生物信息学

蛋白质结构预测

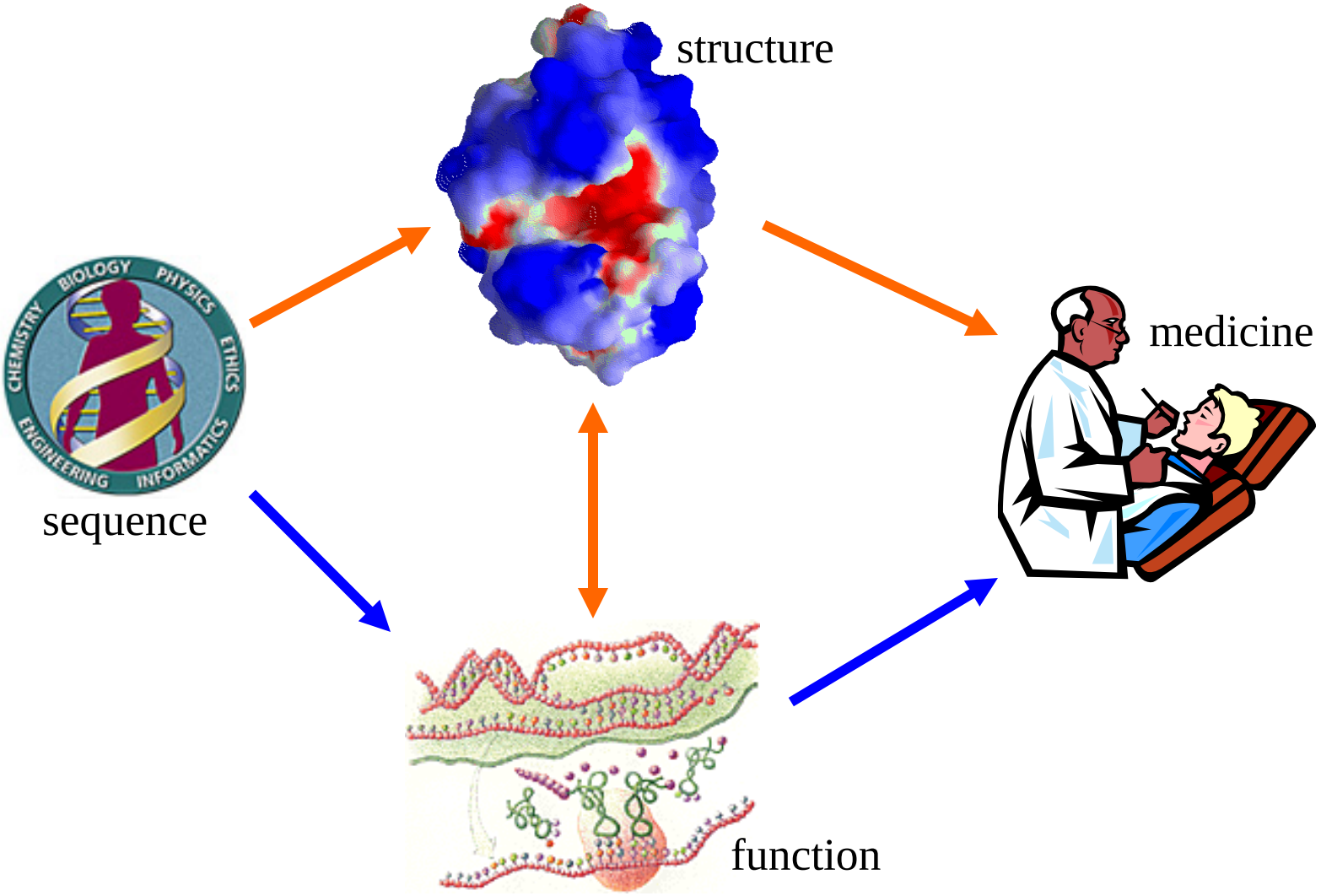
吴凌云

中国科学院数学与系统科学研究院

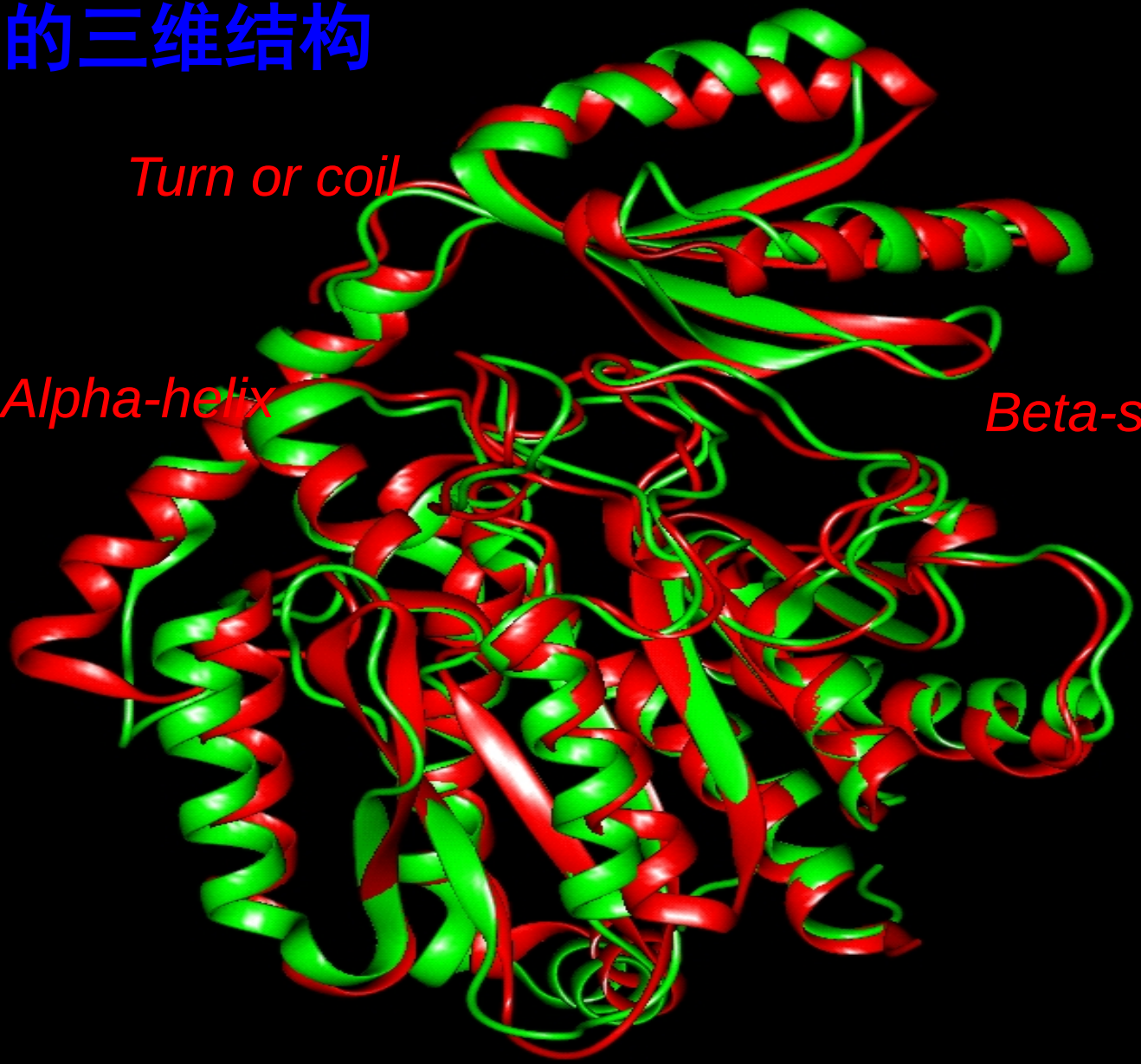
目录

- 蛋白质的三维结构
- 蛋白质结构预测的重要性
- 蛋白质二级结构预测方法
- 蛋白质三级结构预测方法
- 研究趋势

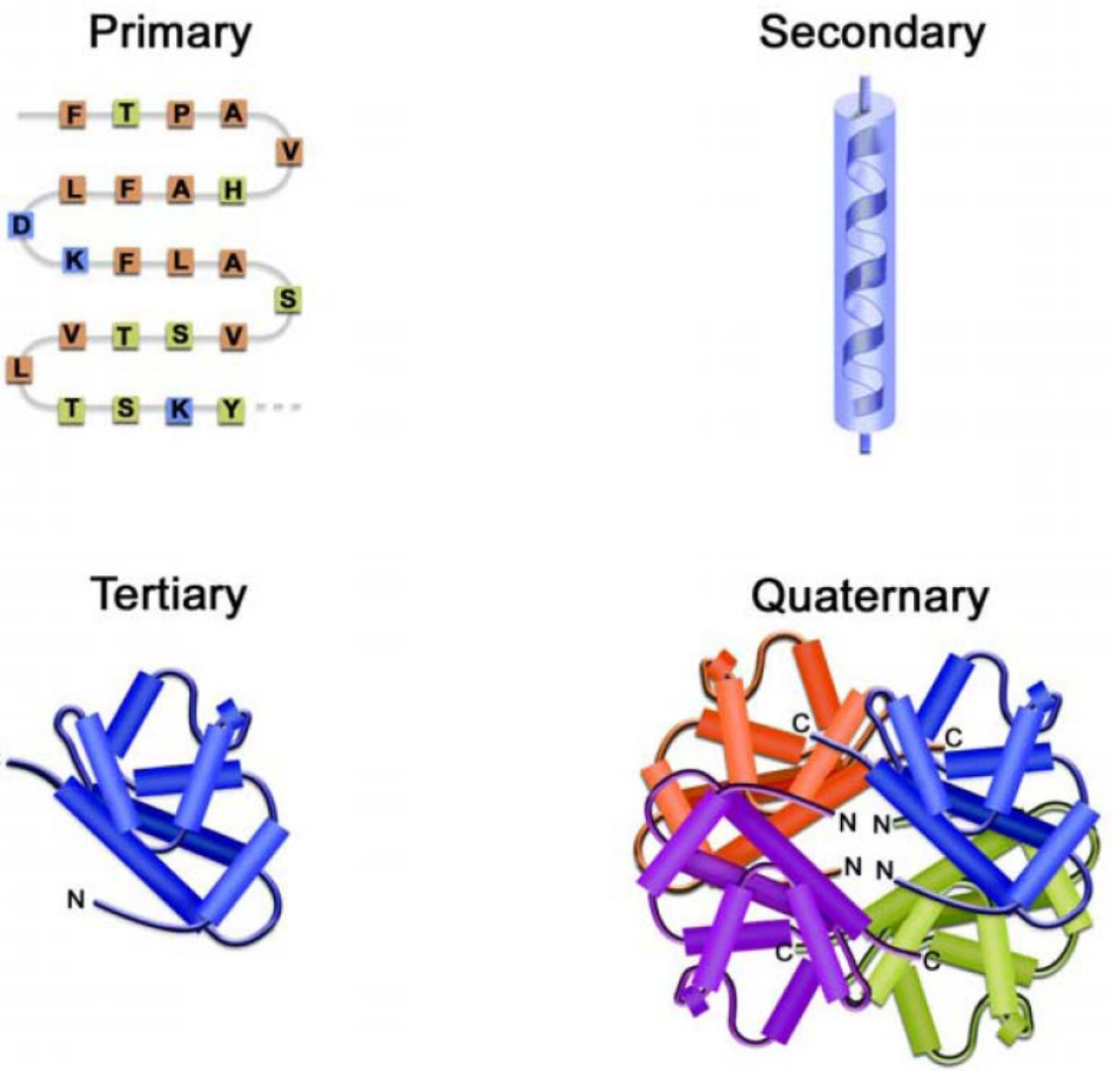
蛋白质的结构



蛋白质的三维结构

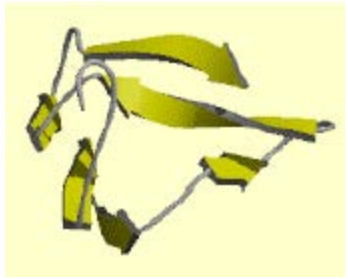


蛋白质的结构层次


Primary
F T P A
V
L F A H
D
K F L A
S
V T S V
L
T S K Y
Secondary
Tertiary
C
N
Quaternary
C
C
N N
N N
C
C

蛋白质二级结构

- α-helix (30-35%)
  α-螺旋
- β-sheet / β-strand (20-25%)
  β-折叠
- Coil (40-50%) 无规则卷曲
- Loop 环
- β-turn β-转角

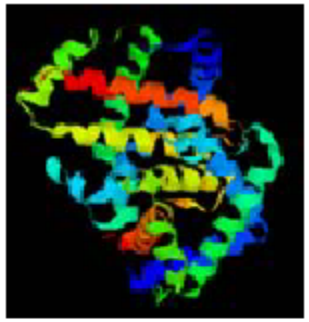


蛋白质结构分类法

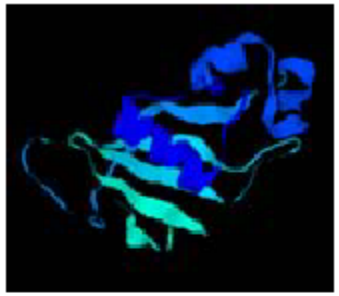
- Class (<10) 结构类
- Folds (<1000) 折叠子
- Superfamily 超家族
  - 序列或结构相似
- Family 家族
  - 序列相似性 > 25% – 30%
  - 同源 Homology

蛋白质结构类

Class α

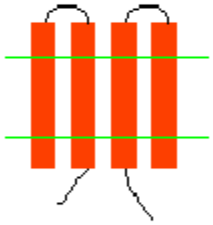


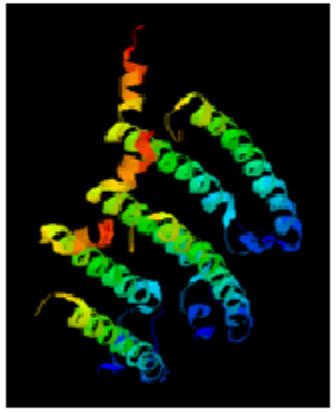
Class α + β



Class β

Membrane





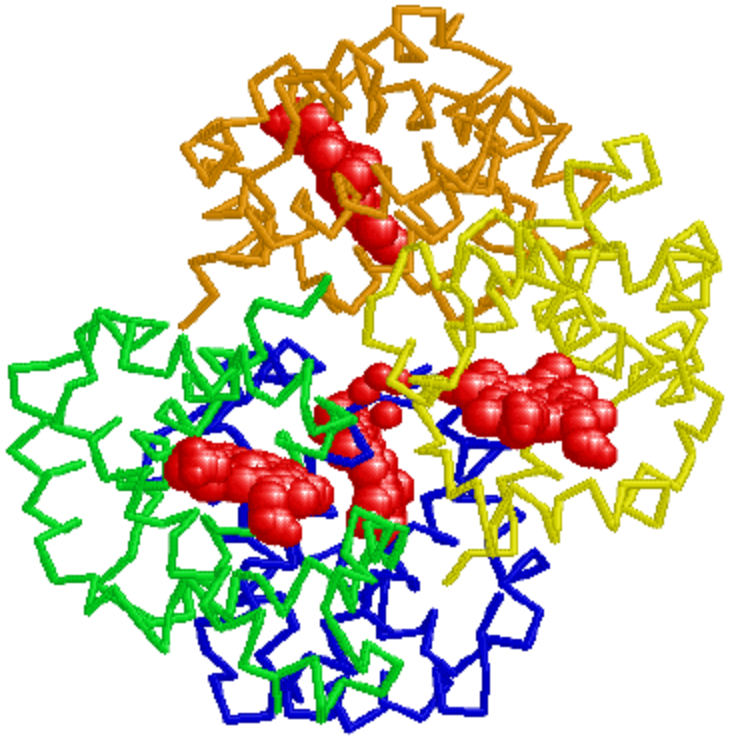
Class α / β

目录

- 蛋白质的三维结构
- **蛋白质结构预测的重要性**
- 蛋白质二级结构预测方法
- 蛋白质三级结构预测方法
- 研究趋势

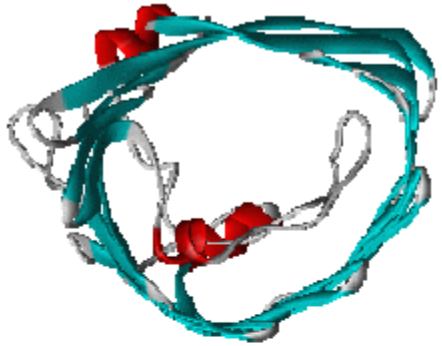
蛋白质结构与功能

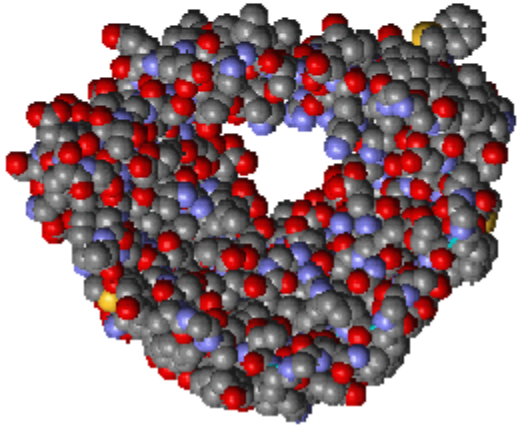
- Hemoglobin (血红蛋白)
  - 运输氧气



蛋白质结构与功能

- Porin (孔蛋白)
  - 穿透细胞膜的运输



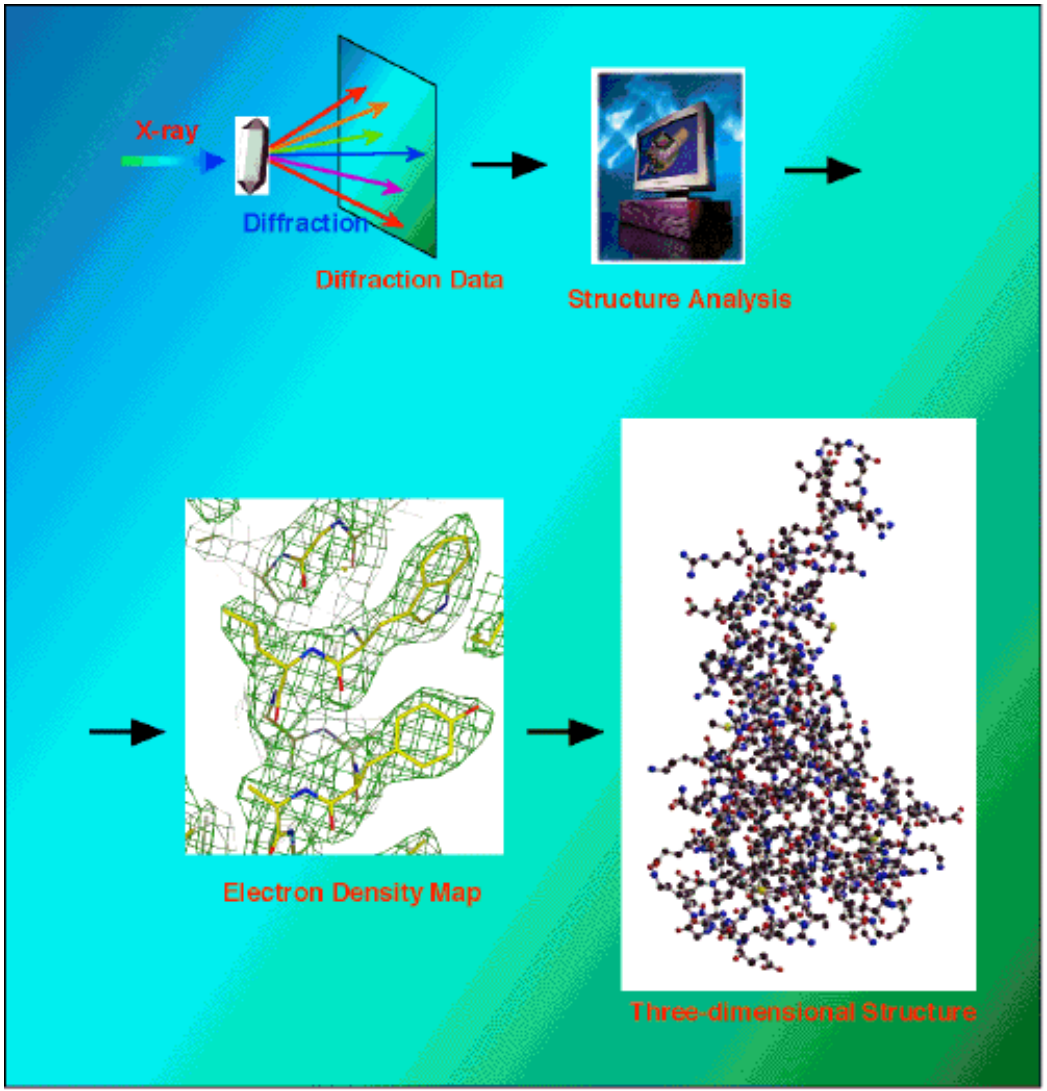


蛋白质三维结构的测定

X-射线衍射法

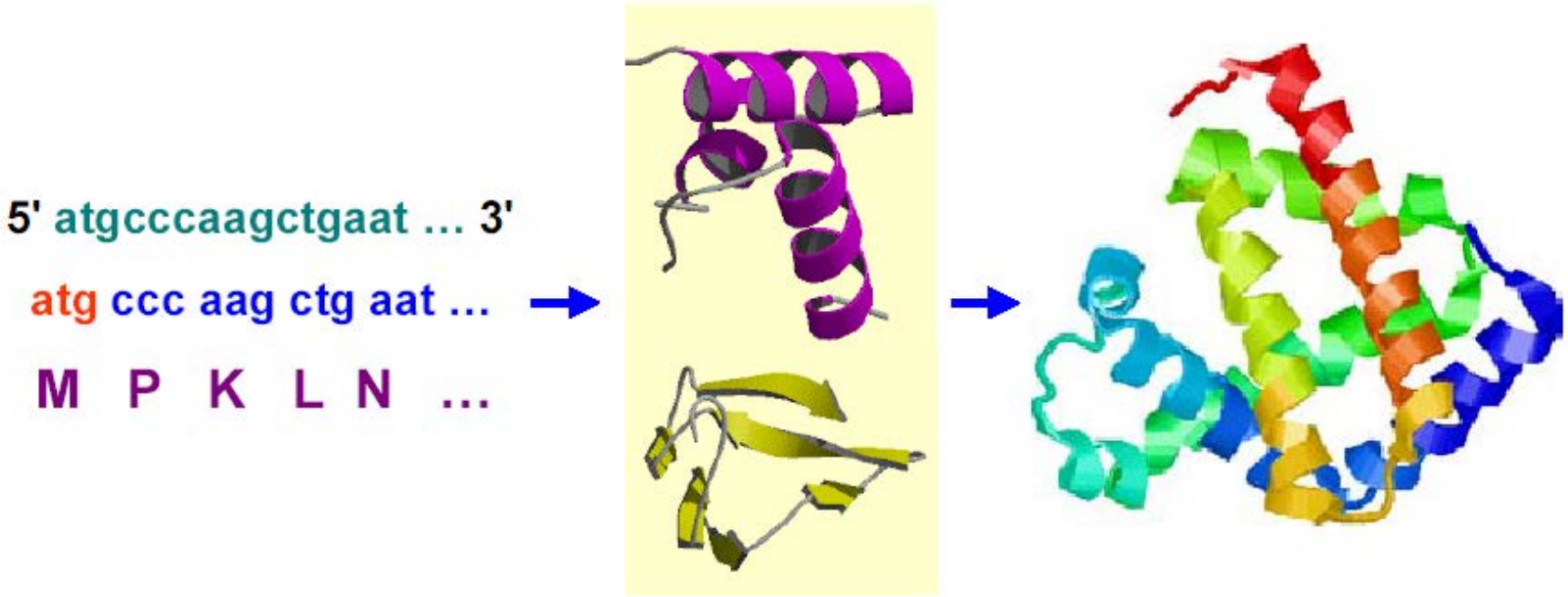
核磁共振法

- 费用昂贵
- 速度慢
- 对某些蛋白质无法使用



蛋白质结构预测

- 从蛋白质一级序列（氨基酸）序列预测蛋白质三维结构
- 序列 → 二级结构 → 三维结构 → 功能



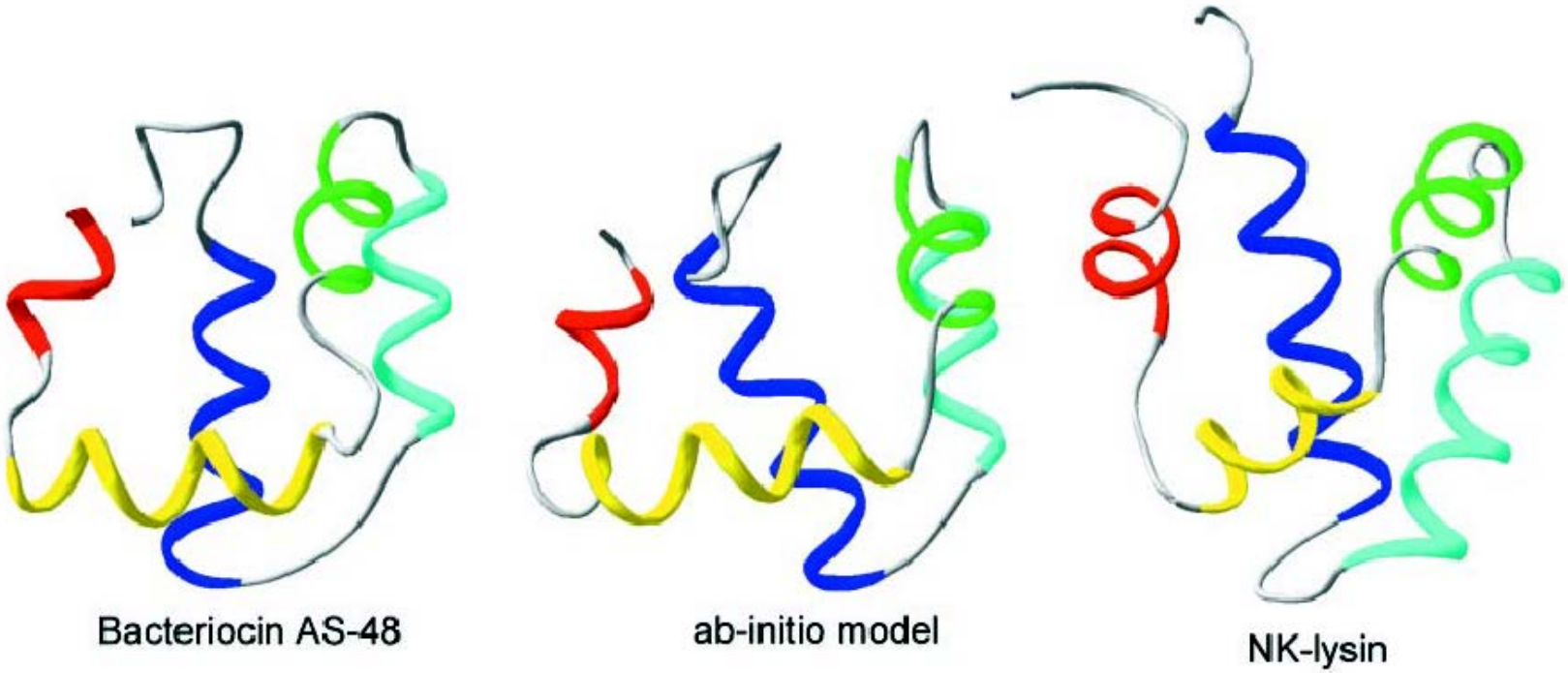
必要性与可行性

- 必要性
  - DNA序列数据 >> 蛋白质序列数据 >> 蛋白质结构数据

	1994	1997	2002	2005	2010
序列 (Swiss-Prot)	40,000	68,000	114,033	192,799	515,203
结构 (PDB)	4,045	7,000	18,838	32,434	64,500

  - 意义重大：蛋白质功能预测、药物设计...
- 可行性
  - 蛋白质的序列信息唯一地决定三维结构
  - 序列相似性意味着结构相似性

序列—结构—功能



左边的Bacteriocin是CASP-4测试中的T0102试题。中间的结果是Baker研究小组用从头预测方法Rosetta获得的。右边是用中间的结构在数据库中进行搜索得到的最相似的蛋白质结构。

目录

- 蛋白质的三维结构
- 蛋白质结构预测的重要性
- **蛋白质二级结构预测方法**
- 蛋白质三级结构预测方法
- 研究趋势

二级结构预测

- Given a protein sequence $a_1a_2…a_N$, secondary structure prediction aims at defining the state of each amino acid a_i as being either H (helix), E (extended=strand), or O (other) (Some methods have 4 states: H, E, T for turns, and O for other).
- The quality of secondary structure prediction is measured with a “3-state accuracy” score, or Q3. Q3 is the percent of residues that match “reality” (X-ray structure).

二级结构的确定

- Determine Secondary Structure positions in known protein structures using DSSP or STRIDE:
  - Kabsch and Sander. Dictionary of Secondary Structure in Proteins: pattern recognition of hydrogen-bonded and geometrical features. Biopolymer 22: 2571-2637 (1983) (DSSP)
  - Frischman and Argos. Knowledge-based secondary structure assignments. Proteins, 23:566-571 (1995) (STRIDE)

Q3

ALHEASGPSVILFGSDVTVPPASNAEQAK Amino acid sequence

hhhhhooooeeeeoooeeeooooohhhhh

Actual Secondary Structure

ohhhooooeeeeooooooeeeoooohhhhhh **Q3=22/29=76%**

(useful prediction)

hhhhhoooohhhhooohhhooooohhhhh **Q3=22/29=76%**

(terrible prediction)

- Q3 for random prediction is 33%
- Secondary structure assignment in real proteins is uncertain to about 10%; Therefore, a “perfect” prediction would have Q3=90%.

早期的二级结构预测方法

- Chou and Fasman
  - Chou and Fasman. Prediction of protein conformation. Biochemistry, 13: 211-245, 1974
- GOR
  - Garnier, Osguthorpe and Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol., 120:97-120, 1978

Chou and Fasman

Start by computing amino acids **propensities** to belong to a given type of secondary structure:

$$\frac{P(i|Helix)}{P(i)} \qquad \frac{P(i|Beta)}{P(i)} \qquad \frac{P(i|Turn)}{P(i)}$$

Propensities > 1 mean that the residue type *i* is likely to be found in the corresponding secondary structure type.

Amino Acids Propensity

Amino Acid	α-Helix	β-Sheet	Turn
Ala	1.29	0.90	0.78
Cys	1.11	0.74	0.80
Leu	1.30	1.02	0.59
Met	1.47	0.97	0.39
Glu	1.44	0.75	1.00
Gln	1.27	0.80	0.97
His	1.22	1.08	0.69
Lys	1.23	0.77	0.96
Val	0.91	1.49	0.47
Ile	0.97	1.45	0.51
Phe	1.07	1.32	0.58
Tyr	0.72	1.25	1.05
Trp	0.99	1.14	0.75
Thr	0.82	1.21	1.03
Gly	0.56	0.92	1.64
Ser	0.82	0.95	1.33
Asp	1.04	0.72	1.41
Asn	0.90	0.76	1.23
Pro	0.52	0.64	1.91
Arg	0.96	0.99	0.88

Favors α-Helix

Favors β-strand

Favors turn

Chou and Fasman

- Predicting helices:
  - find nucleation site: 4 out of 6 contiguous residues with P(α)>1
  - extension: extend helix in both directions until a set of 4 contiguous residues has an average P(α) < 1 (breaker)
  - if average P(α) over whole region is >1, it is predicted to be helical

Chou and Fasman

- Predicting strands:
  - find nucleation site: 3 out of 5 contiguous residues with P(β)>1
  - extension: extend strand in both directions until a set of 4 contiguous residues has an average P(β) < 1 (breaker)
  - if average P(β) over whole region is >1, it is predicted to be a strand

Chou and Fasman

Position-specific parameters for turn:

Each position has distinct amino acid preferences.

Examples:

- At position 2, Pro is highly preferred; Trp is disfavored
- At position 3, Asp, Asn and Gly are preferred
- At position 4, Trp, Gly and Cys preferred

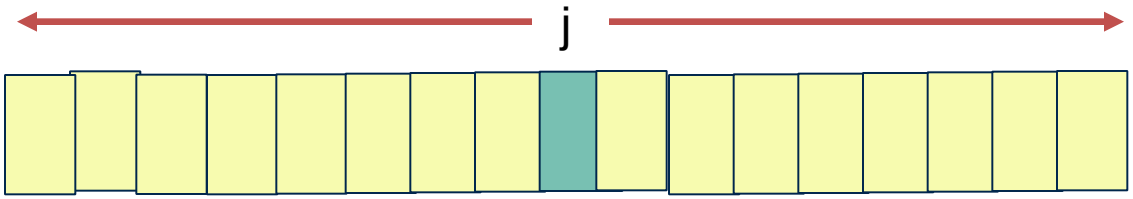
	f(i)	f(i+1)	f(i+2)	f(i+3)
Ala	0.060	0.076	0.035	0.058
Arg	0.070	0.106	0.099	0.085
Asp	0.147	0.110	0.179	0.081
Asn	0.161	0.083	0.191	0.091
Cys	0.149	0.050	0.117	0.128
Glu	0.056	0.060	0.077	0.064
Gln	0.074	0.098	0.037	0.098
Gly	0.102	0.085	0.190	0.152
His	0.140	0.047	0.093	0.054
Ile	0.043	0.034	0.013	0.056
Leu	0.061	0.025	0.036	0.070
Lys	0.055	0.115	0.072	0.095
Met	0.068	0.082	0.014	0.055
Phe	0.059	0.041	0.065	0.065
Pro	0.102	0.301	0.034	0.068
Ser	0.120	0.139	0.125	0.106
Thr	0.086	0.108	0.065	0.079
Trp	0.077	0.013	0.064	0.167
Tyr	0.082	0.065	0.114	0.125
Val	0.062	0.048	0.028	0.053

Chou and Fasman

- Predicting turns:
  - for each tetrapeptide starting at residue i, compute:
    - P_{Turn} (average propensity over all 4 residues)
    - F = f(i)*f(i+1)*f(i+2)*f(i+3)
  - if $P_{Turn} > P_{\alpha}$ and $P_{Turn} > P_{\beta}$ and $P_{Turn} > 1$ and F>0.000075 tetrapeptide is considered a turn.

http://fasta.bioch.virginia.edu/fasta_www/chofas.htm

Position-dependent propensities for helix, sheet or turn is calculated for each amino acid. For each position j in the sequence, eight residues on either side are considered.



A helix propensity table contains information about propensity for residues at 17 positions when the conformation of residue j is helical. The helix propensity tables have 20 x 17 entries.

Build similar tables for strands and turns.

The predicted state of AA_j is calculated as the sum of the position-dependent propensities of all residues around AA_j.

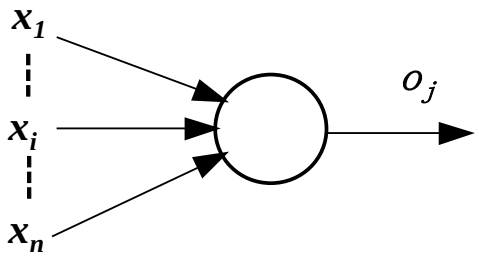
GOR can be used at : http://gor.bb.iastate.edu (GOR5, 2005)

二级结构预测方法的发展

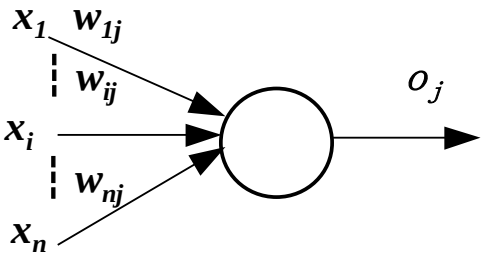
- 基于碱基的统计信息
  - Chou-Fasman, GOR
- 复杂统计信息
  - GOR3, Qian-Sejnowski
- 同源序列信息
  - PHD, PREDATOR

神经元的数学模型

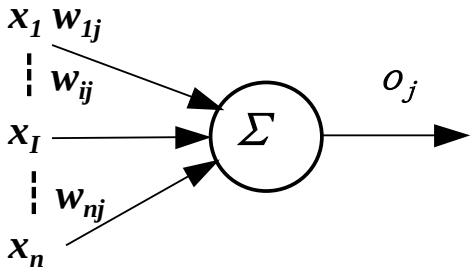
- 模型假设：
  - 每个神经元都是一个多输入单输出的信息处理单元；
  - 神经元输入分兴奋性输入和抑制性输入两种类型；
  - 神经元具有空间整合特性和阈值特性；
  - 神经元输入与输出间有固定的时滞，主要取决于突触延搁；
  - 忽略时间整合作用和不应期；
  - 神经元本身是非时变的，即其突触时延和突触强度均为常数。



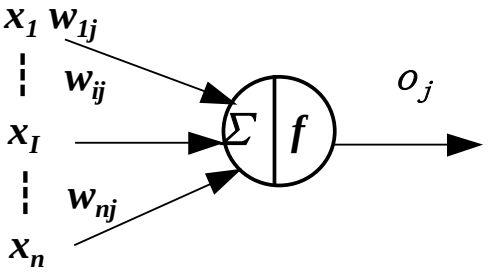
(a)多输入单输出



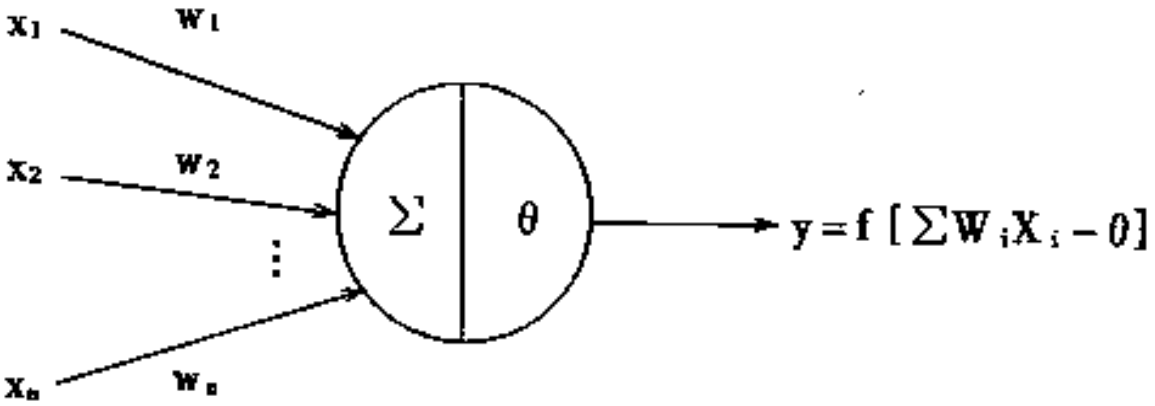
(b)输入加权



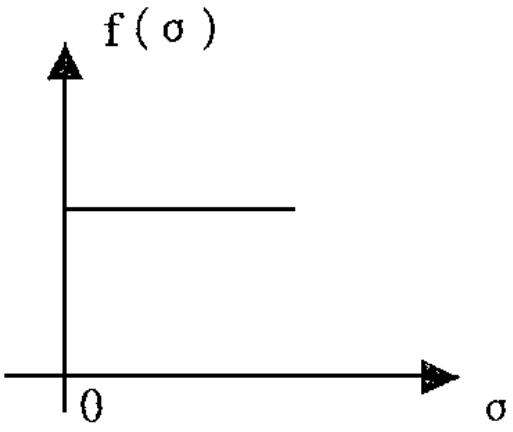
(c)输入加权求和

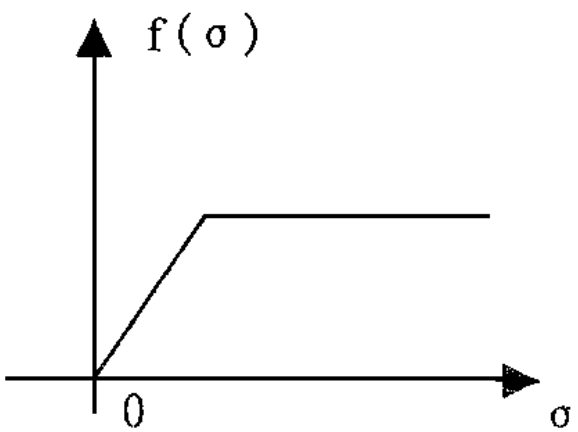


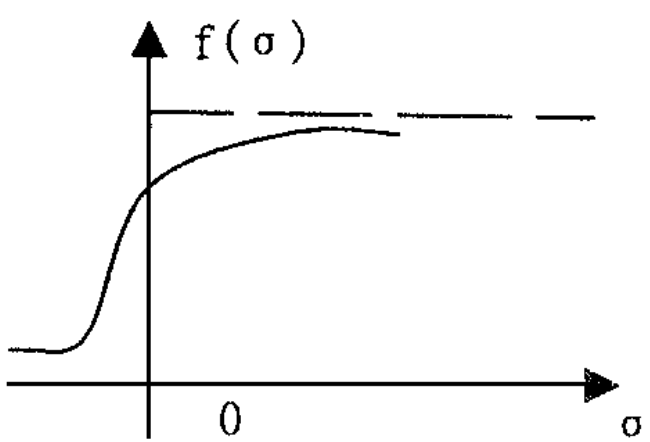
(d)输入-输出函数



神经元函数

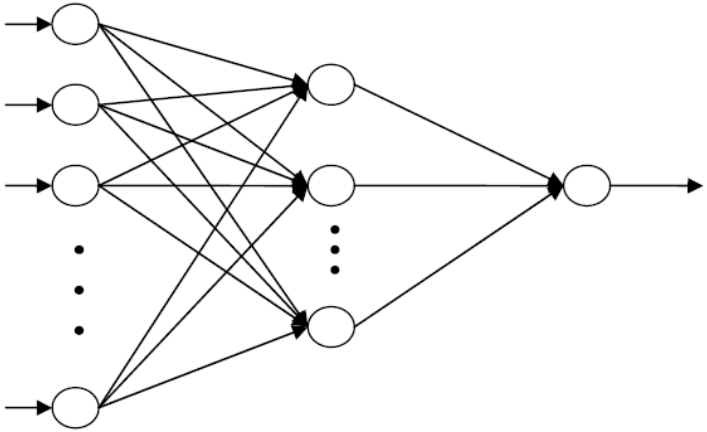


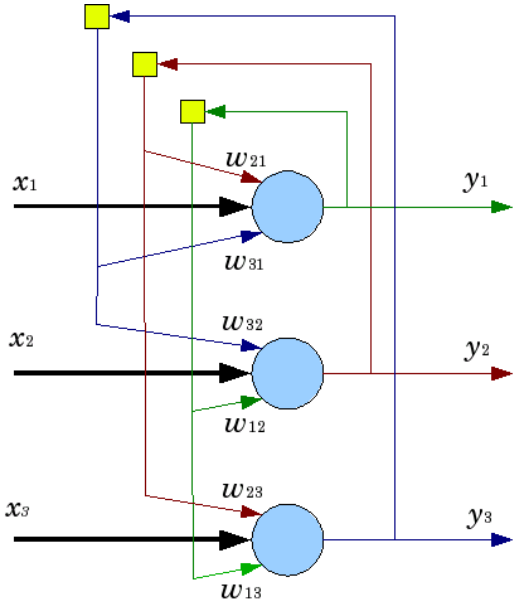




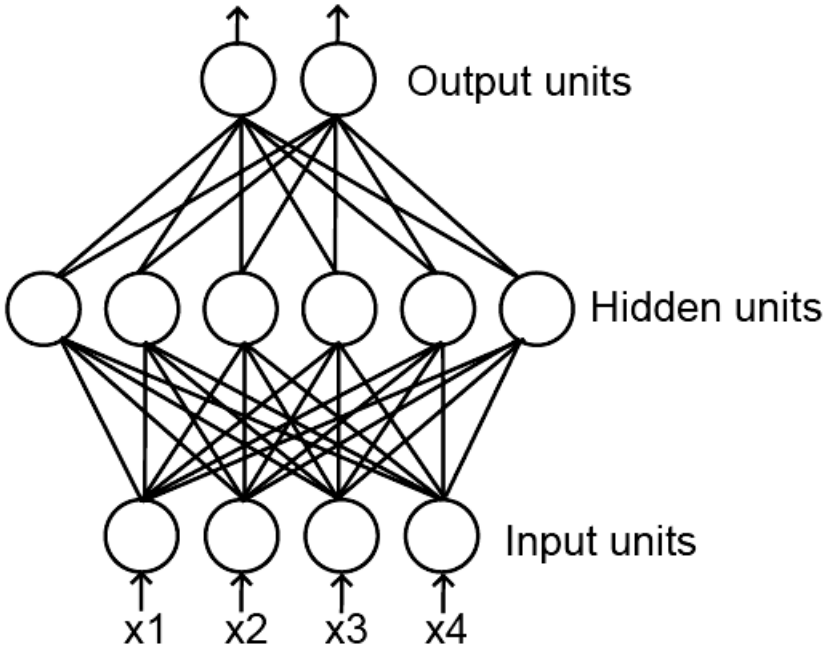
神经网络的分类

- **前向神经网络（BP网）**
- 回馈神经网络（Hopfield网）
- 自组织神经网络（SOM）

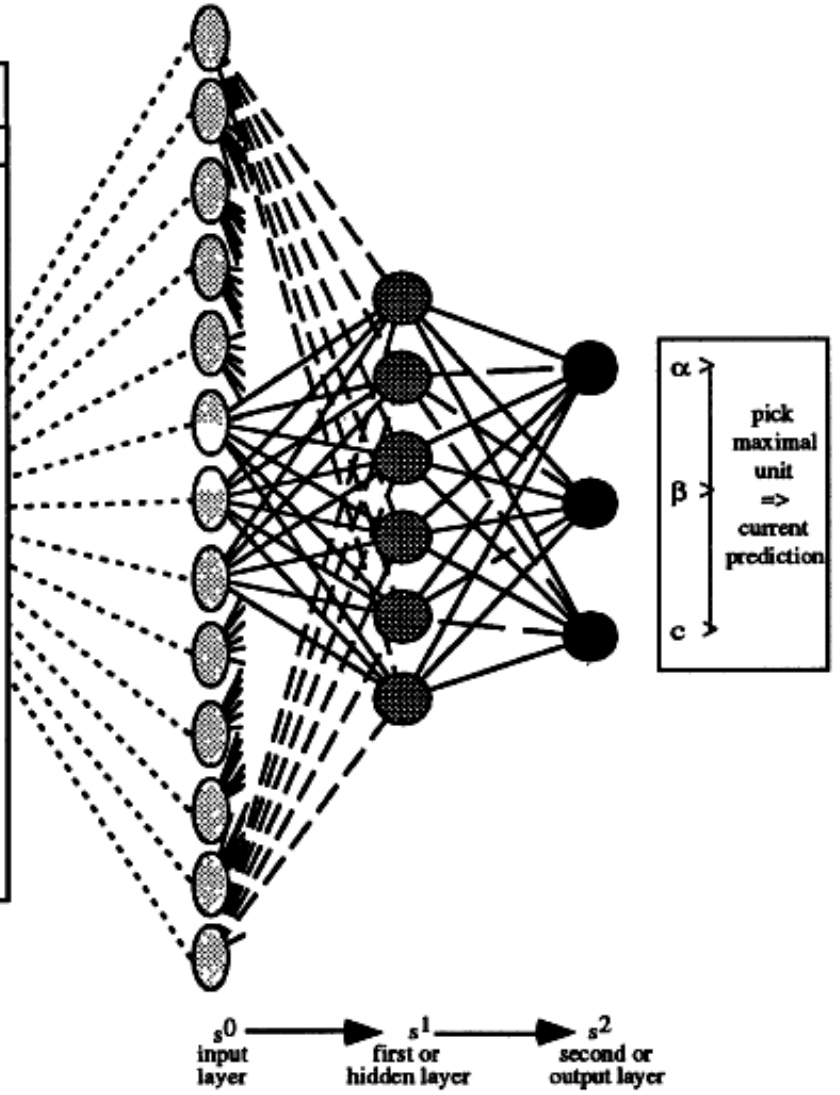




BP神经网络



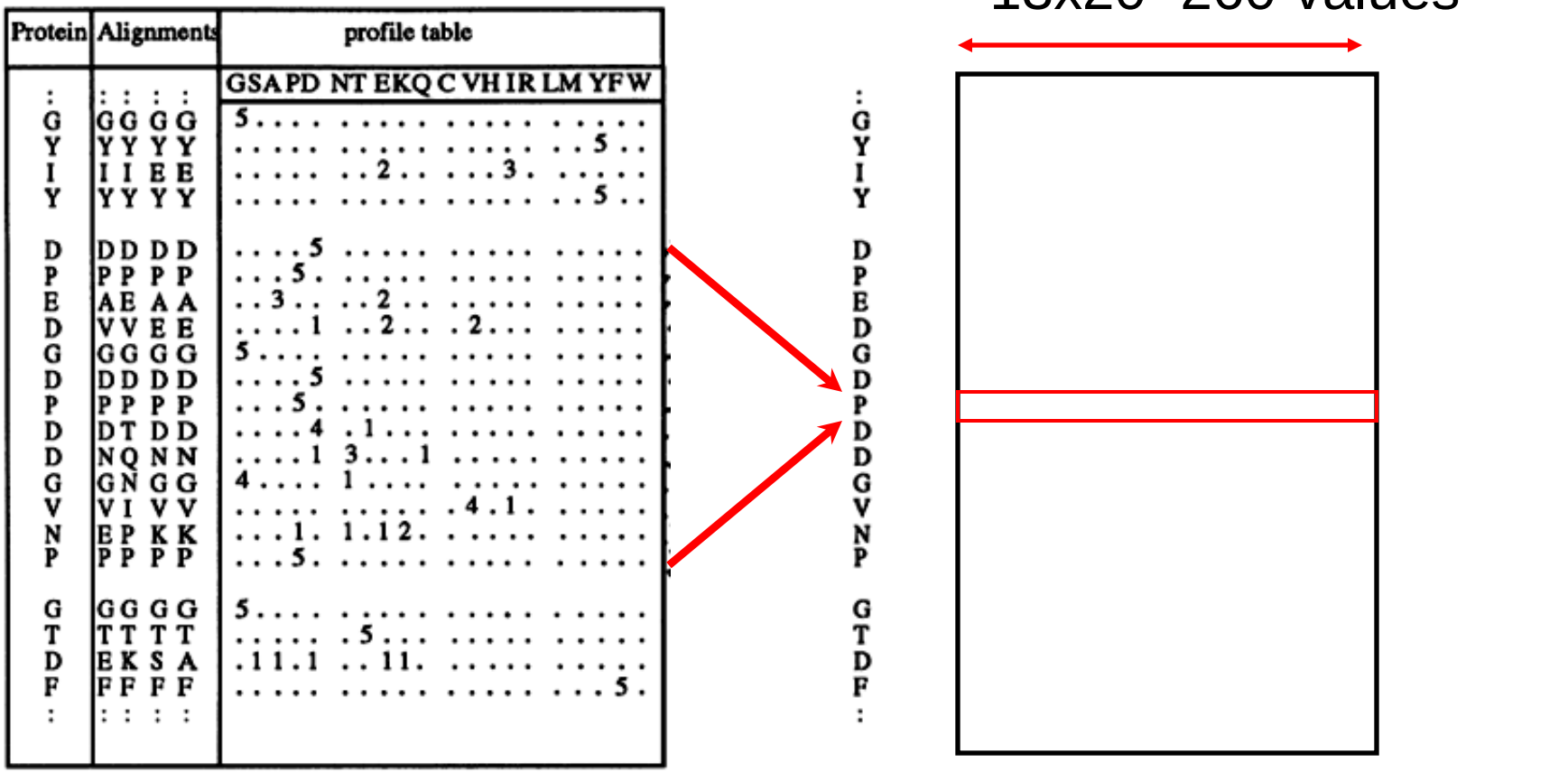
Protein	Alignments	profile table
		GSAPD NTEKQ CVHIR LMYFW
:	: : : :	
G	G G G G	5. . . . .
Y	Y Y Y Y	. . . . 5 . .
I	I I E E	. . . . 2 . . . . . 3 . . . . .
Y	Y Y Y Y	. . . . 5 . .
D	D D D D	. . . . 5 . . . .
P	P P P P	. . . 5 . . . .
E	A E A A	. . 3 . . . . 2 . . . .
D	V V E E	. . . . 1 . . 2 . . . 2 . . . .
G	G G G G	5 . . . .
D	D D D D	. . . . 5 . . . .
P	P P P P	. . . 5 . . . .
D	D T D D	. . . . 4 . 1 . . . .
D	N Q N N	. . . . 1 3 . . . 1 . . . .
G	G N G G	4 . . . . 1 . . . .
V	V I V V	. . . . 4 . 1 . . . . . .
N	E P K K	. . . 1 . 1 . 1 2 . . . .
P	P P P P	. . . 5 . . . .
G	G G G G	5 . . . .
T	T T T T	. . . . . . 5 . . . .
D	E K S A	. 1 1 . 1 . . 1 1 . . . .
F	F F F F	. . . . 5 .
:	: : : :	



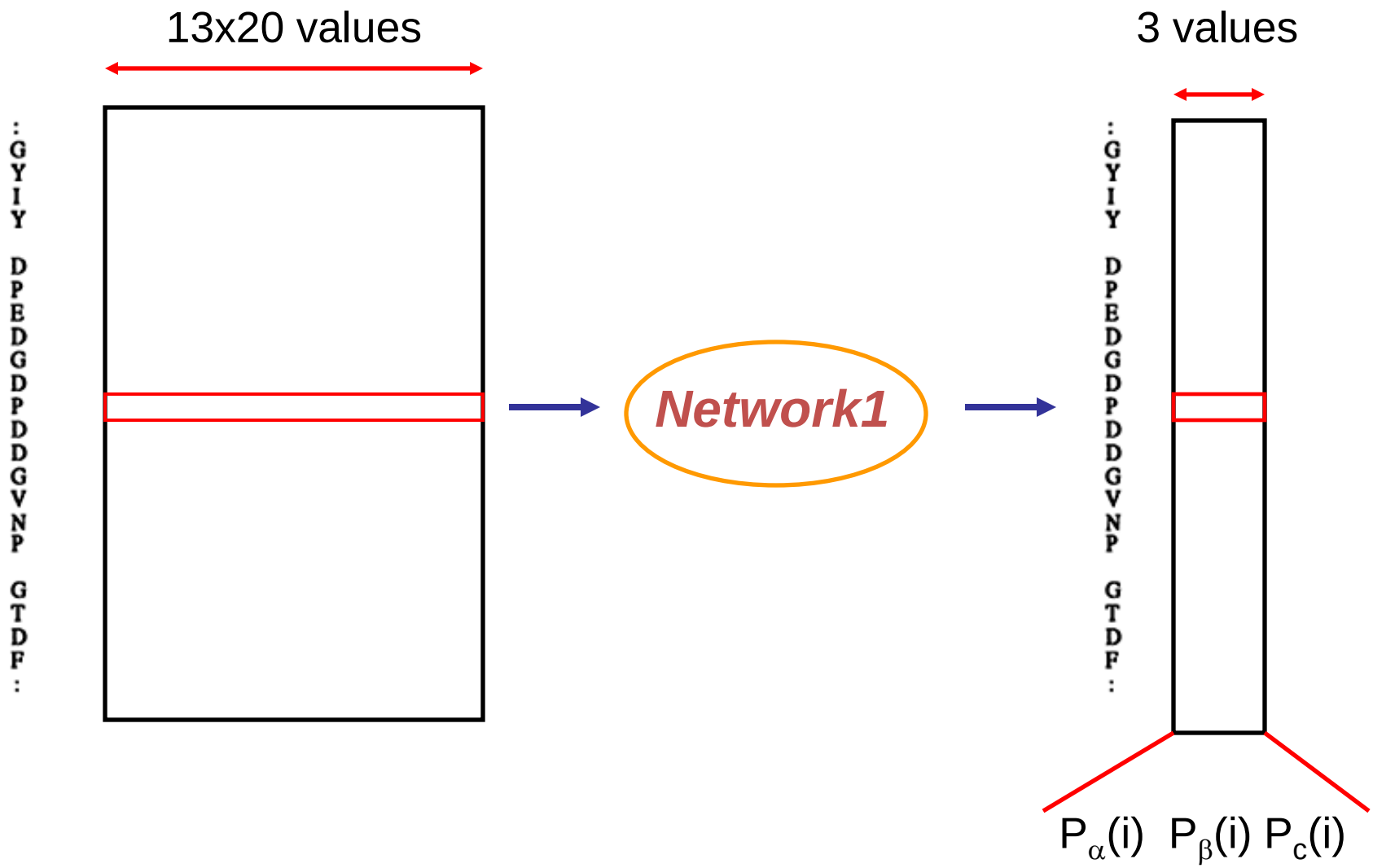
PHD: 输入

For each residue, consider a window of size 13:

13x20=260 values

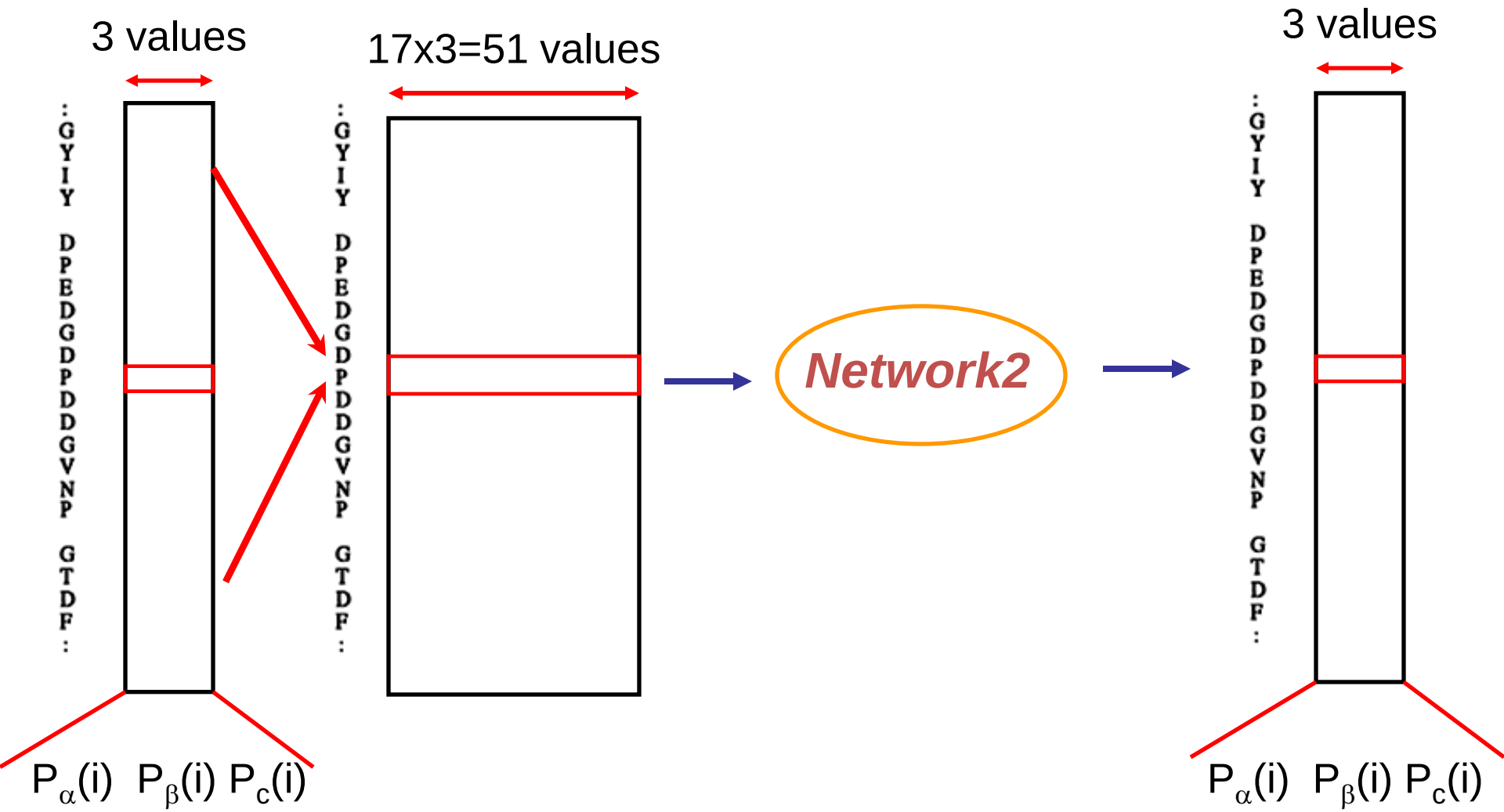


PHD: Network 1（序列→结构）



PHD: Network 2（结构→结构）

For each residue, consider a window of size 17:



二级预测服务器

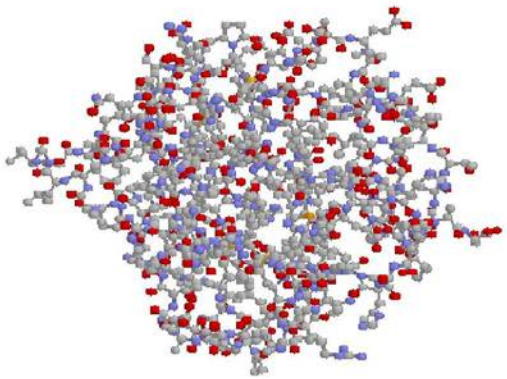
- Available servers:
  - Chou and Fasman: http://fasta.bioch.virginia.edu/fasta_www/chofas.htm
  - GOR: http://gor.bb.iastate.edu
  - PHD: http://cubic.bioc.columbia.edu/predictprotein/
  - JPRED : http://www.compbio.dundee.ac.uk/~www-jpred/
  - PSIPRED: http://bioinf.cs.ucl.ac.uk/psipred/
  - NNPREDICT: http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html

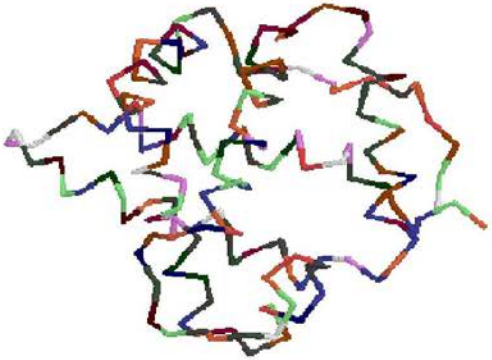
目录

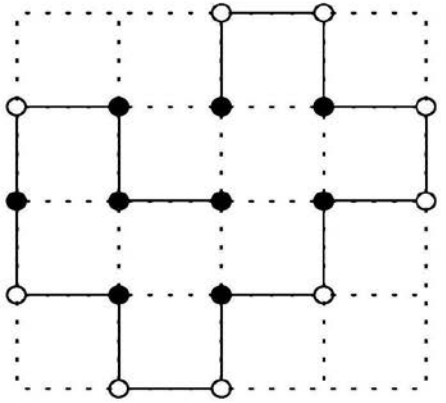
- 蛋白质的三维结构
- 蛋白质结构预测的重要性
- 蛋白质二级结构预测方法
- **蛋白质三级结构预测方法**
- 研究趋势

Lattice Models

- Modeling protein folding



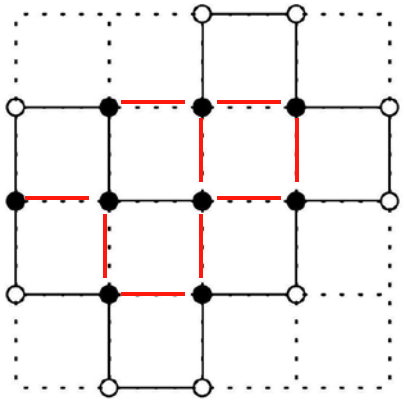




- Simple exact model + approximate algorithm

HP Model

- Twenty amino acids can be divided into two classes:
  - Hydrophobic/Non-polar (H) (疏水)
  - Hydrophilic/Polar (P) (亲水)
- The contacts between H points are favorable



- ● hydrophobic amino acid
- ○ hydrophilic amino acid
- — Covalent bond
- — H-H contact

- Goal: maximize the number of H-H contacts

HP Model

- Reduce computation by limiting degrees of freedom
- Limit α-carbon (Cα) atoms to positions on 2D or 3D lattice
- Protein sequence $\rightarrow$ represented as path through lattice points
- Emphasis on forming hydrophobic core

Complexity

- A combinatorial optimization problem
  - NP-hard problem
  - Long range interaction + global optimization
- GA MC SA ----- time consumed

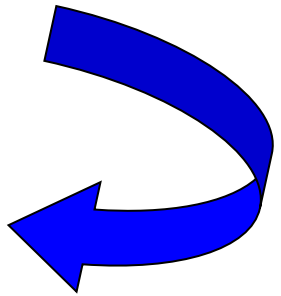
Energy level distribution

Energy level	No. of conformations
0	36.098.079
−1	31.656.934
−2	12.473.446
−3	2.943.974
−4	517.984
−5	77.080
−6	10.364
−7	1194
−8	96
−9	4
Total	83.779.155

How Bad are

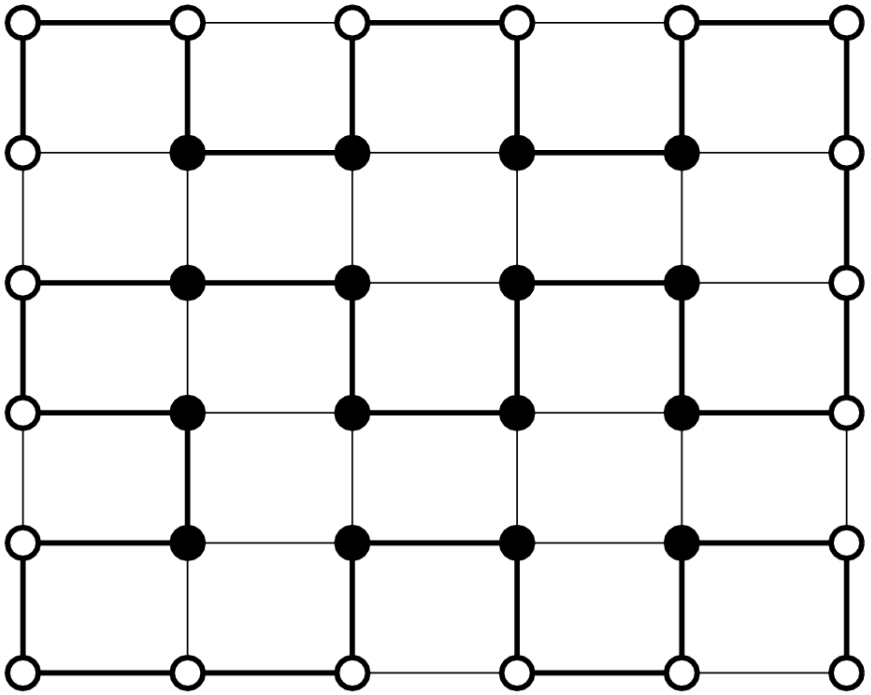
NP-Complete Problems?

Length=20

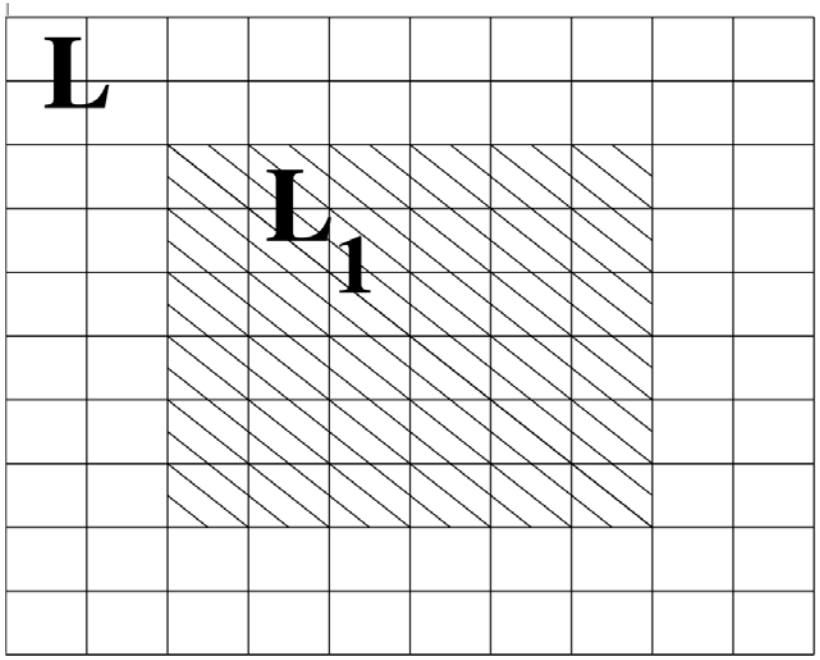


SOM Approach

- Existed SOM solution
  - Motivated by SOM for TSP
  - Incorporation of HP Information
  - Compact lattice



New SOM Approach



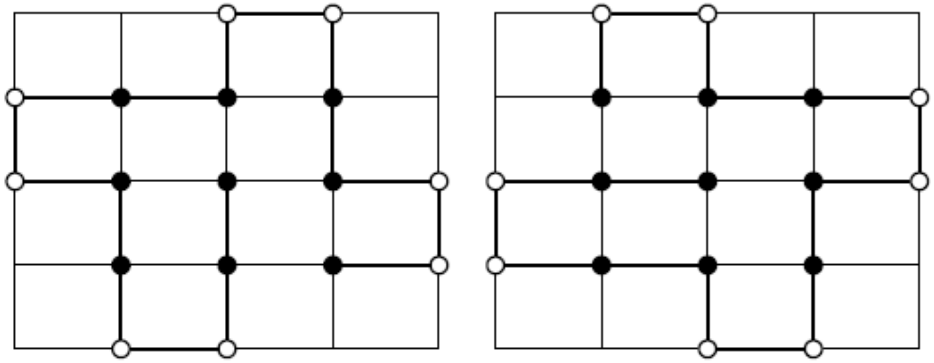
- Motivation
  - Consider a big lattice
    - Multiple map of SOM
    - Feasibility of solutions
  - Equivalent to PCTSP
  - Properly define the lattice distance
  - TSP force + H-H force

New SOM Approach

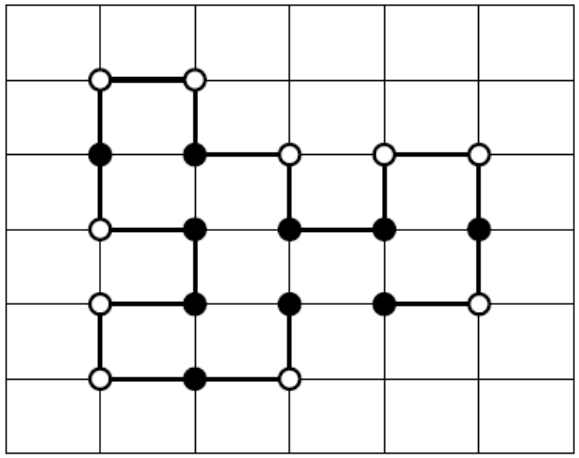
- Approachs
  - Initialization
  - Learning sample set partition strategy
  - Learning sample set reduction strategy
  - Local search procedure

Numerical Results

1. Constructed HP sequences



2. HP benchmark (up to 36 amino acids)

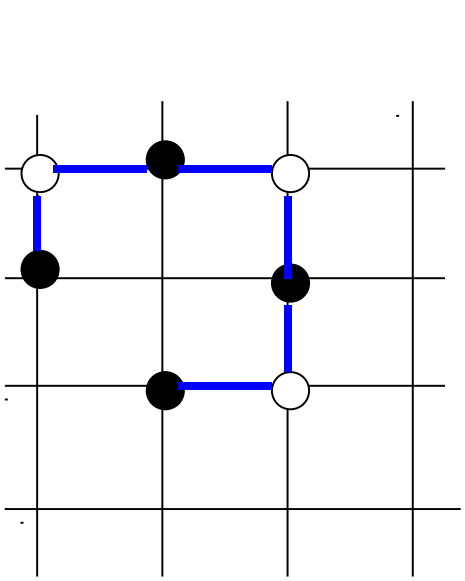


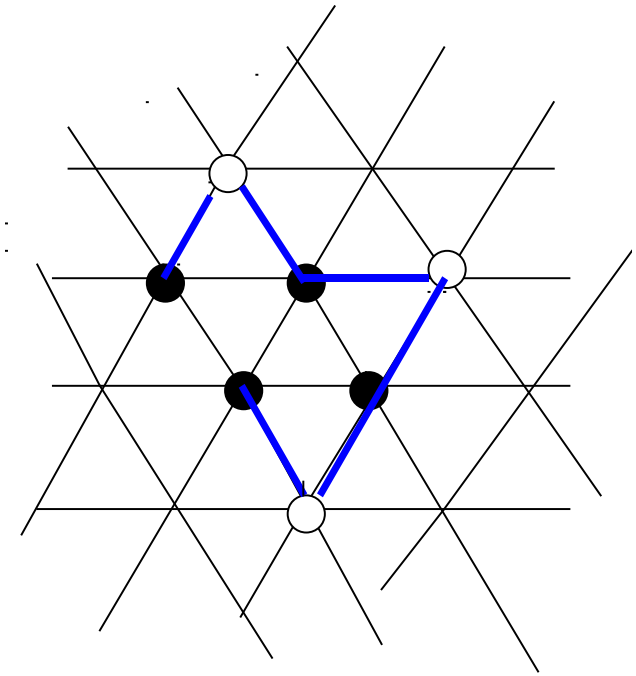
Unique Optimal Folding Problem

- What proteins in the two dimensional HP model have unique optimal (minimum energy) folding? (Brian Hayes, 1998)

- Oswin Aichholzer proved that in square lattice
  - There are closed chains of monomers with this property for all even lengths.
  - There are open monomer chains with this property for all lengths divisible by four.

Square Lattice and Triangular Lattice





Our Results

- For any n = 18k (k is a positive integer), there exists an n-node (open or closed) chain with at least 3^O(n) optimal foldings all with isomorphic contact graphs of size n/2.

- On 2D triangular lattice, for any integer n> 19, there exist both closed and open chains of n nodes with unique optimal folding.

Examples of Optimal Foldings

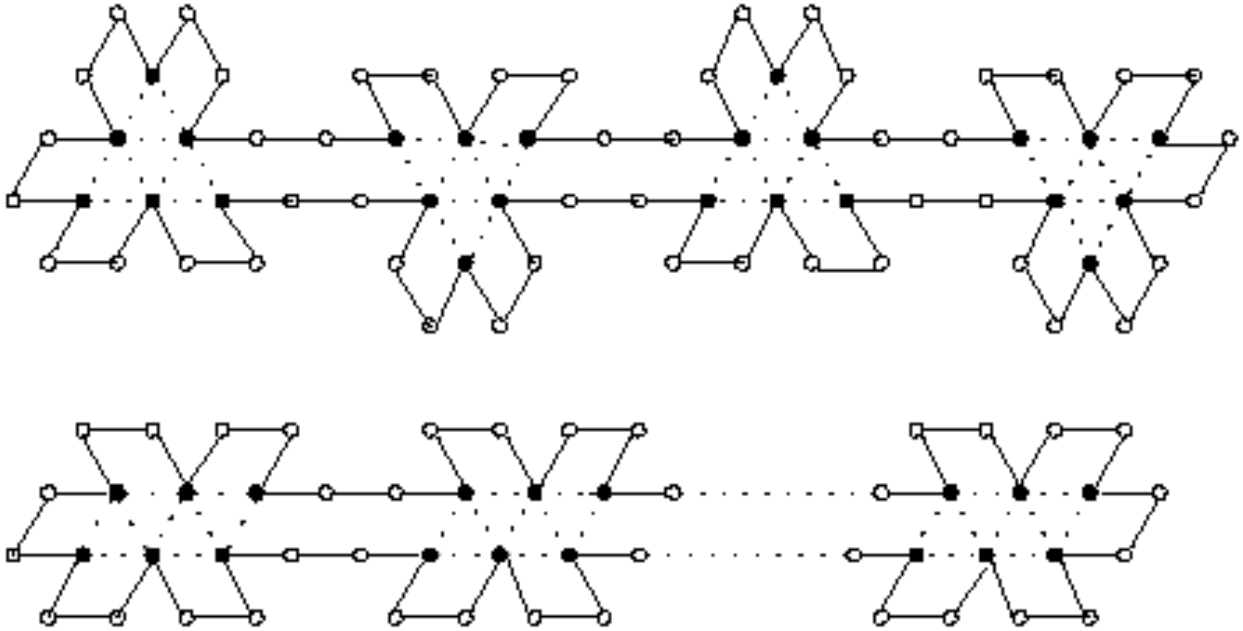


Figure 4: Examples of optimal foldings of $(PHP)^{6k}$

三级结构预测方法

- Comparative (homology) modeling (比较建模法)
  - 通过与已知结构蛋白质的序列比对来预测（序列-序列比对）
  - 适用于>50%的未知结构蛋白质
- Threading (fold recognition) (折叠识别法)
  - 从已知的结构类中找到最适合的（序列-结构比对）
  - 适用于30%左右的未知结构蛋白质
- Ab initio / de novo methods (从头预测法)
  - 不使用已知的结构信息，仅仅利用序列信息来预测
  - 适用于全新结构的蛋白质

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：

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

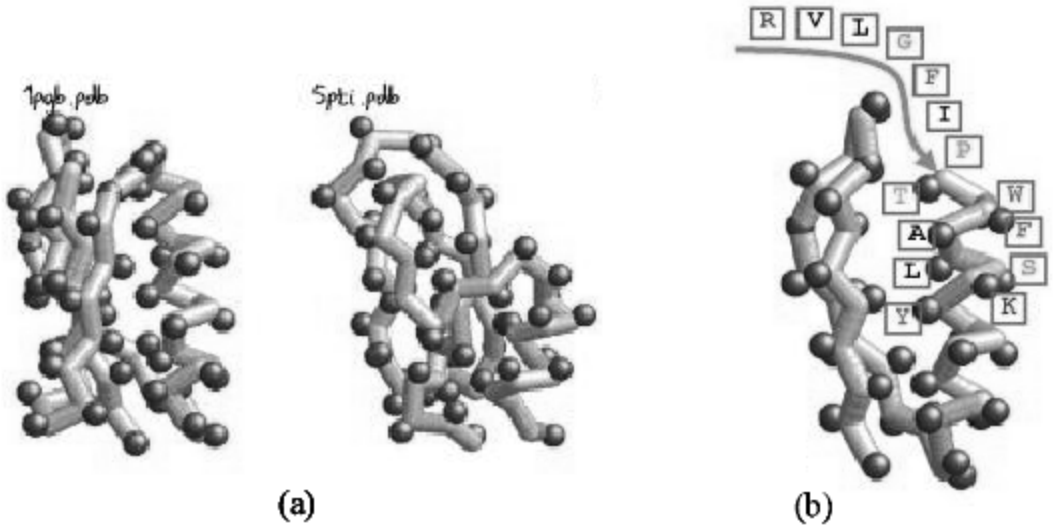
比较建模法

- 原理
  - 序列相似性>25%的蛋白质具有相似的结构
- 基本步骤
  - 将目标序列和模板序列进行比对（关键步骤）
  - 构造主干（backbone）的三维模型
    - 用二级结构预测和手工预测等方法来填补一些空隙
  - 构造环（loops）和卷曲（coils）
  - 安装侧链
    - 搜索侧链数据库，使用分子动力学
  - 优化 / 检验模型
- 关键方法
  - 序列-序列比对

折叠识别法

- 原理
  - 蛋白质具有很少的折叠类型（<1000）
- 基本步骤
  - 将目标蛋白质序列与已知的折叠进行比对
  - 将目标序列“安装”到选择的模板结构上
  - 对模型进行优化、调整
  - 检验模型的合理性
- 关键方法
  - 序列-结构比对

折叠识别法



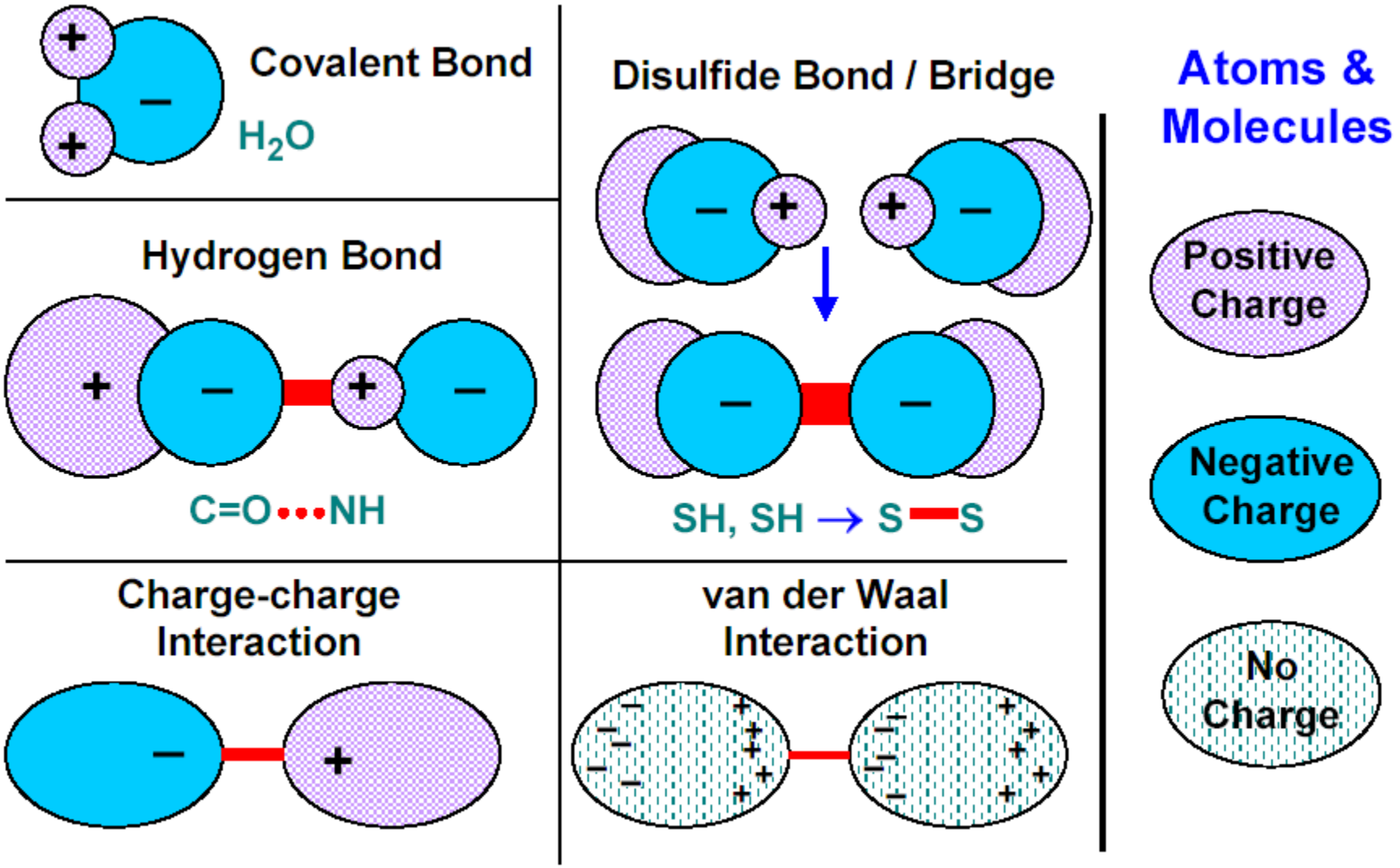
折叠识别法

- 动态规划
- 人工神经网络
- 分支定界法
- 线性规划
- Monte Carlo方法

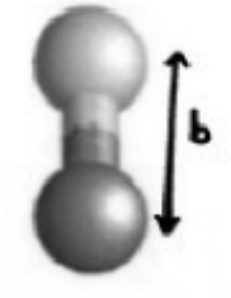
从头预测法

- 理论基础
  - 蛋白质的天然构象是热力学最稳定构象、也是能量最低构象
- 能量函数
  - 通过原子间作用力计算出的热力学能量
  - 精确，但难以计算
- 伪能量函数
  - 根据已知的三维结构知识得到的势能函数
    - 常见的结构 → 低能量
    - 不常见的结构 → 高能量
    - 极罕见的结构 → 极高能量

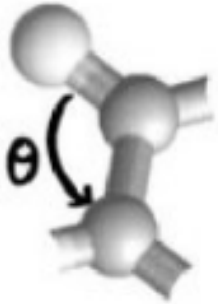
原子间作用力


Covalent Bond
H_2O
Disulfide Bond / Bridge
Hydrogen Bond
C=O•••NH
SH, SH → S—S
Charge-charge Interaction
van der Waal Interaction
Atoms & Molecules
Positive Charge
Negative Charge
No Charge

蛋白质的势能

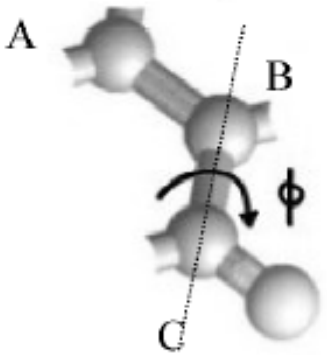


Pair Bonds $U(b) = \frac{1}{2} K_b (b - b_o)^2$



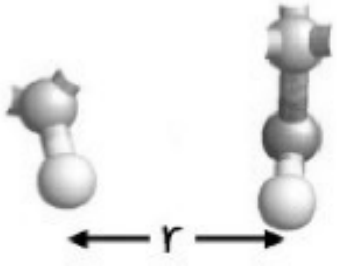
Bond Angles $U(\theta) = \frac{1}{2} K_\theta (\theta - \theta_0)^2$

蛋白质的势能



Torsion angle:Angle between A and D looking through the axis, B-C

$$U(\phi) = K_{\phi}(1 - \cos(n\phi + \delta))$$



Non-bonded pairs

$$U(r) = \varepsilon\left[\left(\frac{r_0}{r}\right)^{12} - 2\left(\frac{r_0}{r}\right)^{6}\right]$$

能量函数

$$E_{\text{AMBER}} = \sum_{\text{bonds}} K_{r_i}(r_i - r_{i,eq})^2 + \sum_{\text{angles}} K_{\theta_\ell}(\theta_\ell - \theta_{\ell,eq})^2 + \sum_{\text{dihedrals}} \frac{V_k}{2}$$

$$\times [1 + \cos(n_k \phi_k - \gamma_k)] + \sum_{i<j} \left(\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + C \frac{q_i q_j}{r_{ij}} \right)$$

$$E_{\text{SOLVATION}} = \sum_{i,j \le N_c} \sum_{k \le M} h_k \exp\left(- \left[\frac{(r_{ij} - c_k)}{w_k} \right]^2 \right)$$

基于能量最小化的从头预测法

- 基本步骤
  - 选择蛋白质能量函数模型
  - 选择三维结构的表示方法
  - 选择三维结构的评价函数
  - 选择寻找最优结构的优化方法
- 关键问题
  - 大规模非线性规划
  - 大量的局部极小点
  - 在计算时间和准确度之间寻找平衡

最优化方法

- 非线性规划
  - 最速下降法
  - 牛顿法
  - 共轭梯度法
- 全局优化方法
  - 遗传算法
  - 分解-结合法
  - 离散化方法

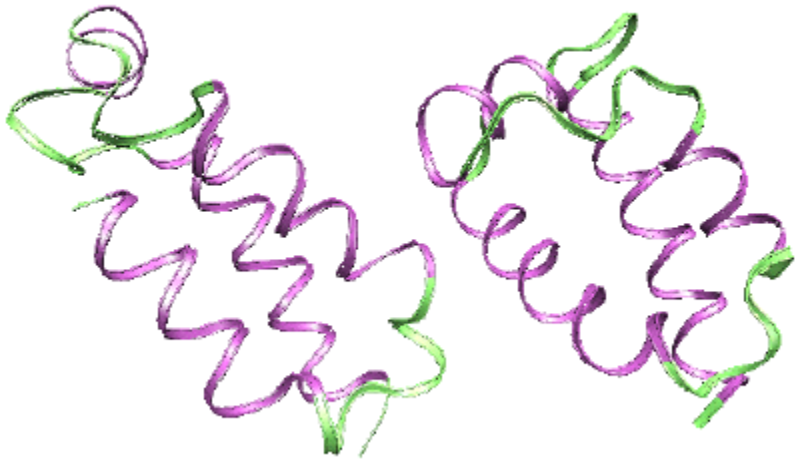
Monte Carlo方法

- Metropolis算法
- 基本步骤
  - 选择一个初始结构A
  - 随机产生一个新的结构B
  - 如果 $E_B < E_A$ 接受B
  - 否则产生一个[0,1]之间的随机数r
  - 如果 $r < e^{-(E_B - E_A)/KT}$，接受B，否则拒绝
  - 重复以上步骤
- 温度T的修改——模拟退火方法

分子动力学（动力系统）

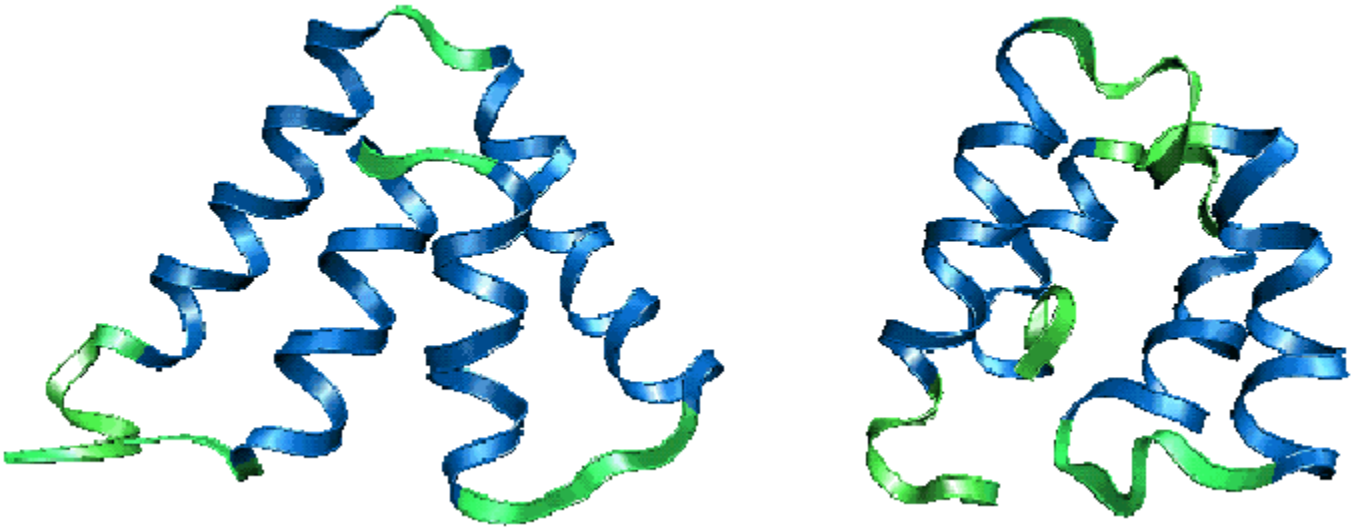
- 方法
  - 对蛋白质中原子间的作用力进行建模
  - 用动力学方程跟踪蛋白质折叠时每个原子的位置
  - 求解这些方程的解析解是极其困难的
  - 用离散动力系统的方法来确定数值解
- 问题
  - 模拟蛋白质折叠过程是非常消耗时间的
  - 模拟 10^{-9} 秒的蛋白质折叠过程大约需要一天时间
  - 蛋白质的实际折叠时间大约在 10^{-4} 秒的数量级或者更多
  - 需要超级计算机

预测结果



E. Eskow, B. Bader, R. Byrd, *et al.* An optimization approach to the problem of protein structure prediction. ***Mathematical Programming***, Vol. 101, No. 3, pp. 497-514, 2004.

预测结果



E. Eskow, B. Bader, R. Byrd, *et al.* An optimization approach to the problem of protein structure prediction. ***Mathematical Programming***, Vol. 101, No. 3, pp. 497-514, 2004.

目录

- 蛋白质的三维结构
- 蛋白质结构预测的重要性
- 蛋白质二级结构预测方法
- 蛋白质三级结构预测方法
- **研究趋势**

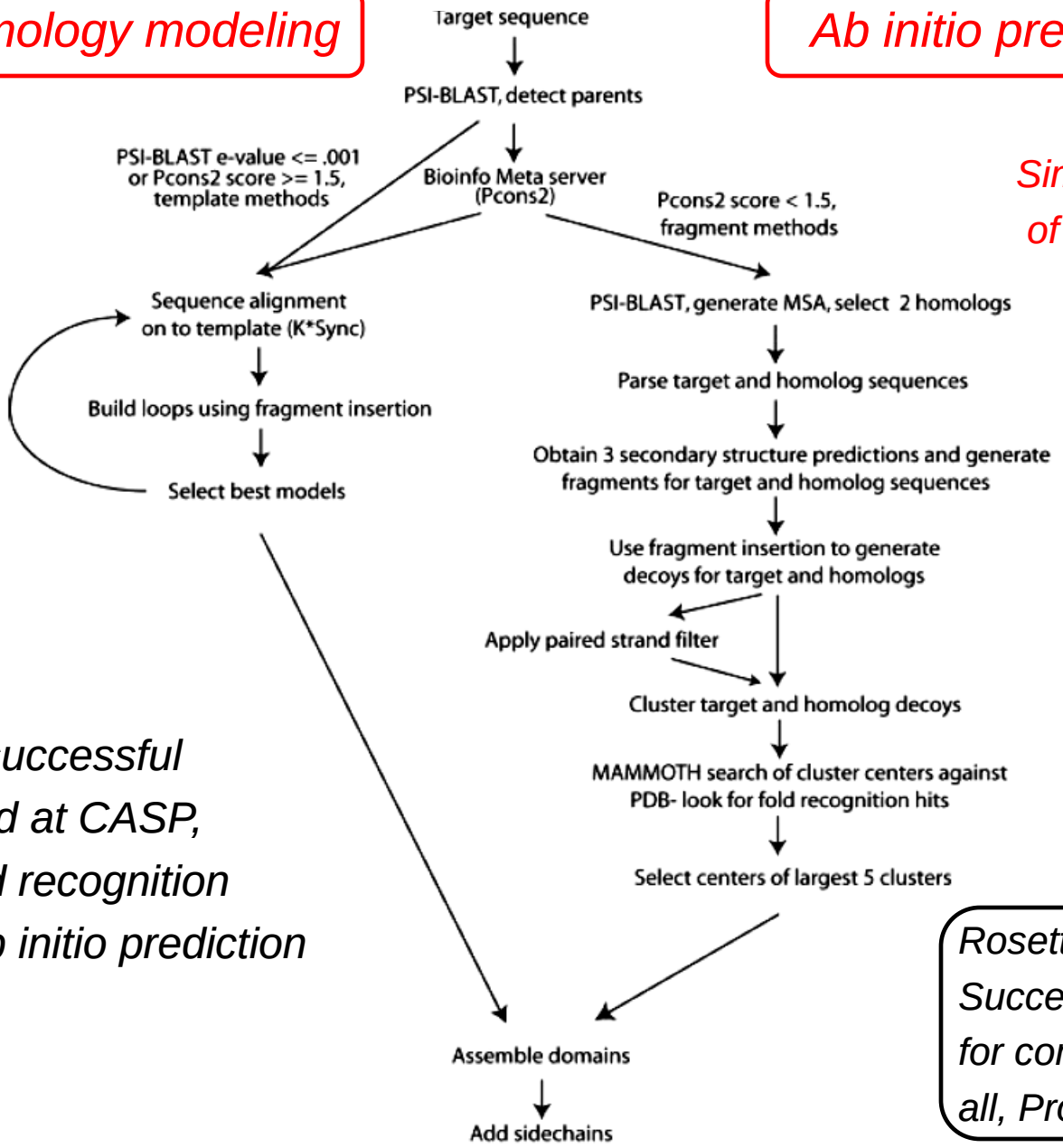
新的趋势

- 混合预测方法
  - 在比较建模法和折叠识别法中使用从头预测法来预测部分难以找到模板的片断
  - 在从头预测法中使用二级结构预测的结果和其他已知结构信息辅助建模
- Meta-predictor
  - 使用多个预测方法
  - 对收集的结果进行综合比较和分析
  - 改进收集的结果

ROSETTA at CASP (David Baker)

Homology modeling

Ab initio prediction



Simultaneous modeling of the target and 2 homologs

Secondary structure prediction

Fragment based approach to generate decoys

Select 5 decoys For prediction

Most successful Method at CASP, for fold recognition and ab initio prediction

Rosetta predictions in CASP5: Successes, failures, and prospect for complete automation. Baker et all, Proteins, 53:457-468 (2003)

参考文献

1. K. Ginalski, N. V. Grishin, A. Godzik and L. Rychlewski. Practical lessons from protein structure prediction. ***Nucleic Acids Research***, Vol. 33, No. 6, pp. 1874-1891, 2005.
2. Jinbo Xu, Ming Li, Dongsup Kim and Ying Xu. Raptor: optimal protein threading by linear programming. ***J. of Bioinformatics and Computational Biology***, Vol. 1, No. 1, pp. 95-117, 2003.
3. E. Eskow, B. Bader, R. Byrd, *et al.* An optimization approach to the problem of protein structure prediction. ***Mathematical Programming***, Vol. 101, No. 3, pp. 497-514, 2004.