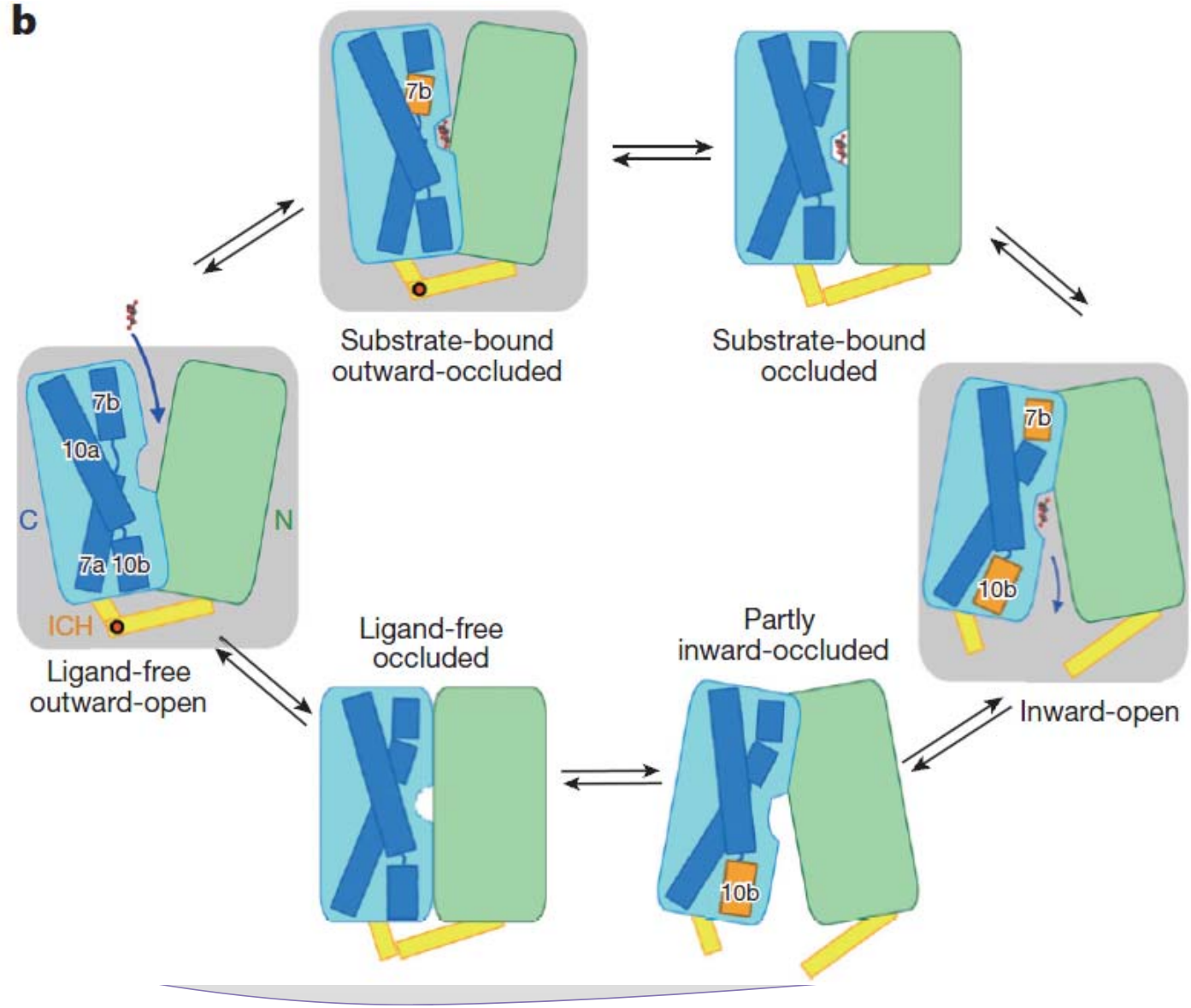


GLUT3

❒ 2015年

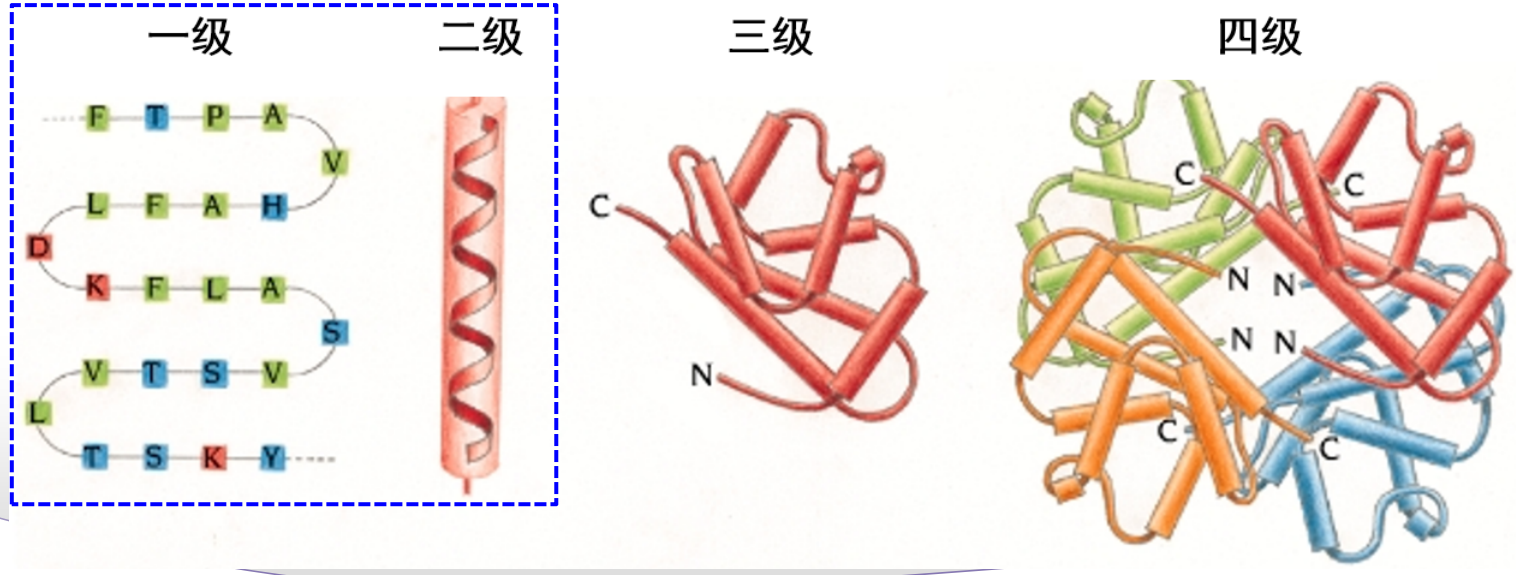

d
C
7b
10a
TM11
TM8
7a
10b
N
TM1e
TM1
TM5
TM2
e
TM8
TM7b
TM10a
N286
N315
E378
Q281
Q280
Q159
(TM5)
W386
TM7a
TM10b
f
TM8
TM7b
N315
N286
TM10a
E378
Q281
W386
Q280
Q159
(TM5)
TM10b
TM7a
g
TM1
I166
F289
TM7b
TM5
I285
F24
I162
F377
TM10b

GLUTs作用机制


b
7b
Substrate-bound
outward-occluded
Substrate-bound
occluded
7b
10a
C
N
7a 10b
ICH
Ligand-free
outward-open
Ligand-free
occluded
Partly
inward-occluded
7b
10b
Inward-open
10b

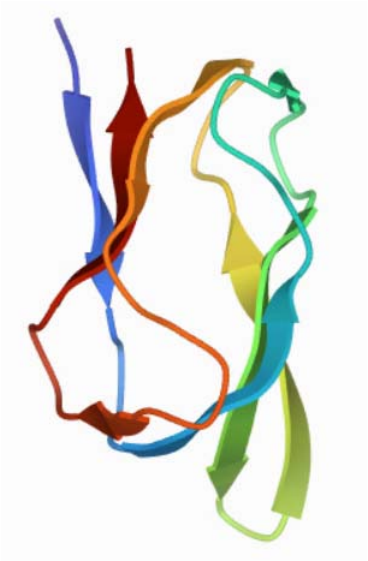
一级和二级结构

- 蛋白质一级结构
  - 氨基酸的线性序列，氨基酸残基之间通过共价键连接
- 二级结构
  - 氨基酸残基局部空间内的排列，残基间存在短程的、非共价的相互作用



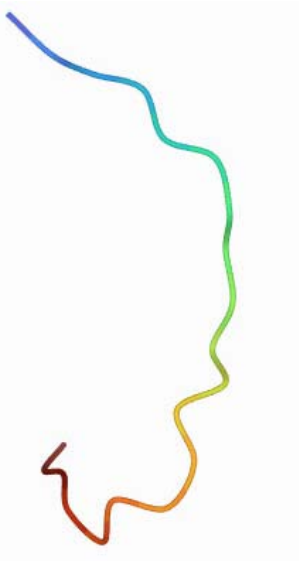
二级结构的4种周期性模式

α-螺旋，1DNG



β-折叠，5GUA

环，1L3Q

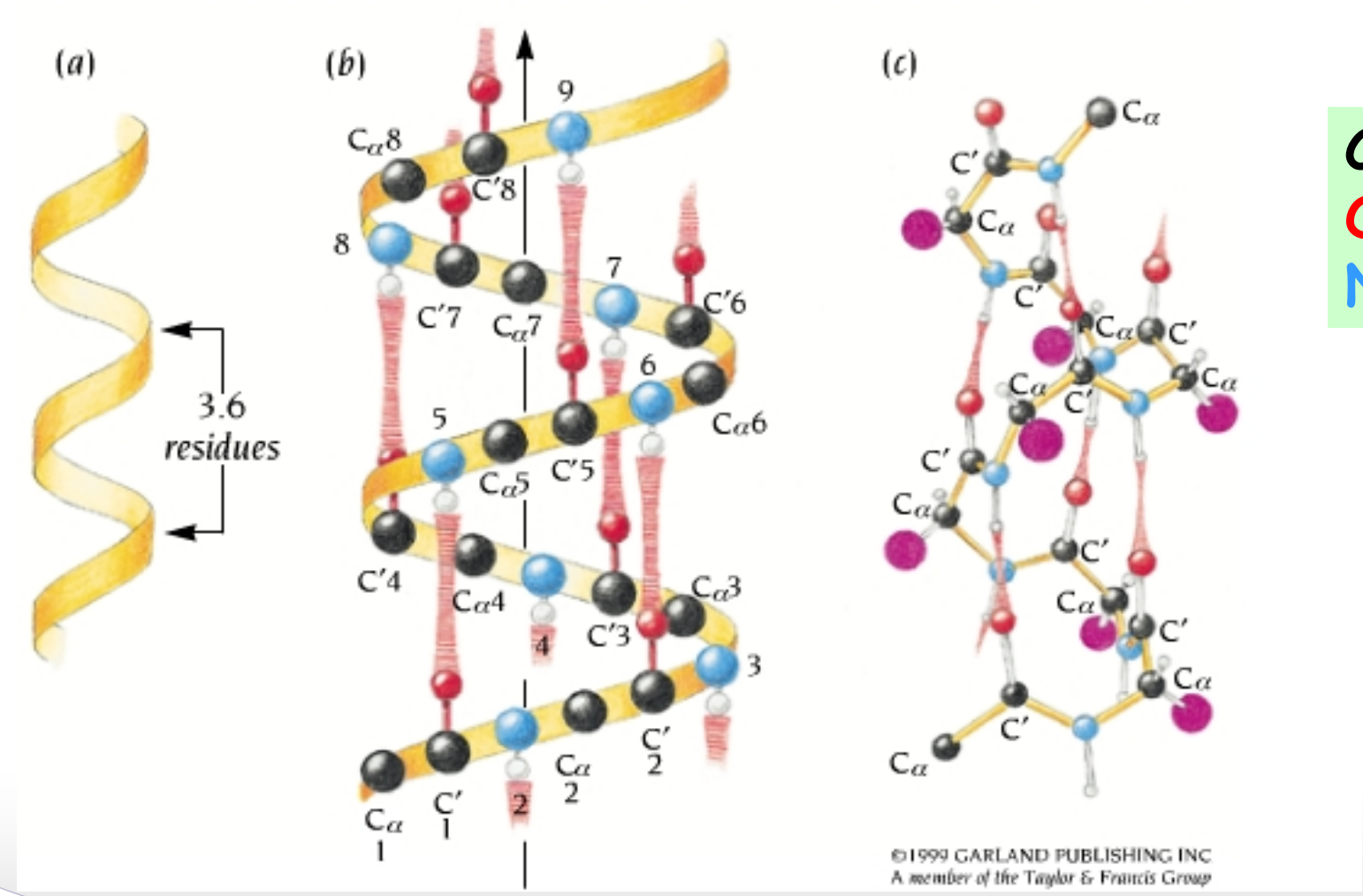


卷曲，1L1K

二级结构的4种周期性模式

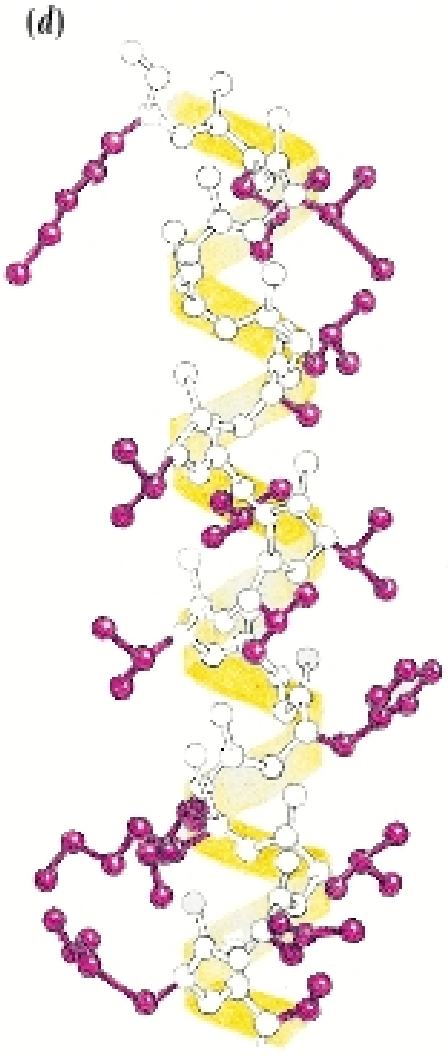
- α-螺旋（ α-helix）
  - 蛋白质中最多的二级结构
  - 平均长度：10个氨基酸残基 (10 A^0)
  - 长度范围为5-40个残基，螺旋每一圈3.6个残基
  - α-螺旋通过氢键稳定结构，通常在蛋白质内核的表面
  - 疏水残基向内，亲水残基向外
  - R-侧基分布在α-螺旋的外侧

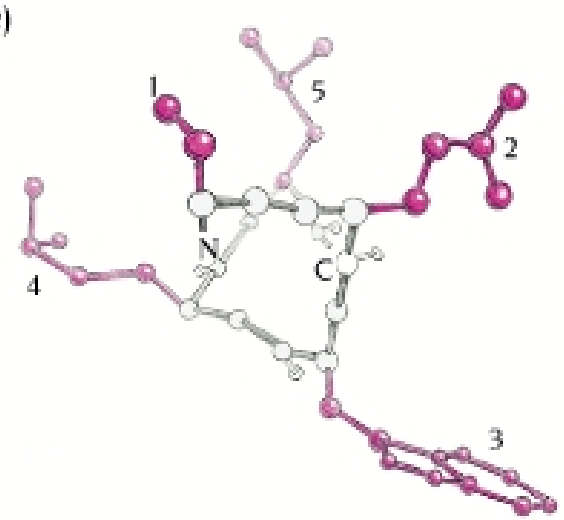
α-螺旋通过氢键稳定结构



C = black
O = red
N = blue

R-侧基分布在α-螺旋的外侧





α-螺旋的氨基酸偏好

- 丙氨酸、谷氨酸、亮氨酸和甲硫氨酸出现的频率高
- 脯氨酸、甘氨酸、酪氨酸和丝氨酸出现的频率低

Table 2.1 Amino acid sequences of three α helices

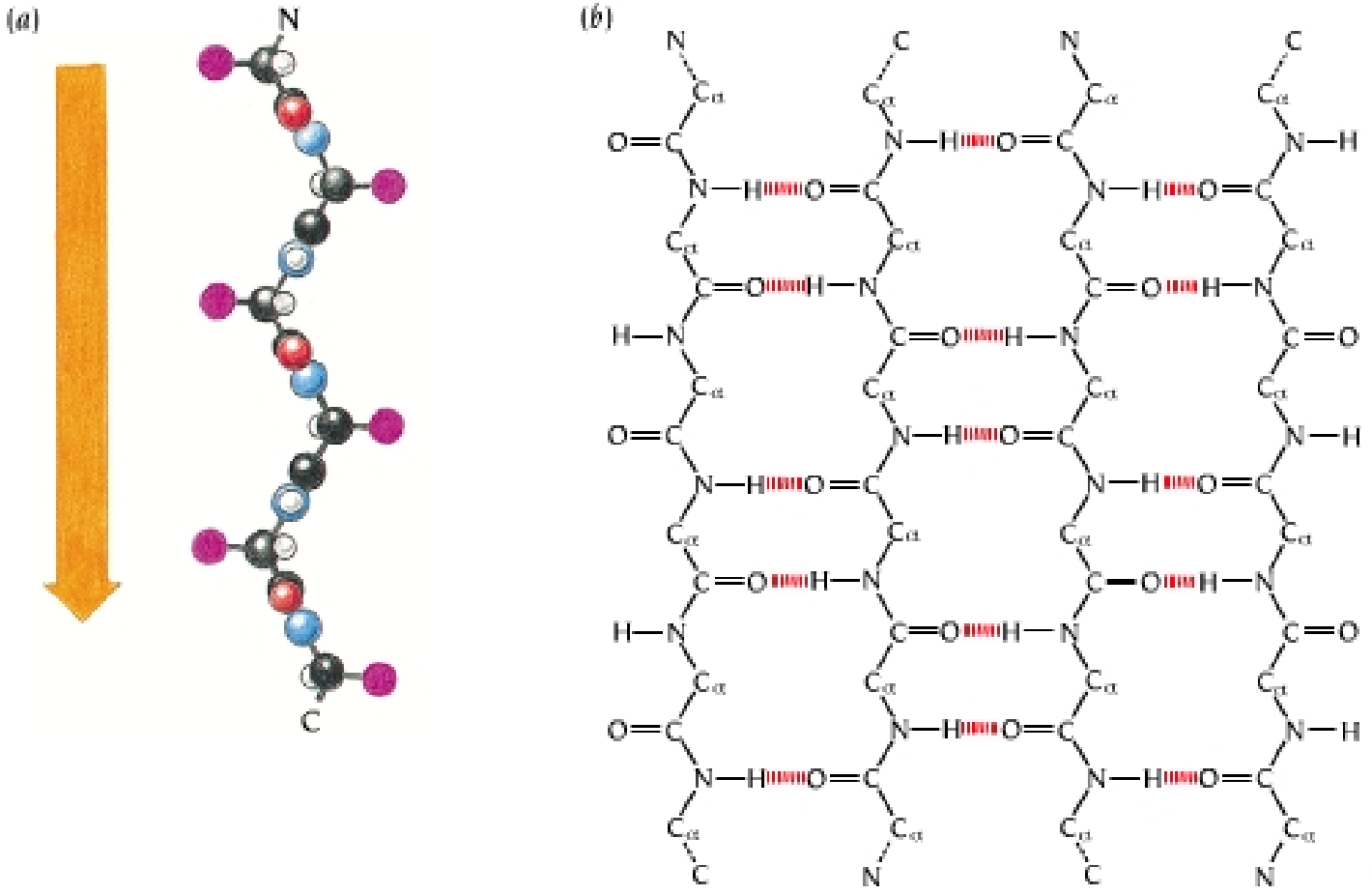
1.	Leu	Ser	Phe	Ala	Ala	Ala	Met	Asn	Gly	Leu	Ala
2.	Ile	Asn	Glu	Gly	Phe	Asp	Leu	Leu	Arg	Ser	Gly
3.	Lys	Glu	Asp	Ala	Lys	Gly	Lys	Ser	Glu	Glu	Glu

The first sequence is from the enzyme citrate synthase, residues 260–270, which form a buried helix; the second sequence is from the enzyme alcohol dehydrogenase, residues 355–365, which form a partially exposed helix; and the third sequence is from troponin-C, residues 87–97 which form a completely exposed helix. Charged residues are coloured red, polar residues are blue, and hydrophobic residues are green.

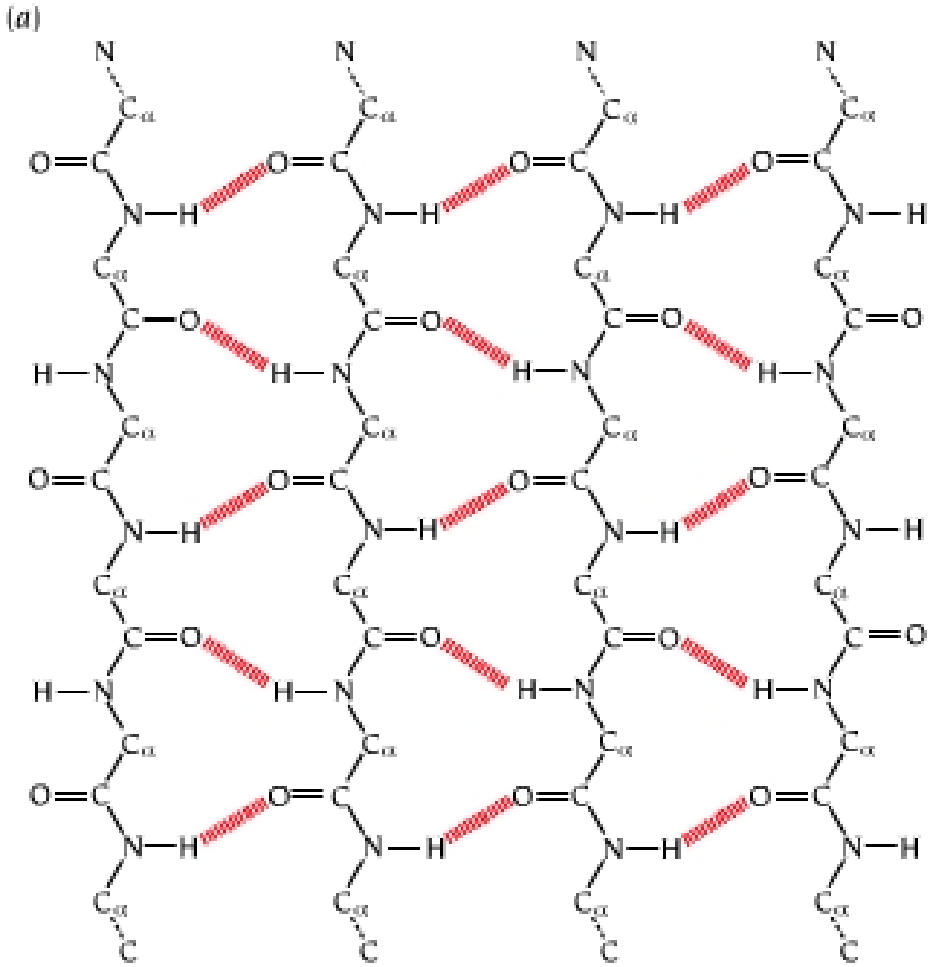
β-折叠（β-sheet）

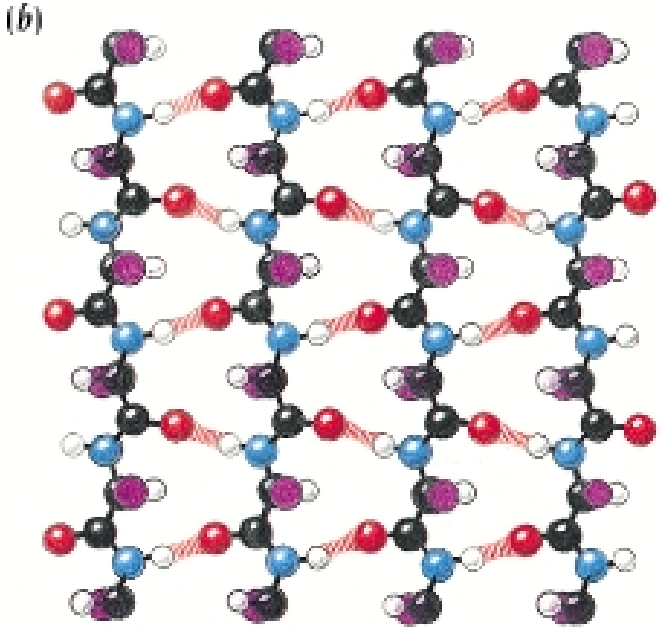
- 一般不单独出现，往往是成对或多个出现
- 不同β-折叠通过氢键连接，从而稳定结构
- β-折叠首尾相连的部分，主要通过短的或长的环连接
- β-折叠的结构包括反平行、平行和混合型

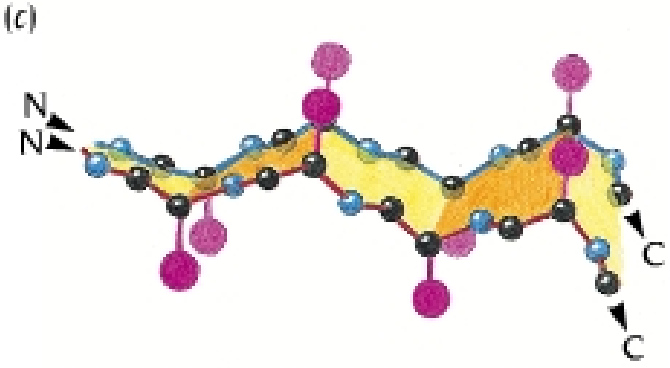
反平行的β-折叠



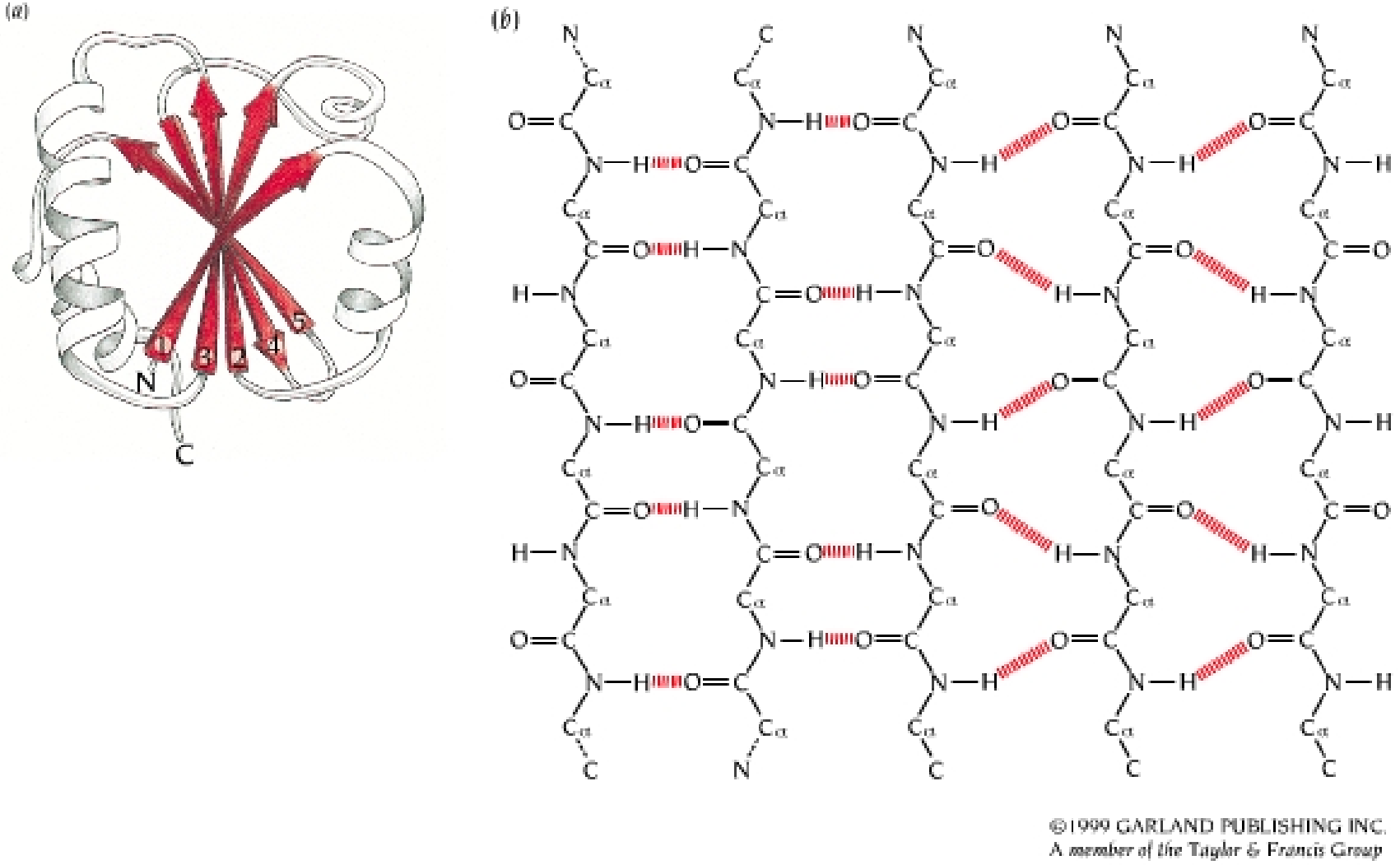
平行的β-折叠







混合的β-折叠

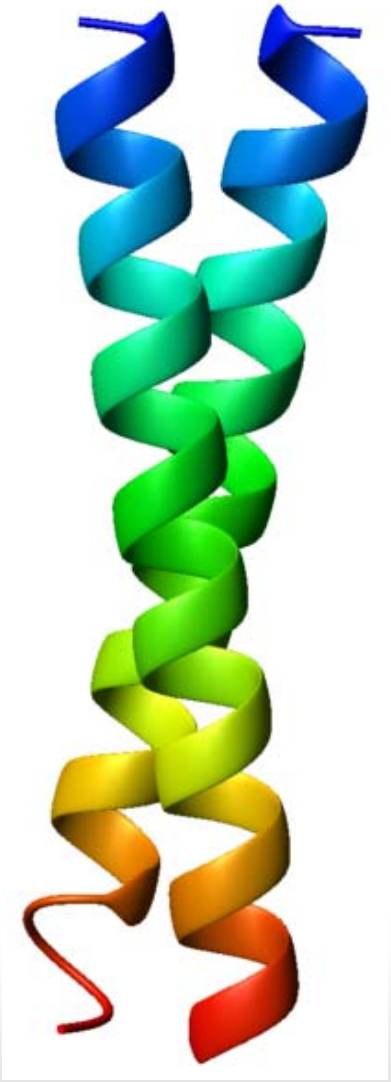

(a)
1
3
2
4
5
N
C
(b)
©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

环（loop）

- 连接α-螺旋和β-折叠
- 长度和三级结构不定
- 通常在蛋白质结构的表面，受点突变的影响较小
- 环的柔性好，构象变化余地大，带电荷、极性的氨基酸比例高
- 倾向于成为蛋白质的活性位点

卷曲（coil）

- 具有无序性，长度范围为4-20个残基
- 与环类似主要是连接α-螺旋和β-折叠

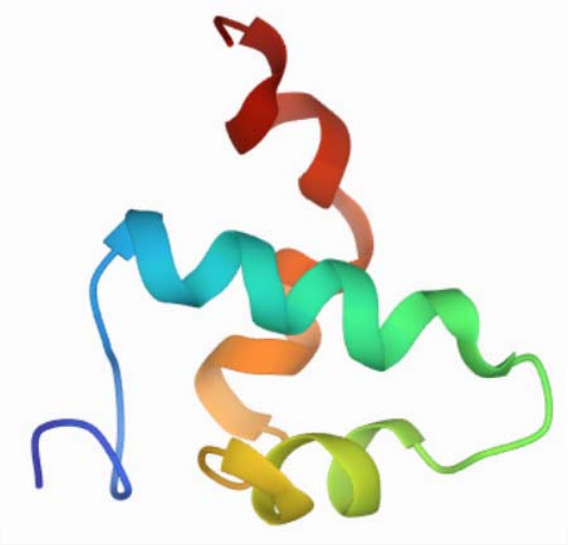


三级和四级结构

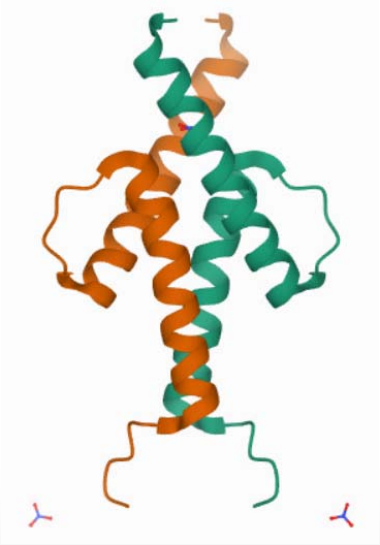
- 蛋白质三级结构
  - 肽链折叠成三维的空间结构
  - 是二级结构在空间上的排布
  - 包括长程的、共价与非共价的相互作用
- 蛋白质的四级结构
  - 多个肽链在空间上的排布
- 蛋白质超二级结构
  - 二级结构和三级结构之间，多个二级结构的组合
  - 也称为“结构模体”（structural motif）

蛋白质超二级结构

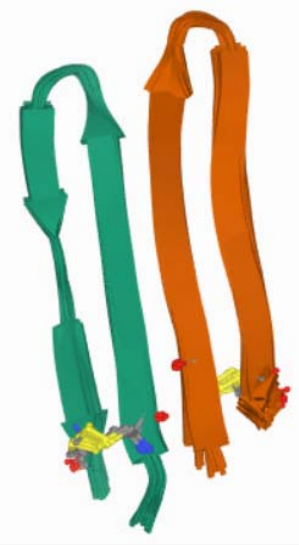
- 螺旋-转角-螺旋（helix-turn-helix，例如DNA结合模体）
- 螺旋-环-螺旋（helix-loop-helix，例如钙离子结合模体）
- β-发卡（β-hairpin）
- 希腊钥匙（Greek key）



螺旋-转角-螺旋，1BW6



螺旋-环-螺旋，3U5V



β-发卡，2L8X

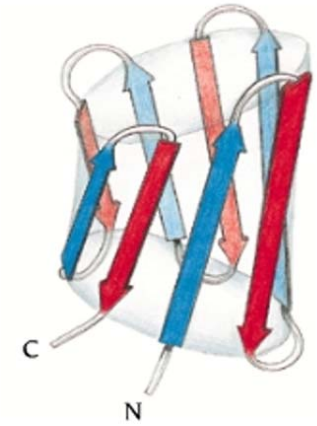
希腊钥匙，4CV7

蛋白质的6种结构类型

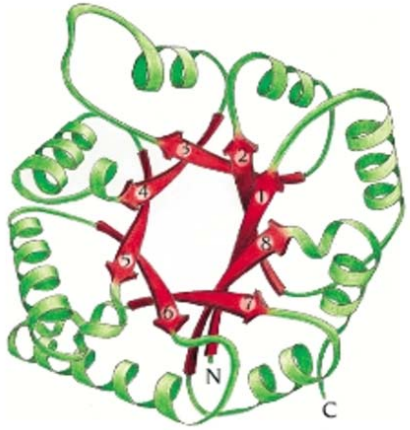
- α结构域，多个α螺旋束通过环连接
- β结构域，主要包含反平行β折叠，两对反平行的β折叠形成三明治（sandwich）结构
- α/β结构域，α螺旋连接的平行的β折叠
- α+β结构域，α螺旋和β折叠各自形成单独的结构
- 多结构域，包含多种α或β结构域
- 膜蛋白或细胞表面蛋白质

蛋白质的主要4种结构类型

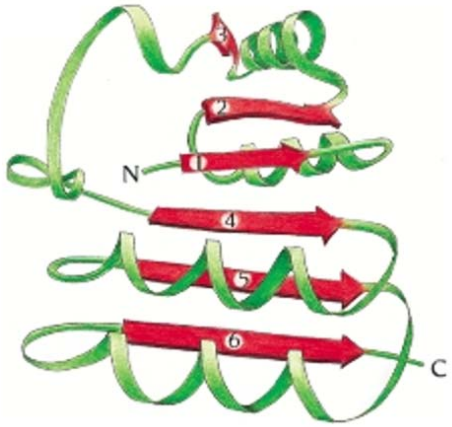
α结构域



β结构域

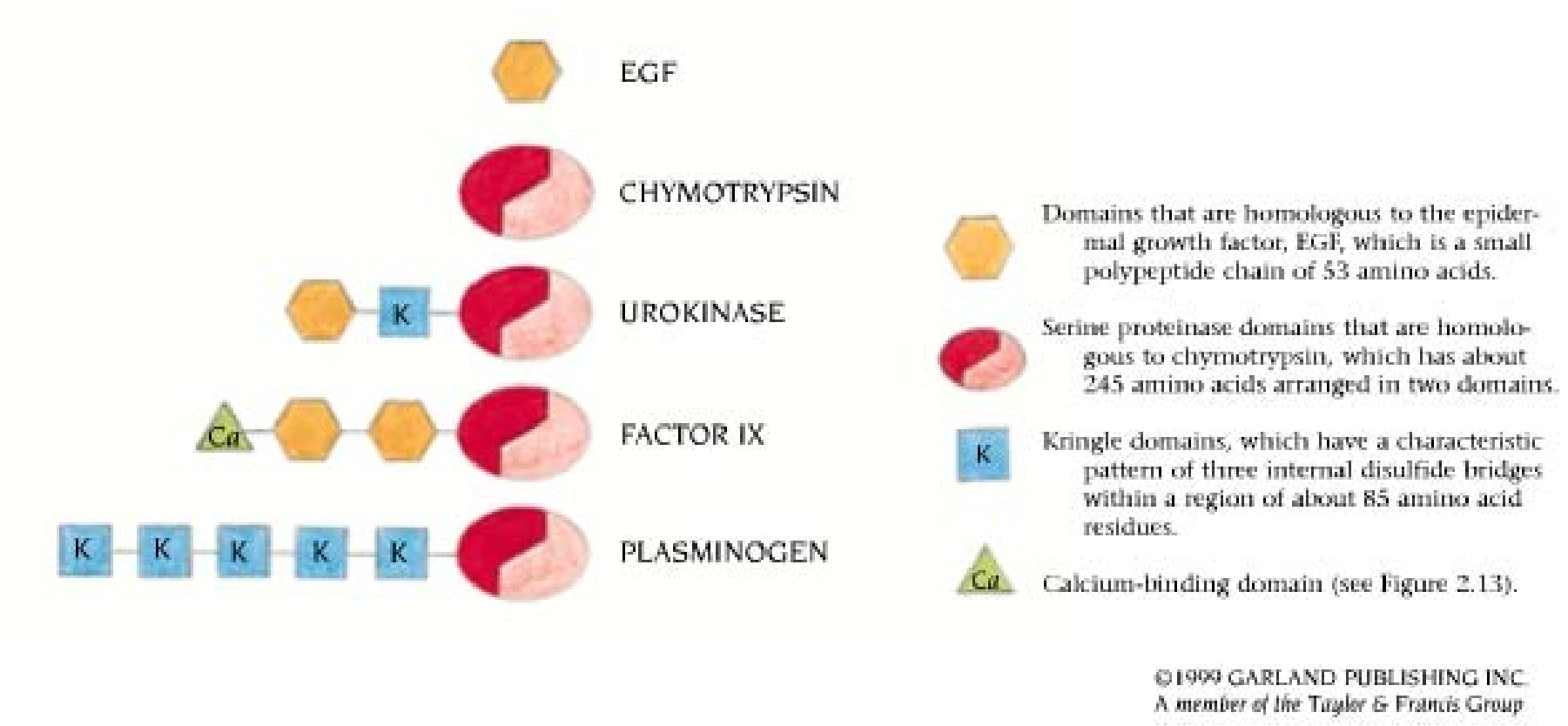


α/β结构域
（TIM barrel）



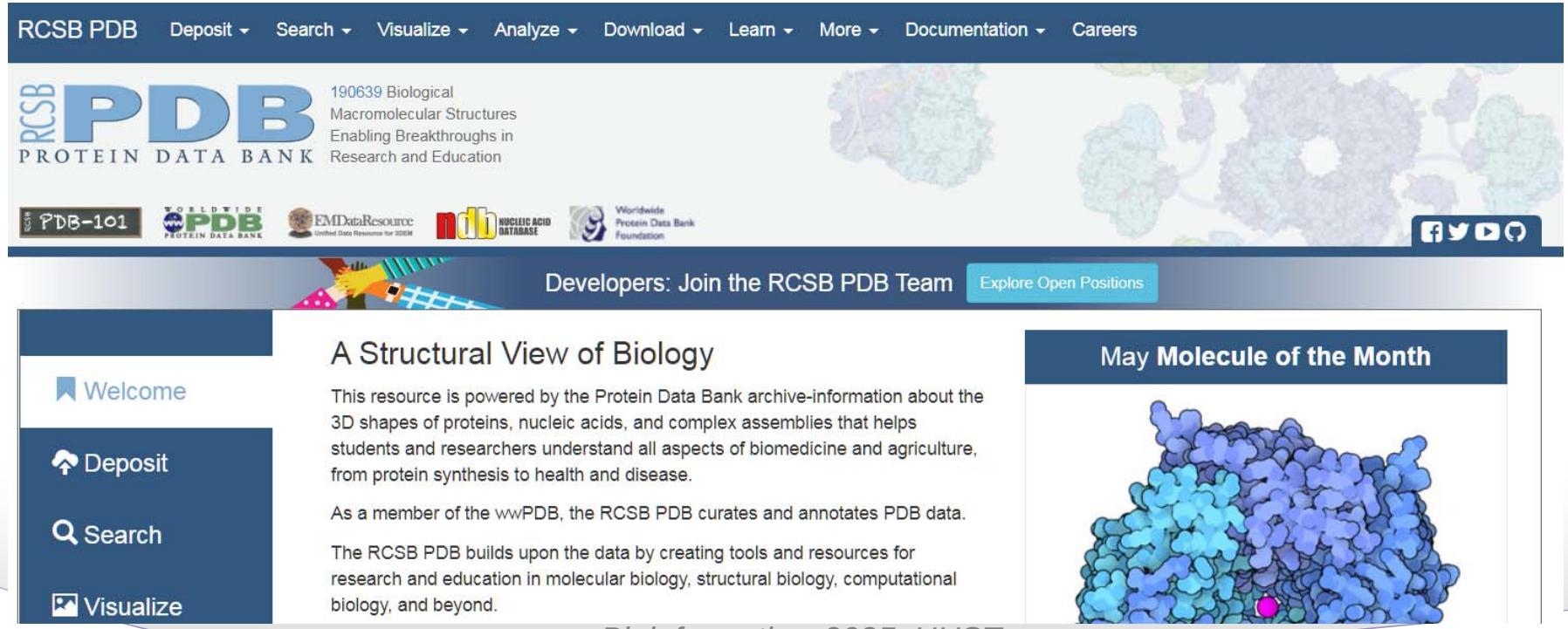
α/β结构域
（Rossman fold）

多结构域


EGF
CHYMOTRYPSIN
K
UROKINASE
Ca
FACTOR IX
K K K K K
PLASMINOGEN
Domains that are homologous to the epidermal growth factor, EGF, which is a small polypeptide chain of 53 amino acids.
Serine proteinase domains that are homologous to chymotrypsin, which has about 245 amino acids arranged in two domains.
K
Kringle domains, which have a characteristic pattern of three internal disulfide bridges within a region of about 85 amino acid residues.
Ca
Calcium-binding domain (see Figure 2.13).
©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

蛋白质结构数据库PDB

- 用于保存生物大分子结构数据的常用档案库
- 1971年，美国Brookhaven国家实验室创建
- 信息每周更新
- 合作伙伴：PDBe、PDBj、BMRB、EMDB



蛋白质结构数据库PDB

- 以文本文件（.pdb文件，可用记事本或富文本浏览器打开）的方式存放数据
- 每个分子各用一个独立的文件,有唯一的PDB-ID
- PDB-ID包含4个字符，由大写字母和数字组成（如血红蛋白的PDB-ID为4HHB）
- 文件中除了原子坐标外，还包括物种来源、化合物名称、结构以及有关文献等基本注释信息
- 还给出分辨率、结构因子、温度系数、蛋白质主链数目、配体分子式、金属离子、二级结构信息、二硫键位置等和结构有关的数据
- PDB格式的文件可以用IQmol或RasMol/OpenRasMol等可视化软件开发，从而直观地观察蛋白质的三维结构

PDB文件的数据格式

4J7B

Download File | View File

Crystal structure of polo-like kinase 1

Xu, J., Shen, C., Wang, T., Quan, J.

(2013) Nat Struct Mol Biol **20** 1047-1053

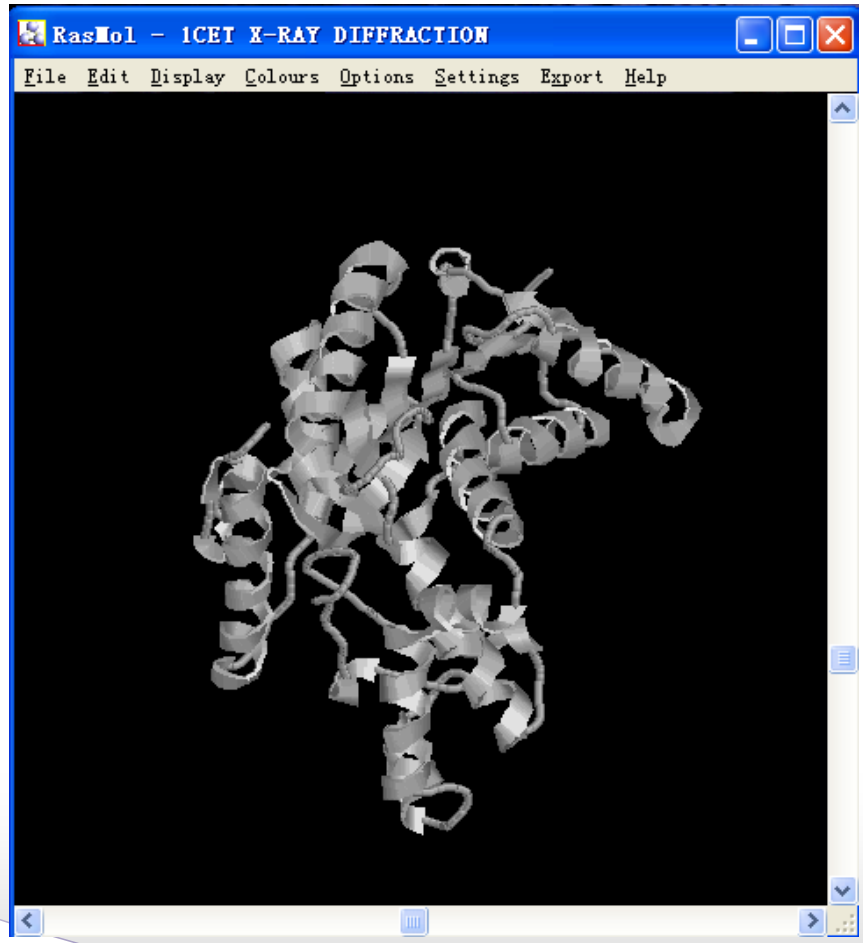
Released: 7/24/2013
Method: X-ray Diffraction
Resolution: 2.3 Å
Residue Count: 1184

Macromolecule:
Polo-like kinase (protein)
Polo-like kinase (protein)
205 kDa microtubule-associated pro ... (protein)
Unique Ligands: --

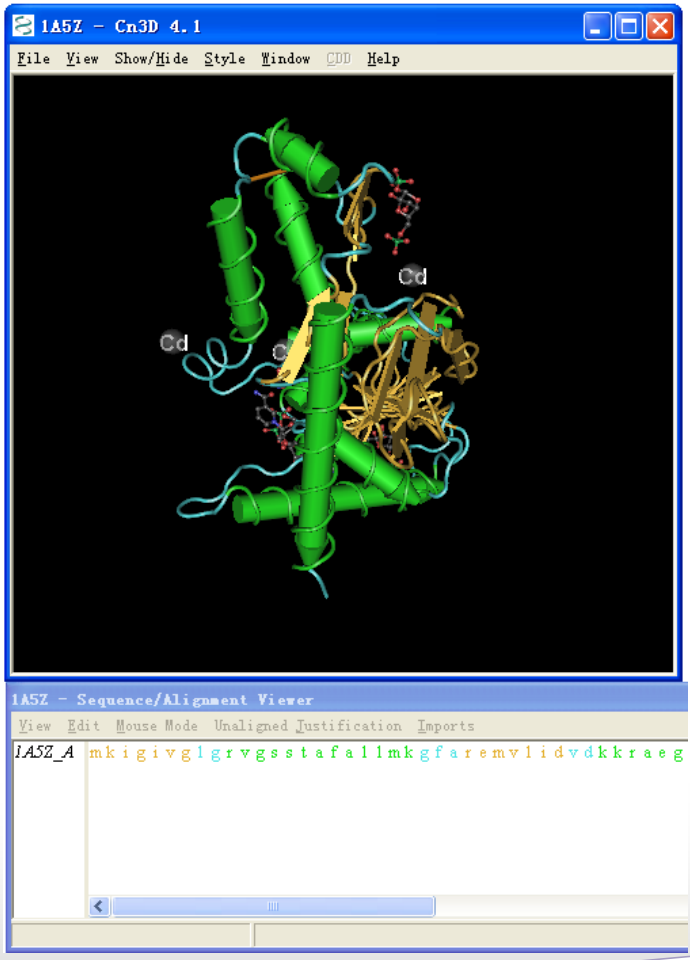
									X	Y	Z												
ATOM	1	N	N	.	PRO	A	1	3	?	24.595	65.444	-36.298	1.00	108.09	?	?	?	?	?	?	18	PRO	A	N	1
ATOM	2	C	CA	.	PRO	A	1	3	?	23.472	64.521	-36.217	1.00	102.09	?	?	?	?	?	?	18	PRO	A	CA	1
ATOM	3	C	C	.	PRO	A	1	3	?	22.492	64.776	-37.341	1.00	101.00	?	?	?	?	?	?	18	PRO	A	C	1
ATOM	4	O	O	.	PRO	A	1	3	?	22.587	65.801	-38.013	1.00	104.89	?	?	?	?	?	?	18	PRO	A	O	1
ATOM	5	C	CB	.	PRO	A	1	3	?	22.856	64.819	-34.845	1.00	100.79	?	?	?	?	?	?	18	PRO	A	CB	1
ATOM	6	C	CG	.	PRO	A	1	3	?	23.867	65.669	-34.113	1.00	105.92	?	?	?	?	?	?	18	PRO	A	CG	1
ATOM	7	C	CD	.	PRO	A	1	3	?	25.097	65.791	-34.965	1.00	110.04	?	?	?	?	?	?	18	PRO	A	CD	1
ATOM	8	N	N	.	LYS	A	1	4	?	21.577	63.840	-37.527	1.00	95.96	?	?	?	?	?	?	19	LYS	A	N	1
ATOM	9	C	CA	.	LYS	A	1	4	?	20.827	63.664	-38.763	1.00	94.57	?	?	?	?	?	?	19	LYS	A	CA	1
ATOM	10	C	C	.	LYS	A	1	4	?	19.461	63.005	-38.506	1.00	89.42	?	?	?	?	?	?	19	LYS	A	C	1
ATOM	11	O	O	.	LYS	A	1	4	?	18.977	62.236	-39.325	1.00	86.89	?	?	?	?	?	?	19	LYS	A	O	1
ATOM	12	C	CB	.	LYS	A	1	4	?	21.646	62.763	-39.712	1.00	94.64	?	?	?	?	?	?	19	LYS	A	CB	1
ATOM	13	C	CG	.	LYS	A	1	4	?	22.140	61.460	-39.061	1.00	91.72	?	?	?	?	?	?	19	LYS	A	CG	1
ATOM	14	C	CD	.	LYS	A	1	4	?	23.061	60.680	-39.989	1.00	92.69	?	?	?	?	?	?	19	LYS	A	CD	1
ATOM	15	C	CE	.	LYS	A	1	4	?	22.290	59.688	-40.863	1.00	88.75	?	?	?	?	?	?	19	LYS	A	CE	1

蛋白质结构的可视化

RasWin

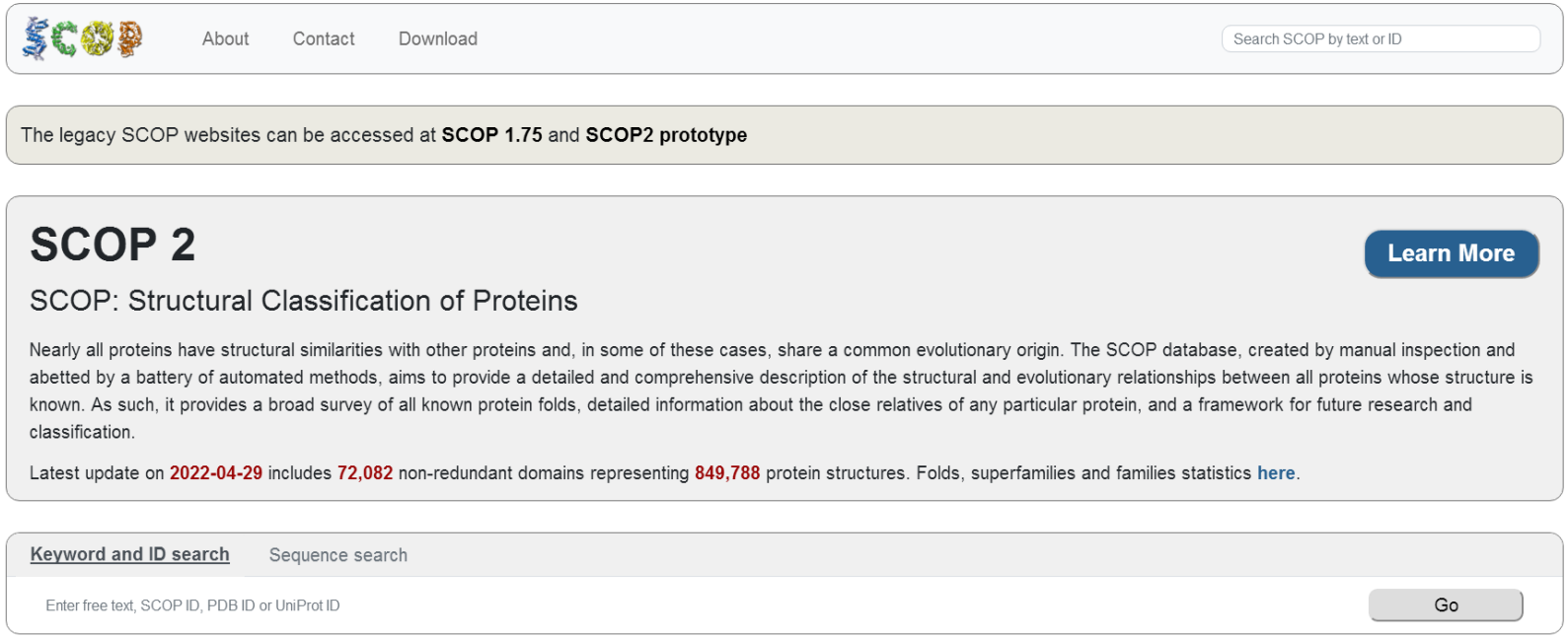


Cn3D



蛋白质结构分类数据库SCOP

- 1994年，英国医学研究委员会建立
- 对已知蛋白质结构进行分类的数据库
- 根据不同蛋白质的氨基酸组成及三级结构的相似性
- 描述已知结构蛋白的功能及进化关系
- 除了使用计算机程序外，主要依赖于人工验证



SCOPe

- SCOP扩展版本，结合自动和人工验证的方法对蛋白质进行注释
- SCOPe提供非冗余的ASTRAIL序列库，用来评估各种序列比对算法

SCOP数据库中的蛋白质分类

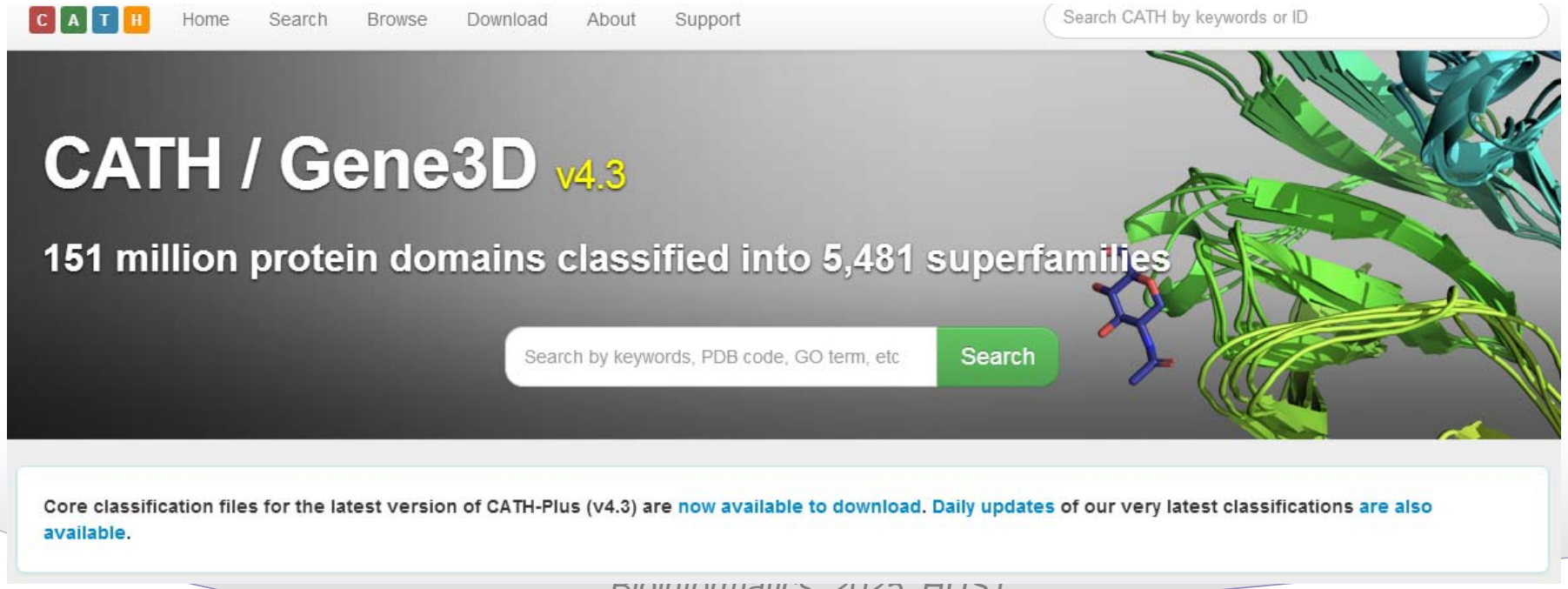
- 树状层级
  - 便于对目标蛋白的结构功能特征进行定位
  - 从根到叶依次为类（class）、折叠类型（fold）、超家族（super family）、家族（family）、蛋白质结构域（protein domain）、来源物种（species）、单个PDB蛋白质结构
- 超家族
  - 描述远源的进化关系, 如果序列相似性较低，但其结构和功能特性表明有共同的进化起源，则将其视作超家族

SCOP数据库中的蛋白质分类

- 家族
  - 描述相近的蛋白质进化关系
  - 通常将序列相似度在30%以上的蛋白质归入同一家族，即其有比较明确的进化关系
  - 某些情况下，尽管序列的相似度低于这一标准，也可以从结构和功能相似性推断其来自共同祖先，而归入同一家族
- 折叠类型
  - 描述空间的几何关系，无论有无共同的进化起源，只要二级结构单元具有相同的排列和拓扑结构，即归入相同的折叠方式
- 类
  - 依据二级结构组成，分为全α螺旋，全β折叠，α螺旋和β折叠，α螺旋＋β折叠以及其它特殊种类

蛋白质结构分类数据库CATH

- 1993年，英国伦敦大学于开发
- CATH：即蛋白质的种类（class，C）、二级结构的构架（architecture，A）、拓扑结构（topology，T）和蛋白质同源超家族（homologous superfamily，H）



CATH数据库中的蛋白质分类

- 种类
  - 全α、全β、α-β（α/β型和α+β型）和低二级结构四类，其中低二级结构类是指二级结构成分含量很低的蛋白质分子
- 构架
  - 第二个层次，主要考虑α螺旋和β折叠形成超二级结构的排列方式，不考虑其连接关系。这一层次的分类主要依靠人工方法
- 拓扑结构
  - 第三个层次，即二级结构的形状和二级结构间的联系，与SCOP中的折叠类型（fold）相当
- 同源性
  - 第四个层次，先通过序列比对再用结构比较来确定的

SCOP与CATH的区别

- SCOP注重从蛋白质进化角度进行分类
- CATH偏重于从结构角度对蛋白质分类，分类基础是蛋白质结构域
- CATH数据库根据序列相似度将结构域分为同一序列家族（sequence family，>=35%）、直系家族（orthologous Family，>=60%）、相似结构域（like domain，>=95%）或相同结构域（identical domain，=100%）

蛋白质二级结构预测

- 二级结构预测的目标
  - 蛋白质中约85%的残基处于三种稳定二级结构：α螺旋、β折叠和β转角
  - 根据蛋白质一级序列判断残基是否处于特定的3种二级结构之一
- 二级结构预测的重要性
  - 预测结果提供结构信息
  - 蛋白质空间结构预测的第一步
  - 是内部折叠、内部残基距离预测的基础
  - 可用于推测蛋白质功能和预测蛋白质结合位点等

蛋白质二级结构预测

- 基本依据
  - 每段相邻的氨基酸残基具有形成一定二级结构的倾向
  - 通过统计和分析发现这些倾向或者规律
  - 二级结构预测问题可转化为模式分类和识别问题
- 三态总体每残基准确性
  - Three-state overall per-residue accuracy，Q3
  - 评估蛋白质二级结构预测准确性的主要方法
  - 预测正确正确的具有α螺旋、β折叠或环/卷曲的残基数量，除以三种预测残基的加和

蛋白质二级结构的预测方法

- 第一代方法
  - 基于单个氨基酸残基统计分析
  - 从有限的数据集中提取各种残基形成特定二级结构的倾向
- 第二代方法
  - 统计的对象是长度为11到21个氨基酸片段，
  - 以之体现中心残基所处的环境
  - 以残基在特定环境形成特定结构的倾向作为依据预测中心残基的二级结构
  - 算法包括：基于统计信息；基于物理化学性质；基于序列模式；基于多层神经网络；基于多元统计；基于机器学习的专家规则；最邻近算法

Chou-Fasman法

- 基于统计信息，1974年，Peter Y. Chou和Gerald D. Fasman提出
- 根据氨基酸残基在各种二级结构中出现的频率来预测三种主要的二级结构α螺旋、β折叠和卷曲
- 训练数据：15个已知三维构象的蛋白质结构，共2473个氨基酸残基
- 作者定义了三种二级结构的蛋白质构象参数（protein conformational parameter）P_α、P_β和P_c

氨基酸在各种二级结构中的频率

TABLE I: Amino Acid Residues in the Helix, Inner Helix,[a] β-Sheet, and Coil Regions of 15 Proteins.

Amino Acid	No. of Residues	Residues in Helix	Residues in Inner Helix	Residues in β Region	Residues in Coil Region
Ala	228	119	62	38	71
Arg	78	22	9	12	44
Asn	133	35	12	15	83
Asp	111	39	10	15	57
Cys	54	15	3	12	27
Gln	95	40	16	20	35
Glu	113	62	28	5	46
Gly	232	45	22	32	155
His	74	33	11	9	32
Ile	106	38	22	29	39
Leu	196	94	64	41	61
Lys	175	67	34	22	86
Met	28	12	6	8	8
Phe	82	33	16	18	31
Pro	85	18	0	9	58
Ser	202	57	24	25	120
Thr	156	47	21	32	77
Trp	44	18	10	9	17
Tyr	100	22	10	22	56
Val	181	74	44	51	56
Total	2473	890	424	424	1159

[a] The three helical end residues on both N- and C-terminals of a helical region are omitted.

P_α, P_β, P_c的计算

$$P = \frac{f_i}{\frac{\sum f_j}{20}}$$

TABLE II: Frequency of Helical, Inner Helical,[a] β, and Coil Residues in 15 Proteins with Their Conformational Parameters P_α, $P_{\alpha i}$, P_β, and P_c.

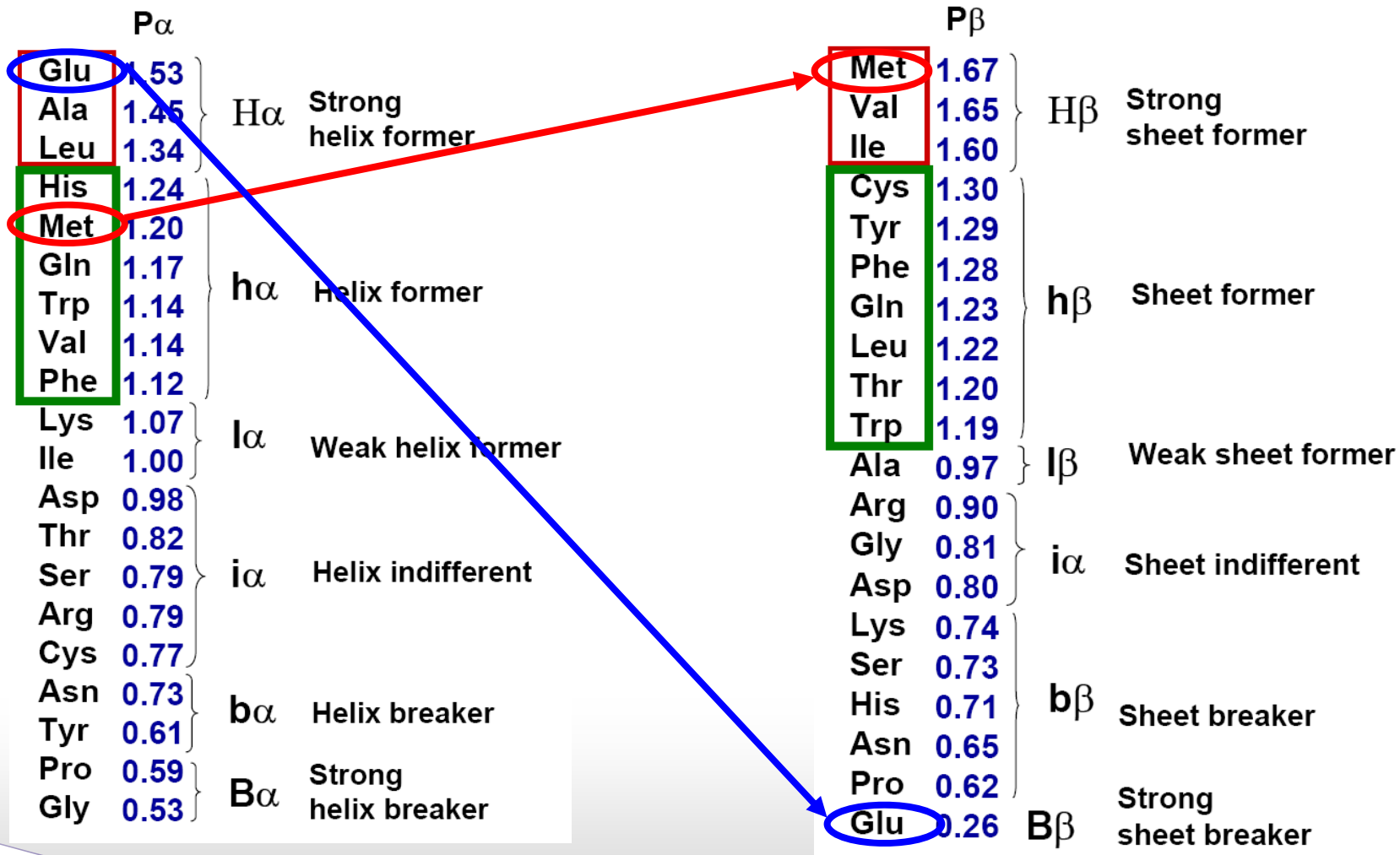
Amino Acid	f_α[b]	P_α[c]	$f_{\alpha i}$[b]	$P_{\alpha i}$[c]	f_β[b]	P_β[c]	f_c[b]	P_c[c]
Ala	0.522	1.45	0.272	1.59	0.167	0.97	0.311	0.66
Arg	0.282	0.79	0.115	0.67	0.154	0.90	0.564	1.20
Asn	0.263	0.73	0.090	0.53	0.113	0.65	0.624	1.33
Asp	0.351	0.98	0.090	0.53	0.137	0.80	0.514	1.09
Cys	0.278	0.77	0.056	0.33	0.222	1.30	0.500	1.07
Gln	0.421	1.17	0.168	0.98	0.211	1.23	0.368	0.79
Glu	0.549	1.53	0.248	1.45	0.044	0.26	0.407	0.87
Gly	0.190	0.53	0.091	0.53	0.138	0.81	0.668	1.42
His	0.446	1.24	0.149	0.87	0.122	0.71	0.432	0.92
Ile	0.358	1.00	0.208	1.22	0.274	1.60	0.368	0.78
Leu	0.480	1.34	0.327	1.91	0.209	1.22	0.311	0.66

f_i和f_j为三种二级结构之一中不同氨基酸出现的频率，P_α、P_β和P_c根据相应的不同氨基酸出现的频率分别进行计算，用来反映20种氨基酸残基在不同二级结构中的重要性

Trp	0.409	1.14	0.227	1.33	0.203	1.19	0.386	0.82
Tyr	0.220	0.61	0.100	0.58	0.220	1.29	0.560	1.19
Val	0.409	1.14	0.243	1.42	0.282	1.65	0.309	0.66

$\langle f_\alpha \rangle^c = 0.359$ $\langle P_\alpha \rangle^f = 1.00$ $\langle f_{\alpha i} \rangle^c = 0.171$ $\langle P_{\alpha i} \rangle^f = 1.00$ $\langle f_\beta \rangle^d = 0.171$ $\langle P_\beta \rangle^e = 1.00$ $\langle f_c \rangle^d = 0.469$ $\langle P_c \rangle^e = 1.00$

P_α & P_β



判定规则

- Chou-Fasman法的两条判定规则
  - 对于给定的一条长度大于6 aa的片段，若P_α值大于1.03，且P_α值大于P_β值，则判定为α螺旋
  - 对于给定的一条长度大于6 aa的片段，若P_β值大于1.05，且 P_β值大于P_α值，则判定为β折叠
- Chou-Fasman法的Q3准确性为~50-60%，对于β折叠的预测性能较差

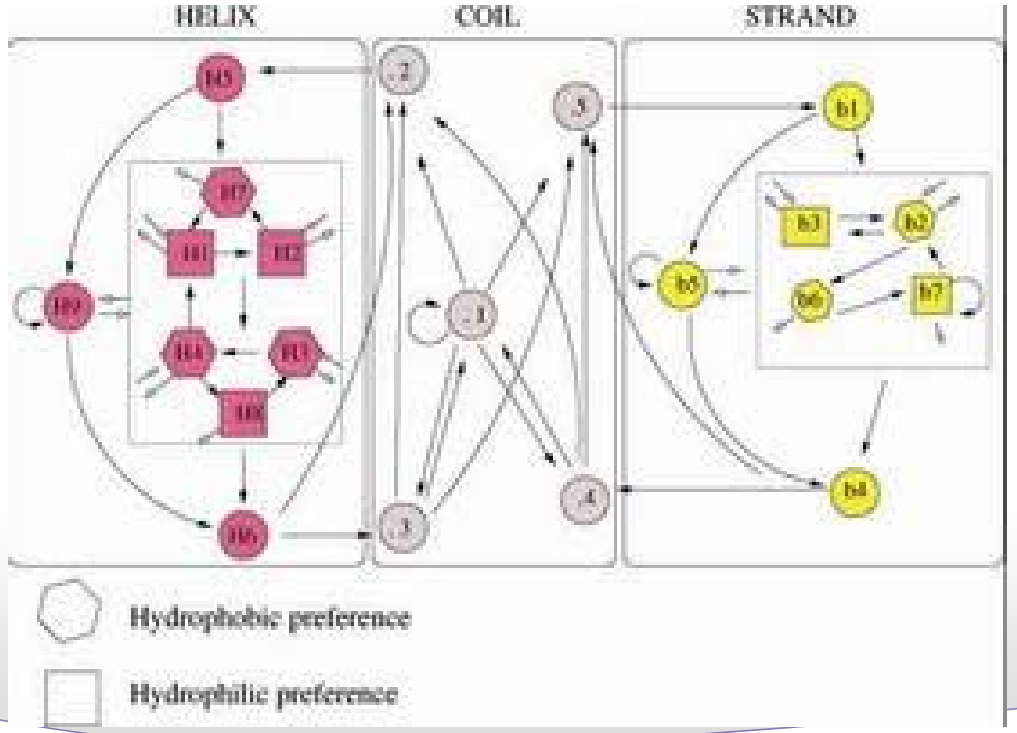
Garnier, Osguthorpe and Robson (GOR)：HMM

GOR Scoring Tables (original)

3 states – α-helix, β-sheet, turn

-8	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7	+8
W	R	Q	I	C	T	V	N	A	F	L	C	E	H	S	Y	K

准确性~65%



前两代方法的局限

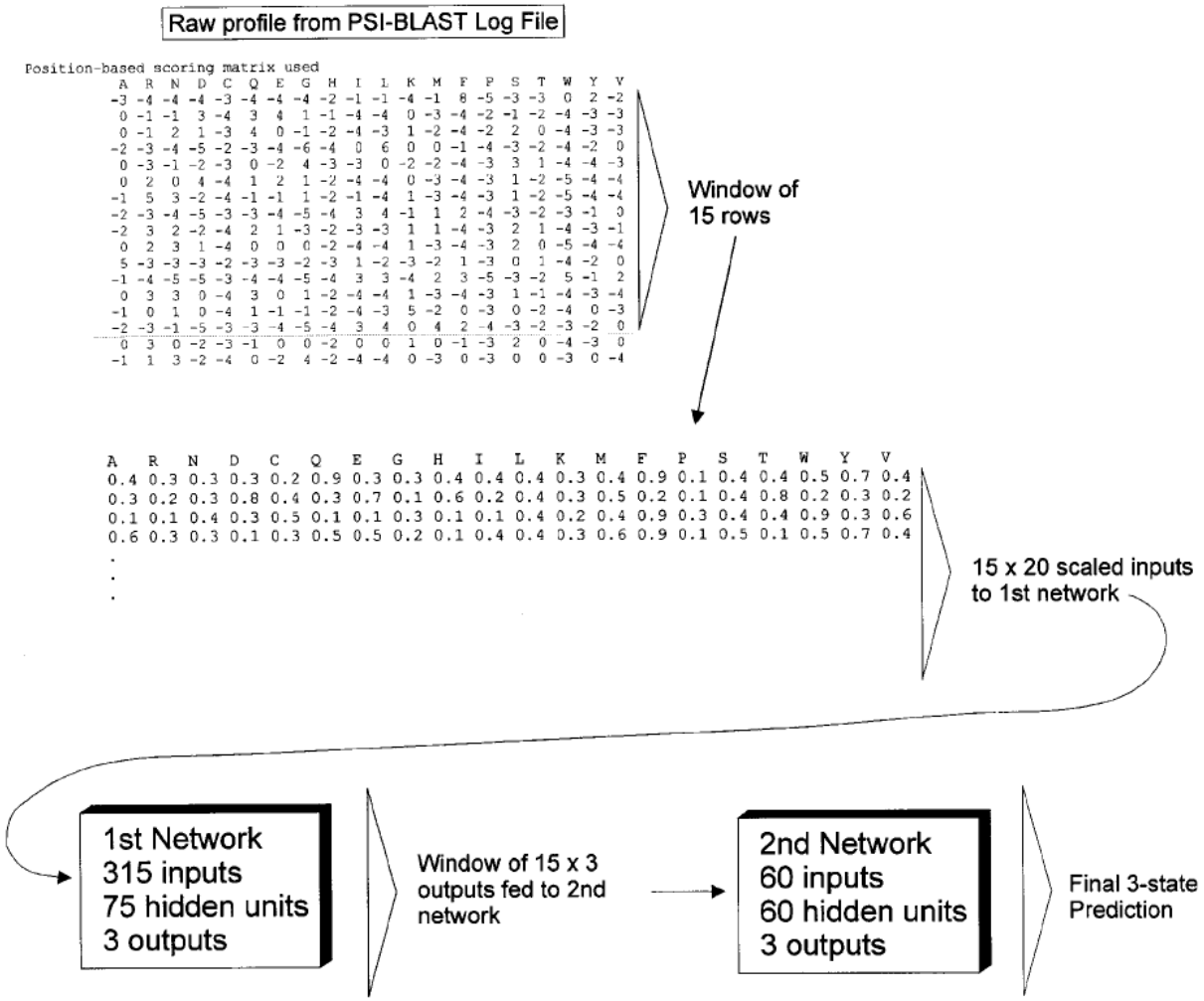
- 只利用局部，最多预测20个残基信息
- 统计分析表明局部信息仅包含65%左右的二级结构信息，长程相互作用不容忽视
- 只利用局部信息的二级结构预测方法准确率都小于70%
- 对β折叠预测的准确率仅为28%至48%

第三代方法

- 蛋白质家族的多序列比对
  - 得到进化信息，计算各残基的保守程度
  - 引入长程信息，描述其结构特征
  - 准确率能达到70%到75%
  - 对β折叠，预测结果与实验观察趋于一致
- 基于统计的神经网络方法PHDsec（第三代）
  - 首先达到70％的准确性
  - 将提交的靶序列利用BLASTP查询Swiss-Prot数据库得到同源序列
  - 将查询结果过滤后再进行CLUSTALW多序列比对，得到的进化信息作为神经网络的输入值进行计算
  - 同时采用20种氨基酸描述蛋白质序列的全局信
  - 综合考虑局部序列间关系和整体蛋白质性质来预测残基二级结构

David T. Jones: PSSM

❒ PSIPRED: PSSM + Neural Network

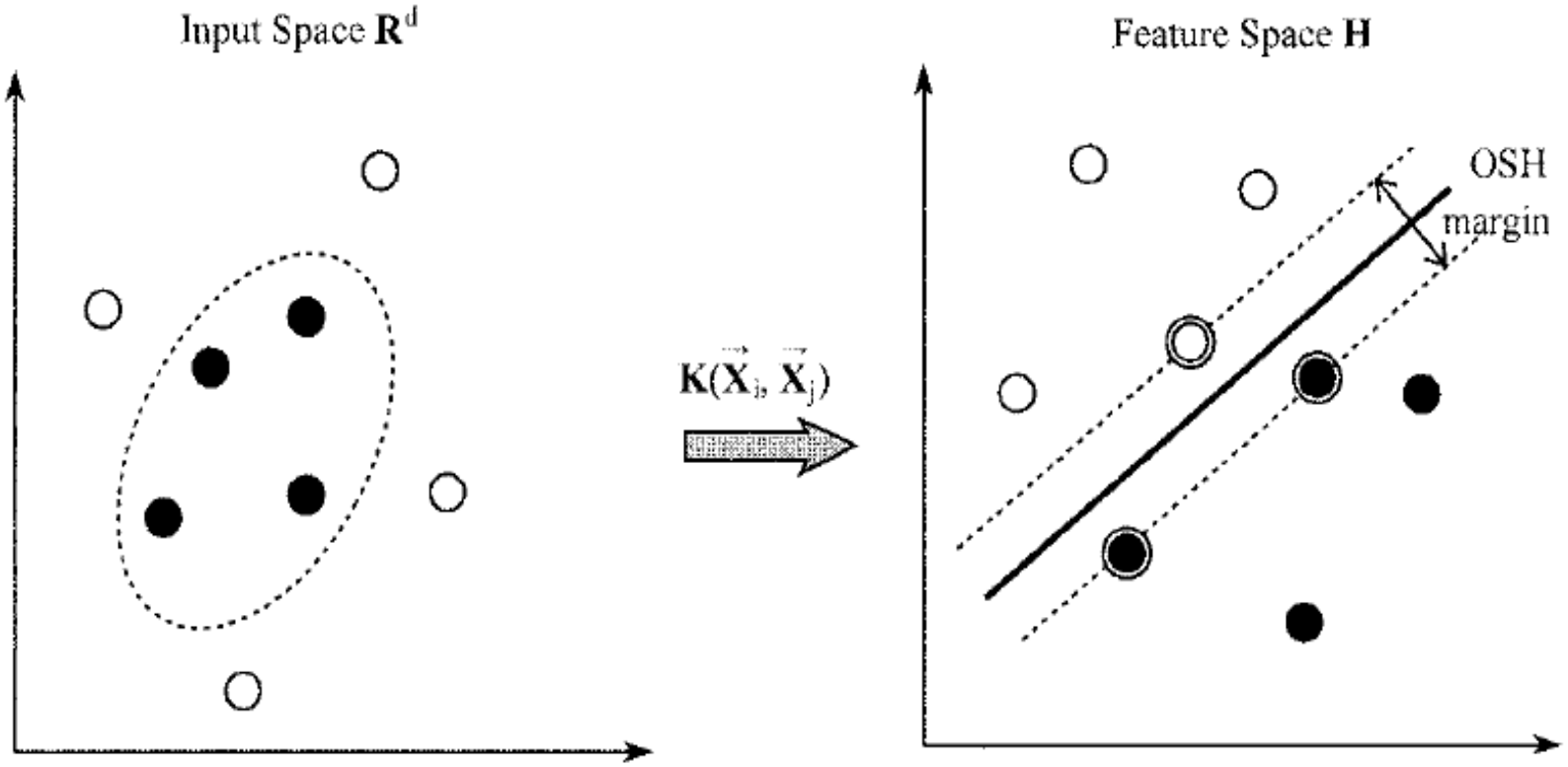


准确性
76.5%~
78.3%

支持向量机算法

- 2001年，清华大学孙之荣和华苏军开发
  - 训练集和测试集：序列相似性低于25%的126条蛋白质链（RS126数据集）和序列相似性低的513条蛋白质链
- 算法原理
  - 将问题转化成6组二分类预测，包括α螺旋/非α螺旋；β折叠/非β折叠；卷曲/非卷曲；α螺旋/β折叠；β折叠/卷曲；卷曲/α螺旋
  - 考虑了长度为5-17个氨基酸片段对中心残基所处的环境的影响
  - 序列编码方式是正交二进制编码（orthogonal binary coding，OBC），如丙氨酸的OBC向量为[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
  - 将编码之后的向量输入支持向量机中，即可训练计算模型
  - 参考了PHDsec，整合了蛋白质多序列比对获得的进化信息，Q3准确性达到73.5%

支持向量机算法



准确性~73.5%

第四代方法

- 使用深度学习（deep learning）技术来训练模型
- 2014年，华人学者程建林设计了DNSS法
- 预测二级结构的Q3准确性为80.7%

Comparison of Secondary Structure Predictions

	CASP9		CASP10		Combined CASP		Min Score	
Method	Q_3 (%)	Sov (%)	Q_3 (%)	Sov (%)	Q_3 (%)	Sov (%)	Q_3 (%)	Sov (%)
DNSS	81.1	74.7	80.2	73.6	80.7	74.2	50.4	46.1
PSSpred	83.3	72.0	81.0	70.4	82.2	71.3	41.8	33.7
SSpro	79.6	72.6	78.8	71.9	79.2	72.3	49.6	34.0
PSIPRED	80.9	69.3	81.2	68.6	81.0	69.0	33.8	23.2
RaptorX	78.1	70.4	77.9	70.3	78.0	70.3	45.6	33.0

蛋白质三级结构预测

- 结构基因组学
  - 人的基因组中包含>22,00个基因
  - 细胞内：通常>3,000种蛋白质
  - 目标：通过实验或者计算的手段解析所有蛋白质在自然条件下的三级结构
- 蛋白质折叠的动力学
- 蛋白质三级结构的预测：具有最小自由能的构象
  - 同源建模 (Homology modeling)
  - 穿针引线 (Threading)
  - 从头预测 (*Ab initio* Prediction)

蛋白质折叠

- 动力学过程：蛋白质一级序列发生折叠形成能量更低的三维构象的过程
- 在细胞内，蛋白质的折叠可以是自发的，也可以由酶或伴侣蛋白的介导
- 折叠过程中，蛋白质的熵与焓都发生改变
- 许多蛋白质在体外不能自发折叠
- 蛋白质折叠具有高度的动态性，其结构在自然条件下并不是固定的
- 蛋白质的功能常常依赖其构象的改变
- 自然条件下的蛋白质结构，与蛋白质变性之后的能量差非常小（~5-15 kcal/mol），大约等于1-2个氢键的能量

三级结构的相关预测

- 蛋白质结构预测或“蛋白质折叠”预测，即给定一条蛋白质的氨基酸序列，预测其三级结构，其基本原理为预测具有最小自由能的构象
- 蛋白质设计（protein engineering）或“反向折叠"，即给定一个蛋白质的三级结构，找出所有符合该结构的氨基酸序列
- 目前，蛋白质折叠动力学的机制尚未完全清楚，难以快速、准确地建立序列和结构之间的对应关系
- Li和Scheraga等曾用随机搜索方法确定多肽构象，但单纯构象搜索对于结构和自由度复杂得多的蛋白质无能为力

三级结构预测的发展方向

- 物化理论分析
  - 根据蛋白质天然构象处于热力学最稳定、能量最低状态的理论，计算蛋白质分子中所有原子间相互作用及蛋白质和溶剂间的相互作用
  - 通过能量最小化方法获得体系能量最低的构象，即从头预测（*ab initio* prediction）方法
  - 目前还缺乏有效的方法计算蛋白质构象的全局能量最小点
  - 复杂生物环境中，热力学条件下蛋白质周期性或无规则动态运动，与溶剂分子或配体小分子的相互作用，不同的溶剂环境、浓度环境，或者膜、凝胶、多孔吸附材料等界面条件，都可能导致实际条件下的蛋白质构象与理论计算的最稳结构有较大差距
  - 理论计算所得的结构也往往能在一定程度上反映出实际结构中的大部分特征，如二级结构、超二级结构、结构域的组成和相互作用等

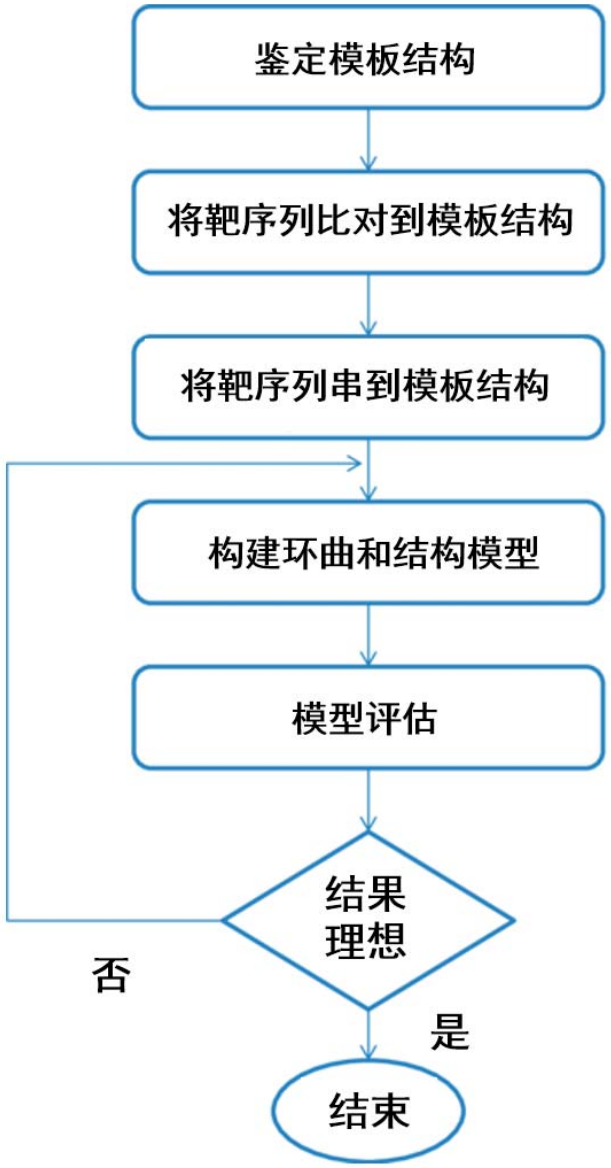
三级结构预测的发展方向

- 统计学方法
  - 对已有蛋白质的构象进行统计分析，从一级序列预测其二级结构进而构建三级结构
  - 目前二级结构预测的准确率已经达到80%左右，能够通过二级结构较为准确地搭建三级结构
  - 已成功地用于蛋白质的同源建模，即从与目标蛋白同源性较好的蛋白质三维结构出发，预测目标蛋白的三维结构
  - 同源建模是当前最被广泛使用的结构预测方法，但当找不到合适模板时，穿针引线法更为实用
  - 穿针引线法不需要同源性模版，并且克服二级结构预测不十分精确的困难，直接得到有参考价值的三维结构
- 两个发展方向的融合
  - 同源建模在寻找模板时作用明显
  - 穿针引线运用序列进化信息提高序列比对的精确度
  - 从头预测算法如AlphaFold2也能达到折叠识别的效果

同源建模法

- 比较建模（comparative modeling）法
  - 基于进化相关的序列具有相似的三维结构，且进化过程中三维结构比序列保守的原理，利用进化相关模板结构信息建模
- 同源建模法的计算流程
  - 将目标序列作为查询序列来搜索PDB和Swiss-Prot等已知蛋白质结构数据库，确定和识别一个同源模板，或选择已知结构的同源序列作为建模的模板
  - 将目标序列和模板序列进行比对，利用多种比对方法或手工校正以改进和优化靶序列和模板结构的比对，比对中可以加入空格
  - 以模板结构骨架作为模型，建立目标蛋白质骨架模型
  - 构建环区和侧链，优化侧链位置
  - 优化和评估产生的模型，使用能量最小化或其它方法优化结构，如利用分子动力学、模拟退火等优化结构

同源建模法的计算流程



靶序列

..pmllhvaaqiasgmrylat..

模板结构

序列比对

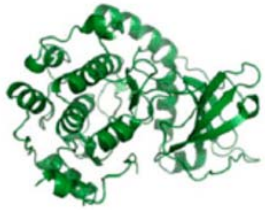
模板序列

..vvllymatqissameylek..

靶序列

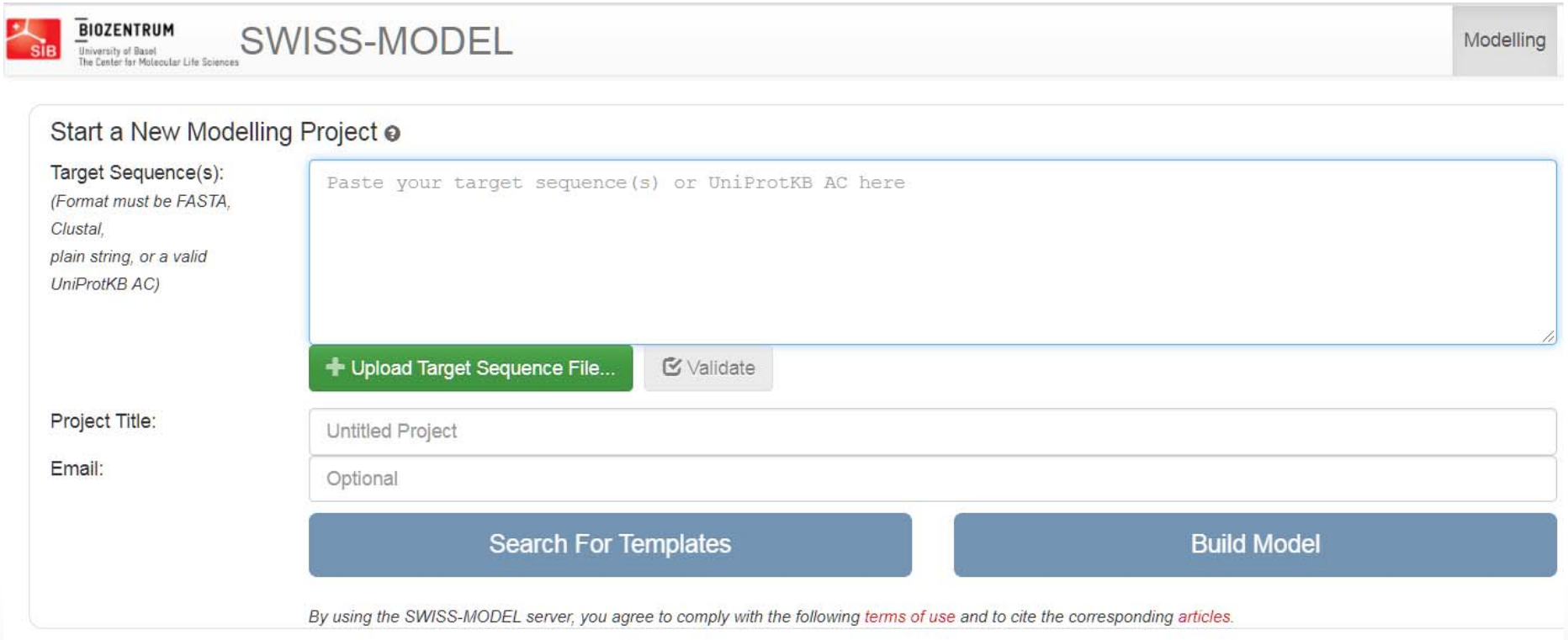
..pmllhvaaqiasgmrylat..

同源的模型



SWISS-MODEL

- 同源建模法的常用工具



同源建模法的适用范围

- 传统同源建模通过PSI-BLAST找到已知结构的相关蛋白
- 最大挑战是对模板链进行空隙和插入的建模
- 目标蛋白与模板结构保守性的程度及序列比对的正确性严重影响预测模型的准确性
  - 与模板一致性超过50%的序列建模通常较为可靠，其Cα原子位置与实验结构的平均偏差约1Å
  - 蛋白质序列一致性在30%～50%时，至少可共有80%的结构，在该范围的最好模型与实验结构中Cα原子位置平均偏差<4Å（典型为2～3Å)，且其误差主要在环区
  - 当序列一致性为20%～30%或甚至低于20%时，结构保守性可低至55%
- 比较建模主要在序列一致性大于30%的序列间进行

穿针引线法

- 结构在进化上的保守性要高于序列
  - 蛋白质折叠的类型有限，只有约1000种
  - 因此蛋白质三维结构预测问题可转化为：根据不同的模版，预测给定蛋白质的折叠类型，并进一步拼装成三级结构
  - 穿针引线法综合考虑两部分信息：能量函数和模板库（template library）
- 穿针引线法的计算流程
  - 将目标蛋白序列和已知的折叠进行匹配，根据比对的进化信息在已知的结构中找到一个或几个匹配最好的折叠结构，作为建模的模板
  - 将目标序列的“线”穿到模板的折叠结构上，拼装出最好的匹配模型，并使用能量函数优化模型。
  - 关键在于目标序列与已有折叠模板的比对，目标序列与折叠模板的相似性越高，预测模型就越可靠

穿针引线法的计算流程

靶序列

ALKKGF…HFDTSE

结构模板

1. 将给定序列与模板库做序列比较 （折叠库）
2. 评分准则：给定序列是否与模板的结构吻合（1D-3D谱）
3. 根据打分结果对模板适用性给予排序

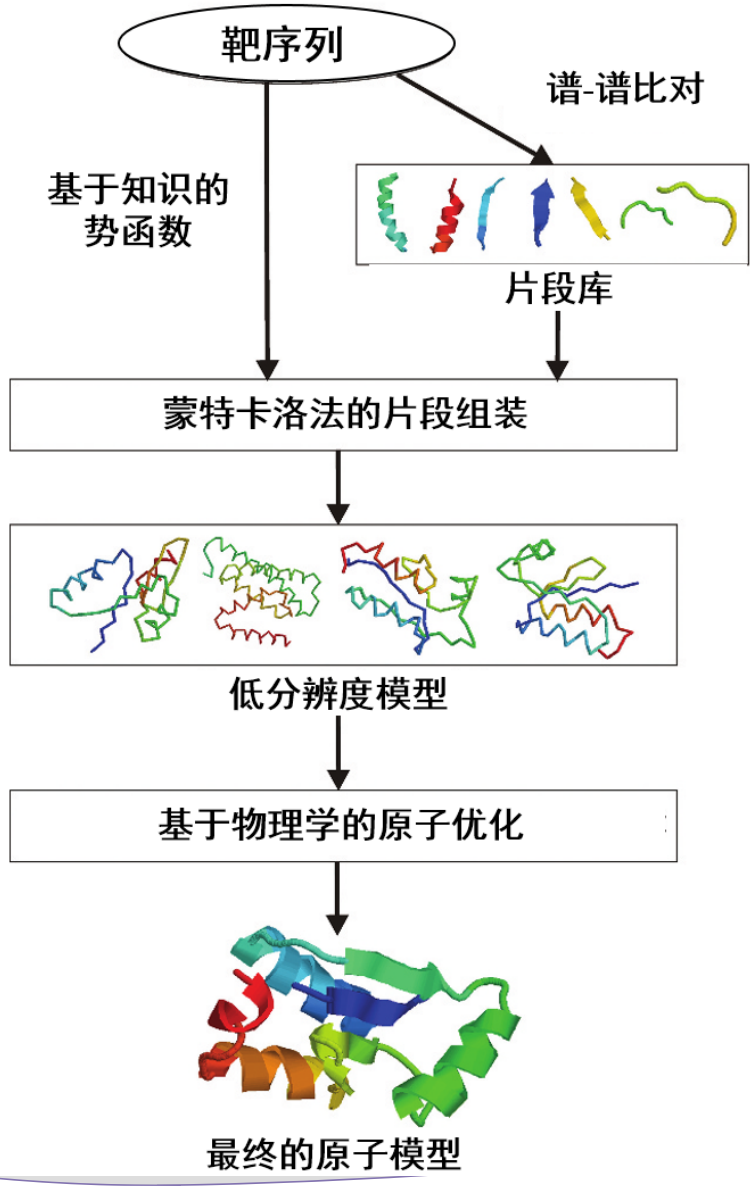
从头预测方法

- 目标蛋白序列缺乏已知结构的同源蛋白质
  - 理论依据：Anfinsen原理，蛋白质的天然结构对应其自由能最低的状
  - 从头预测法：需要设计能量函数，包括氨基酸残基间的键能（bond energy）、键的转角能（bond angle energy）、二面角能（dihedral angle energy）、范德华力（van der Waals energy）和静电力（electrostatic energy）等
- 从头预测依赖以下因素
  - 通过能量优化找到的蛋白质结构具有充分的结构可靠性和计算可控性
  - 符合实际的力场或其它作用力描述方法
  - 高效而准确的搜索构象空间重要区域的算法
  - 对获得结构进行准确评估的方法

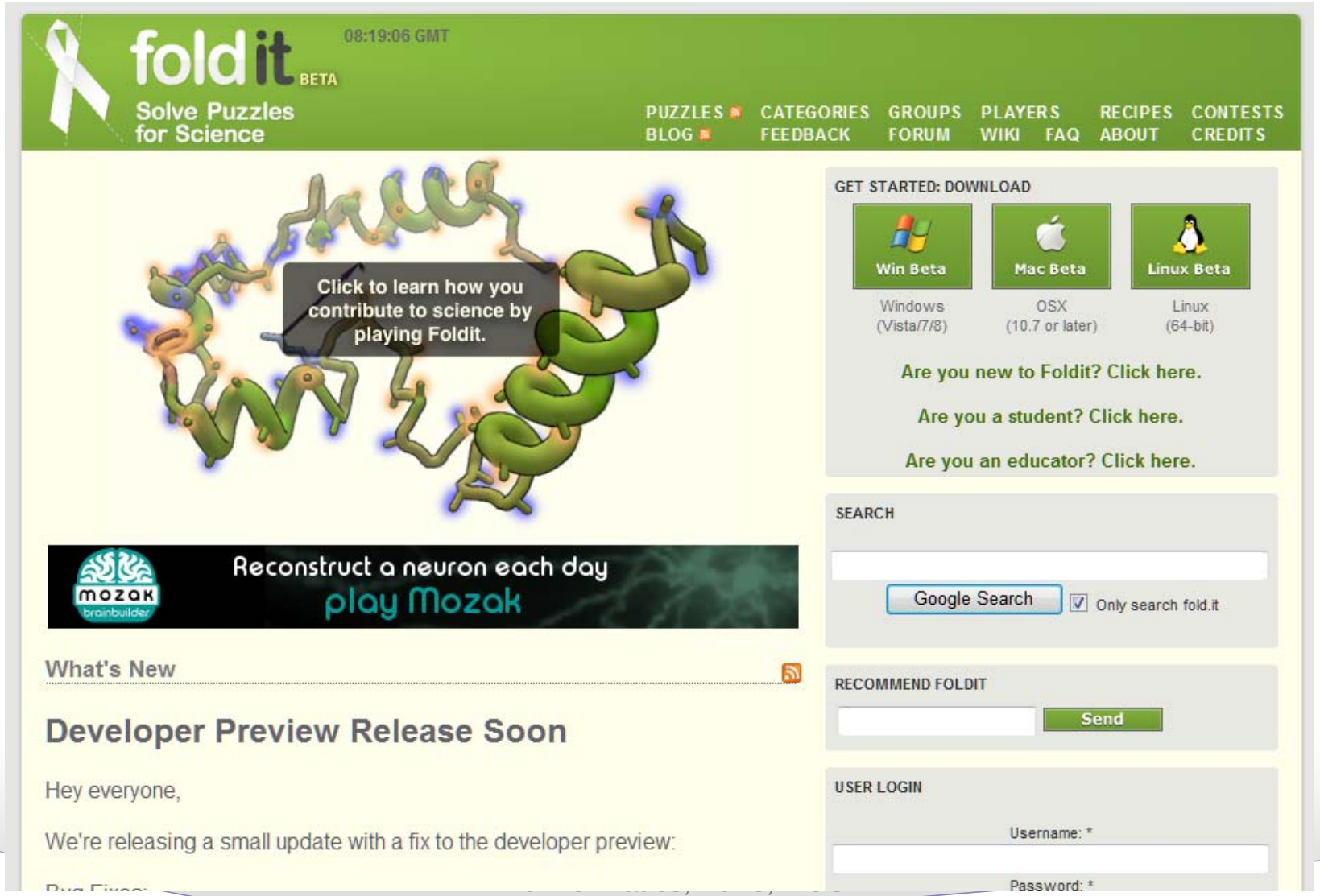
ROSETTA

- 1999年，美国David Baker设计从头预测工具ROSETTA
- 基本原理
  - 长度为3-9个氨基酸残基组成的短肽段库，能够反映各种肽段在局部范围内的三级结构
  - 针对给定蛋白质，寻找各种短肽段组合
  - 结合能量函数予以优化，从而模拟蛋白质的三级结构

ROSETTA法的计算原理



Foldit


foldit BETA
08:19:06 GMT
Solve Puzzles for Science
PUZZLES
BLOG
CATEGORIES
FEEDBACK
GROUPS
FORUM
PLAYERS
WIKI
FAQ
RECIPES
ABOUT
CONTESTS
CREDITS
Click to learn how you contribute to science by playing Foldit.
GET STARTED: DOWNLOAD
Win Beta
Mac Beta
Linux Beta
Windows (Vista/7/8)
OSX (10.7 or later)
Linux (64-bit)
Are you new to Foldit? Click here.
Are you a student? Click here.
Are you an educator? Click here.
mozak brainbuilder
Reconstruct a neuron each day
play Mozak
SEARCH
Google Search
Only search fold.it
What's New
RECOMMEND FOLDIT
Send
Developer Preview Release Soon
USER LOGIN
Hey everyone,
Username: *
We're releasing a small update with a fix to the developer preview:
Password: *

结构基因组学

PROTEIN STRUCTURE

Ovchinnikov *et al.*, *Science* **355**, 294-298 (2017) 20 January 2017

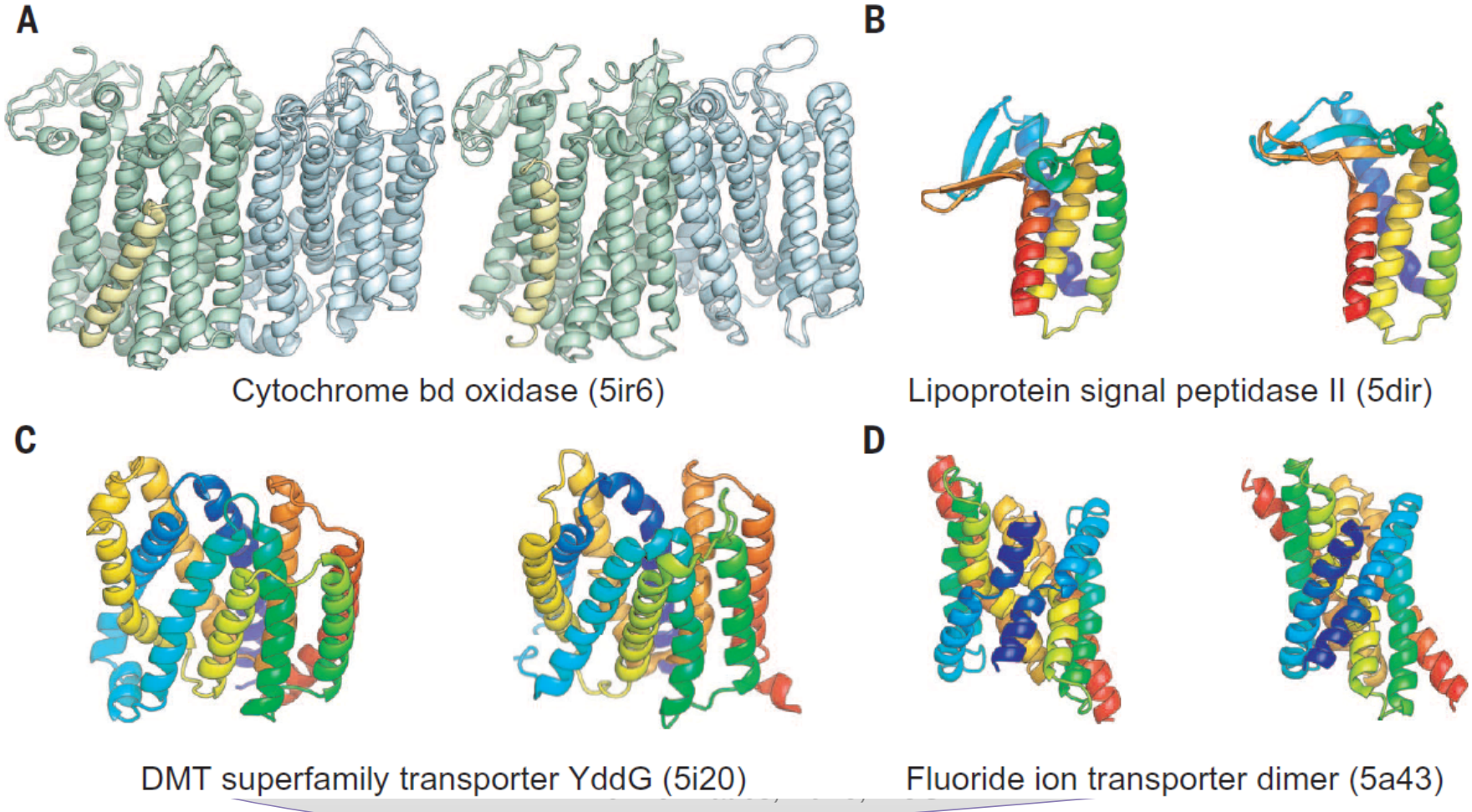
Protein structure determination using metagenome sequence data

Sergey Ovchinnikov,[1,2,3] Hahnbeom Park,[1,2] Neha Varghese,[4] Po-Ssu Huang,[1,2] Georgios A. Pavlopoulos,[4] David E. Kim,[1,5] Hetunandan Kamisetty,[6] Nikos C. Kyrpides,[4,7] David Baker[1,2,5]*

Despite decades of work by structural biologists, there are still ~5200 protein families with unknown structure outside the range of comparative modeling. We show that Rosetta structure prediction guided by residue-residue contacts inferred from evolutionary information can accurately model proteins that belong to large families and that metagenome sequence data more than triple the number of protein families with sufficient sequences for accurate modeling. We then integrate metagenome data, contact-based structure matching, and Rosetta structure calculations to generate models for 614 protein families with currently unknown structures; 206 are membrane proteins and 137 have folds not represented in the Protein Data Bank. This approach provides the representative models for large protein families originally envisioned as the goal of the Protein Structure Initiative at a fraction of the cost.

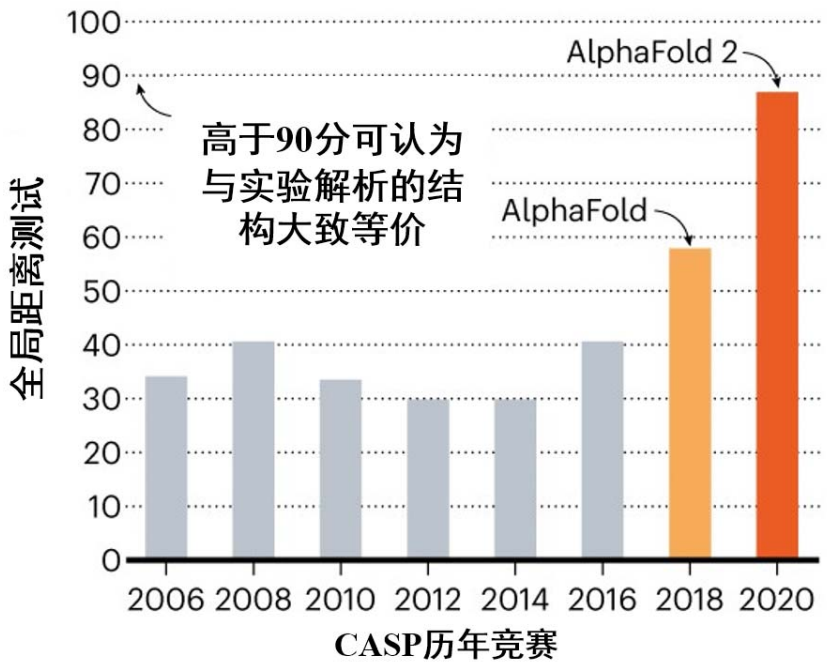
结构基因组学

- 左：预测；右：实验



DeepMind’s AlphaFold2

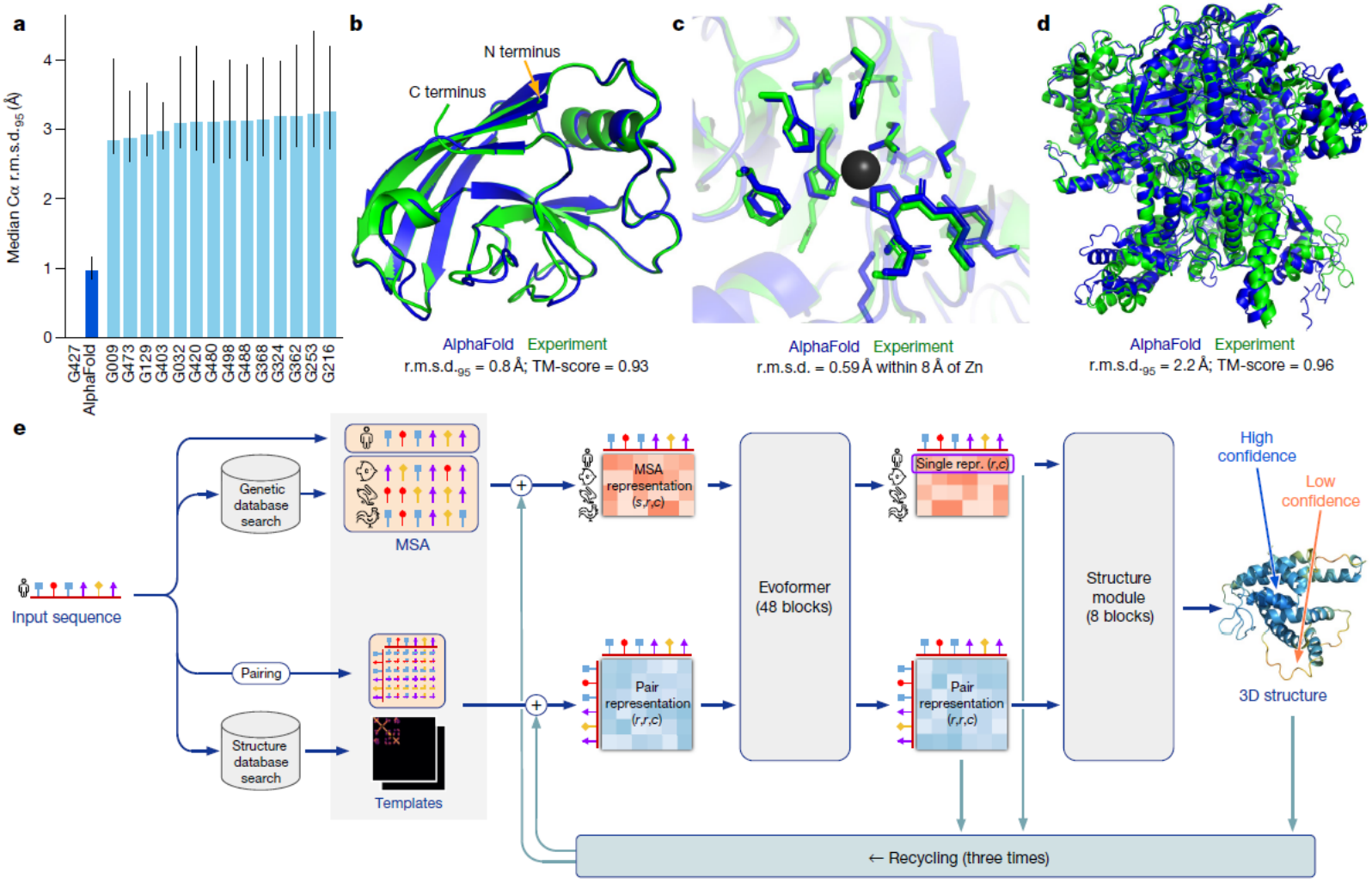
- 2020年，Demis Hassabis等
  - 128 TPU v3
  - 初始模型训练：1周
  - 优化调整：4天
- 综合考虑多序列比对的进化信息
- 氨基酸对的物理和几何限制性特征
- 构建了新的深度学习框架Evoformer
- 准确性与实验技术相当



AlphaFold2在CASP14竞赛中得到90分

算法原理

❒ 特征：多序列比对 & 氨基酸对



Evoformer架构