生物信息学

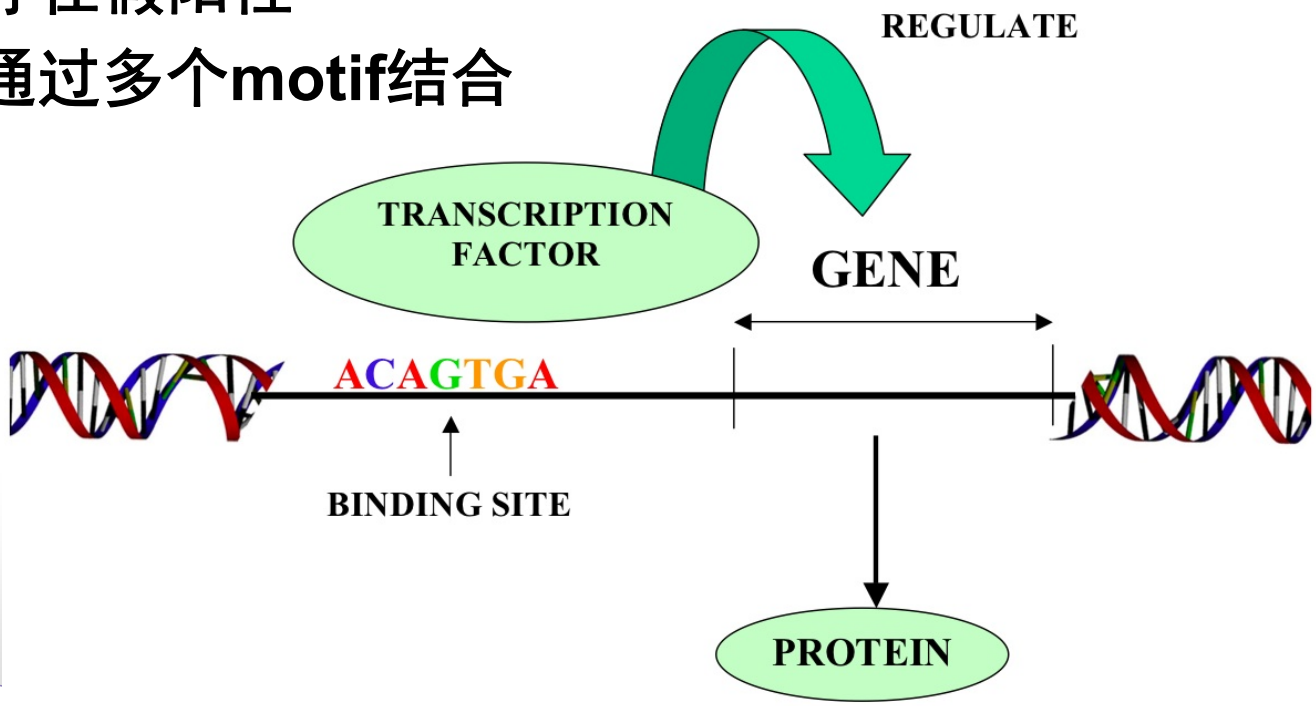
第八章 序列模式识别（2）

有待解决的问题

- 给定的一组序列，可能的motif仅在部分序列中出现，怎么解决？
- 给定一组序列，其中存在某种motif可能在序列上出现两次以上，如何解决？

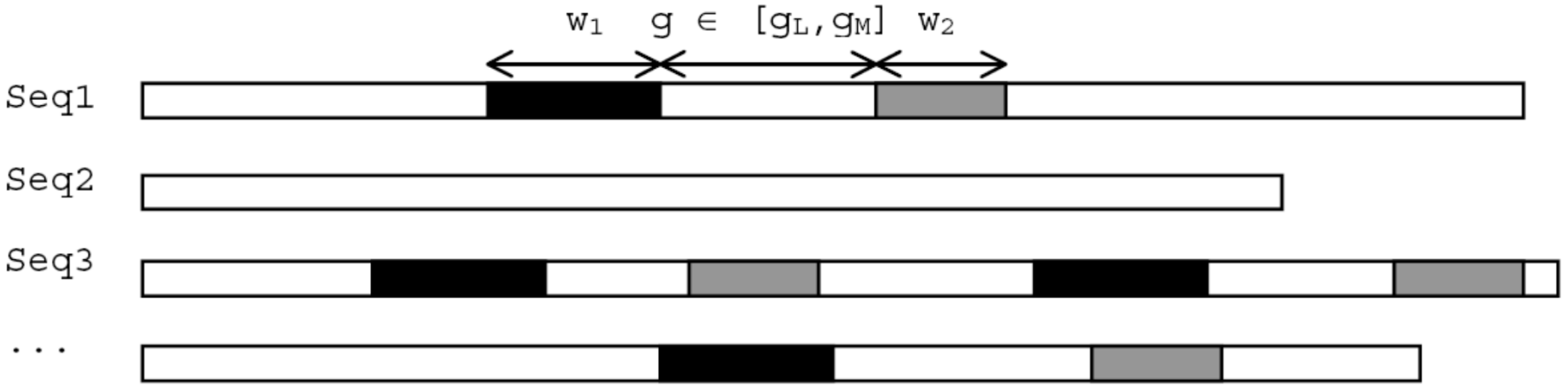
TFBS finding

- TFBS: transcription factor binding site
  - 转录因子识别特定的motif
- CHIP-chip/CHIP-seq: 高通量鉴定与特定转录因子结合的DNA片段
  - 存在假阳性
  - 通过多个motif结合



TFBS finding

- Protein-DNA binding sites finding
- 基本模型
  - 给定n条序列
  - 假设存在若干模体，w_1, w_2,…，其中w_1, w_2之间的间隔为g_L -> g_M
  - w1, w2不一定在每条序列上存在
  - w1, w2可以在正链或反链上存在



如何发现？

- **MDscan: a two-step algorithm**
- **Word enumeration (Seed):**
  - 候选基因排序：结合强度或read数
  - Top sequence (3~20)
  - 统计长度为*w*的word个数 (*w*-mer)
  - 考虑正链和反链
- **“*m*-matches”:**
  - 保证随机产生两个*w*-mer存在*m*个匹配概率<0.15%
  - 可随机生成例如1M对w-mers来计算

TECHNICAL REPORT

An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments

X. Shirley Liu[1], Douglas L. Brutlag[2], and Jun S. Liu[3]*

Published online: 8 July 2002, doi:10.1038/nbt717

如何发现？

- **MDscan: a two-step algorithm**
- **Word enumeration (Seed):**
  - 候选基因排序：结合强度或read数
  - Top sequence (3~20)
  - 统计长度为*w*的word个数 (*w*-mer)
  - 考虑正链和反链
- **“*m*-matches”:**
  - 保证随机产生两个*w*-mer存在*m*个匹配概率<0.15%
  - 可随机生成例如1M对w-mers来计算

TECHNICAL REPORT

An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments

X. Shirley Liu[1], Douglas L. Brutlag[2], and Jun S. Liu[3]*

Published online: 8 July 2002, doi:10.1038/nbt717

m-matches

- 随机生成1,000,000对*w*-mers，不超过1500对存在*m*-matches

																					Un	m
																	15610	156444	578033	**3**	0	3
																3866	50431	261463	683892	**4**	0	4
															999	15788	103945	367525	762862	**5**	1	4
														260	4641	37671	169418	466220	821860	**6**	1	5
													65	1306	12713	70506	243636	555864	867109	**7**	2	5
												12	392	4272	27538	113741	321231	633297	900009	**8**	2	6
											2	105	1407	10118	49012	165815	398689	699797	924887	**9**	3	6
										0	38	420	3620	19766	78290	223229	473498	754994	943284	**10**	3	7
									0	10	133	1234	7510	34257	114473	286401	545232	803518	957935	**11**	4	7
								0	1	37	364	2801	14301	54690	157580	350674	609287	841567	968443	**12**	4	8
							0	1	10	147	966	5626	24230	80470	206308	416197	667823	873497	976093	**13**	5	8
						0	0	6	34	349	2214	10275	38462	111590	258561	479032	719237	899054	982199	**14**	5	9
					0	0	1	15	118	836	4339	17280	56906	148748	313486	538585	763566	919704	986632	**15**	6	9
				0	0	0	4	39	277	1648	7443	27180	79561	189372	369310	593995	802703	936437	990112	**16**	6	10
			0	0	0	1	12	98	627	3007	12156	40168	106776	234404	425856	646959	836795	950011	992629	**17**	7	10
		0	0	0	2	6	35	226	1235	5357	19075	56952	139314	282466	480667	693812	864651	960528	994321	**18**	7	11
	0	0	0	0	0	10	85	457	2350	9107	29009	77848	175328	332612	534945	737045	888473	969147	995810	**19**	8	11
0	0	0	0	1	5	27	177	916	3918	13888	40857	102519	214375	382762	585011	774950	908685	975667	996860	**20**	9	11
20	**19**	**18**	**17**	**16**	**15**	**14**	**13**	**12**	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**w-mers**		

模体的信息量

- PSSM: Seed + "*m*-matches"
- 估算信息量 (Maximum a *posteriori*, MAP)：

$$\frac{x_m}{w} \times \left[\sum_{i=1}^{w} \sum_{j=\mathrm{A}}^{T} p_{ij} \log p_{ij} - \frac{1}{x_m} \sum_{\text{all segments}} \log(p_0(s)) - \log(\text{expected bases/site}) \right]$$

- X_m: 模体/PSSM里存在*m*-matches的个数；
- p_{ij}:模体/PSSM里第*i*位的第*j*个核酸的出现频率
- $p_0(s)$: 从随机数据里产生一个模体*s*的概率
- log(expected bases/site): top sequences里"期望"获得的*m*-matches个数 (未知)

模体的信息量

$$\frac{\log(x_m)}{w}\left[\sum_{i=1}^{w}\sum_{j=\mathrm{A}}^{T} p_{ij}\log p_{ij} - \frac{1}{x_m}\sum_{\text{all segments}}\log(p_0(s))\right]$$

- **Markov background model:**
- **A. 随机数据**
- **B. 针对每一个w-mer，计算其背景概率**

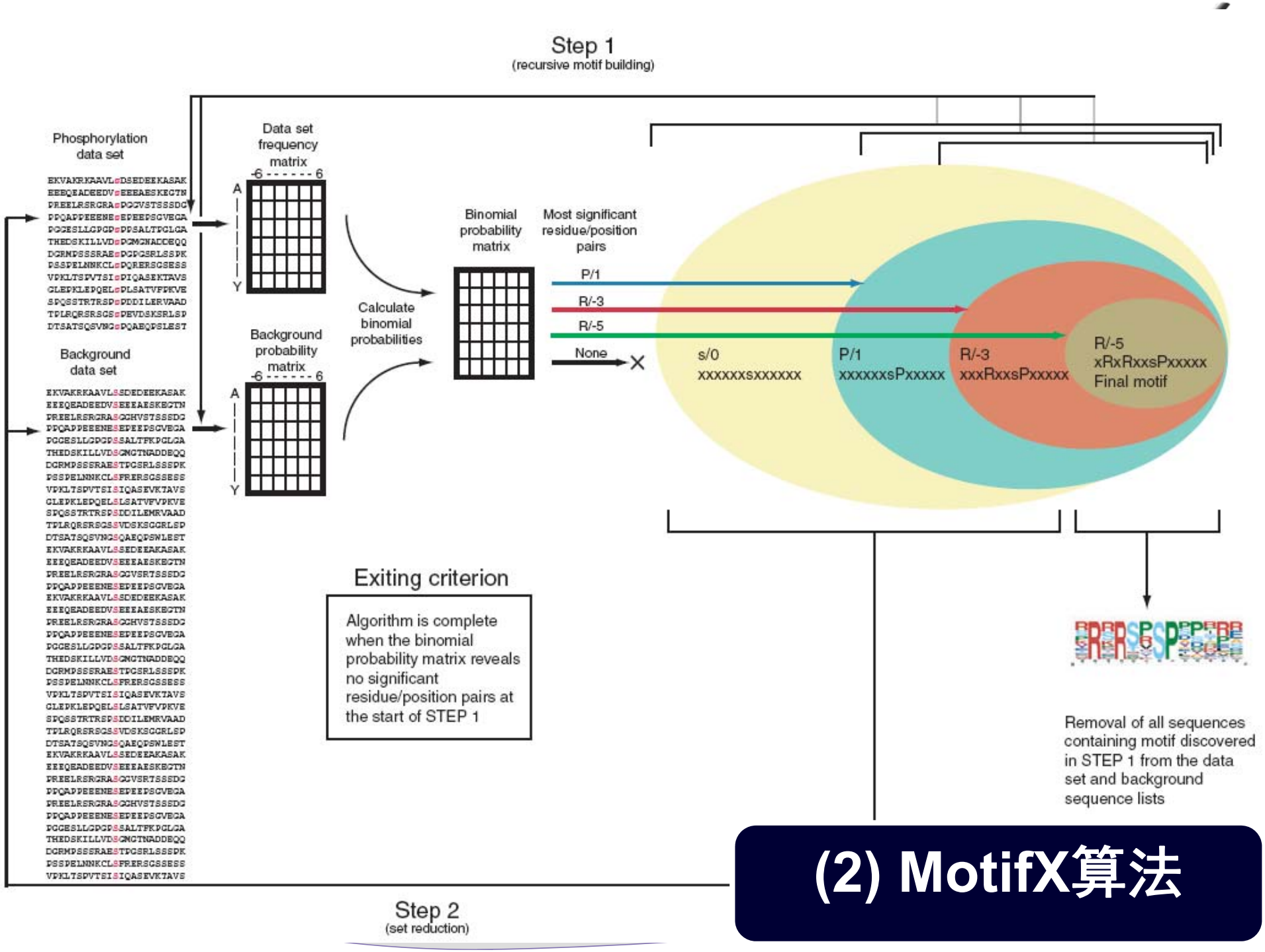
$$\begin{aligned} p_0(ATGTA) = {} & p(A|\text{previous 3 bases } CTT) \times p(T|\text{previous 3 bases } TTA) \times \\ & p(G|\text{previous 3 bases } TAT) \times p(T|\text{previous 3 bases } ATG) \times \\ & p(A|\text{previous 3 bases } TGT) \end{aligned}$$

模体/PSSM的优化

- 仅保留10-50个MAP最高的模体
- 利用生成的模体/PSSM对剩余的序列进行打分
- 将所有结果按照得分排序
- 依次将结果加入模体/PSSM
- 重新计算MAP，若分值提高则保留，反之放弃

问题

- 过表达某转录因子，其中500个基因上调，200个下调，如何分析？
- CHIP-array实验中，发现与转录因子X结合的基因有500个，如何分析？


Step 1
(recursive motif building)
Phosphorylation data set
Data set frequency matrix
-6 ------ 6
A
Y
Background data set
Background probability matrix
Calculate binomial probabilities
Binomial probability matrix
Most significant residue/position pairs
P/1
R/-3
R/-5
None
s/0
xxxxxxsxxxxxx
P/1
xxxxxxsPxxxxx
R/-3
xxxRxxsPxxxxx
R/-5
xRxRxxsPxxxxx
Final motif
Exiting criterion
Algorithm is complete when the binomial probability matrix reveals no significant residue/position pairs at the start of STEP 1
Removal of all sequences containing motif discovered in STEP 1 from the data set and background sequence lists
Step 2
(set reduction)

(2) MotifX算法

马尔科夫及隐马尔科夫模型

- 20世纪初，俄国数学家Andrey Markov提出马尔科夫链

Andrey Markov
1856-1922

- 马尔科夫模型
- 马尔科夫链
- 隐马尔科夫模型

马尔科夫模型

- 马尔科夫模型：随机过程的一种，主要特点为“无后效性”，即根据当前的状态即可完全确定将来的状态

马尔科夫性 & 马尔科夫链

- 定义：对于离散的随机过程$X_1, X_2, X_3 \ldots$，其马尔科夫性定义为：

$$P(X_{n+1} = j \mid X_1 = x_1, X_2 = x_2, \ldots X_n = x_n) = P(X_{n+1} = j \mid X_n = x_n)$$

(for all x_i, all j, all n)

- 具有马尔科夫性的过程称为马尔科夫过程
- 时间(先后顺序)和状态都离散的马尔科夫过程称为马尔科夫链

马尔科夫模型：参数估计

$$P_{-2}(A|C) = \frac{N_{CA}^{(-3,-2)}}{N_{C}^{(-3)}}$$

$$S = S_1\, S_2\, S_3\, S_4\, S_5\, S_6\, S_7\, S_8\, S_9$$

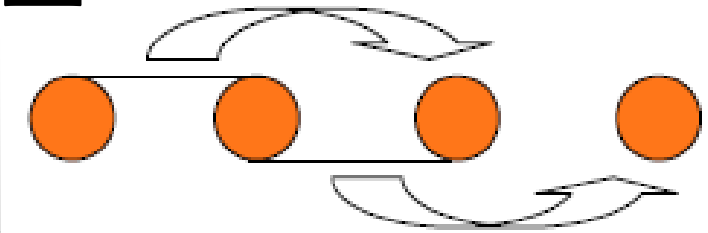
$$R = \frac{P(S|+)}{P(S|-)} = \frac{P_{-3}(S_1)P_{-2}(S_2|S_1)P_{-1}(S_3|S_2)\cdots P_{6}(S_9|S_8)}{P_{bg}(S_1)P_{bg}(S_2|S_1)P_{bg}(S_3|S_2)\cdots P_{bg}(S_9|S_8)}$$

K-order 马尔科夫模型

- 一阶马尔科夫模型：当前位置仅依赖前一位

$$P_{-2}(A|C) = \frac{N_{CA}^{(-3,-2)}}{N_{C}^{(-3)}}$$

- k阶马尔科夫模型：当前位置依赖前一位，而前一位依赖前两位,…,前k-1位依赖前k位
- 0阶马尔科夫模型：位点独立

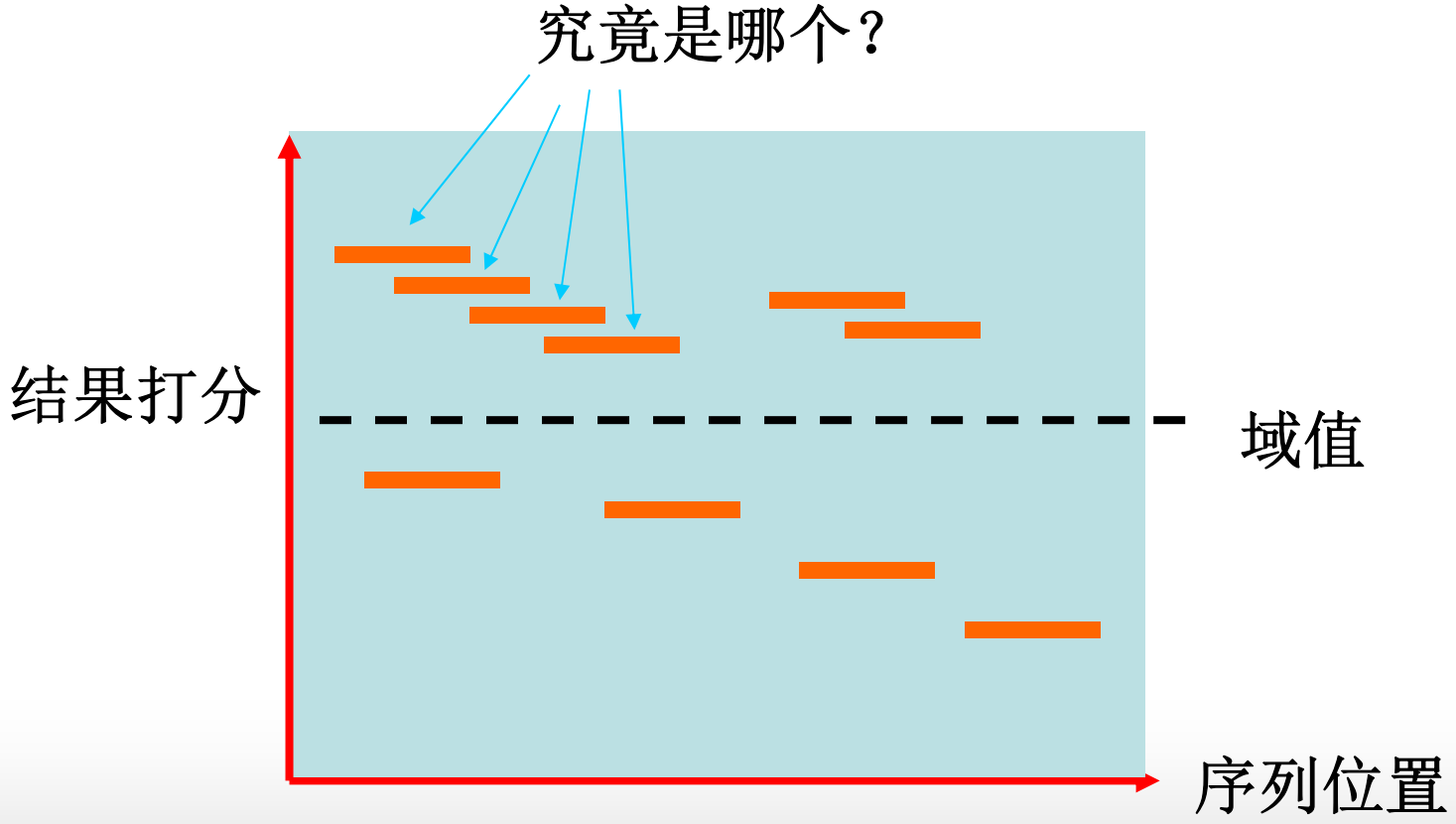


Markov & PSSM

- 对真实的数据进行训练，PSSM =~ 0阶马尔科夫模型
- 对新序列的扫描：从头至尾，每次移动1~n位（窗口滑动的方法）
- 分别计算窗口内的序列，是(+)和(-)的概率，计算log-odds ratio
- 设定域值，若高于域值，则保留结果

然而

- 对给定的一段序列进行扫描并预测，结果可能如下：



另外

- CpG序列的预测
- 蛋白质跨膜结构域的预测
- Gene及基因结构预测

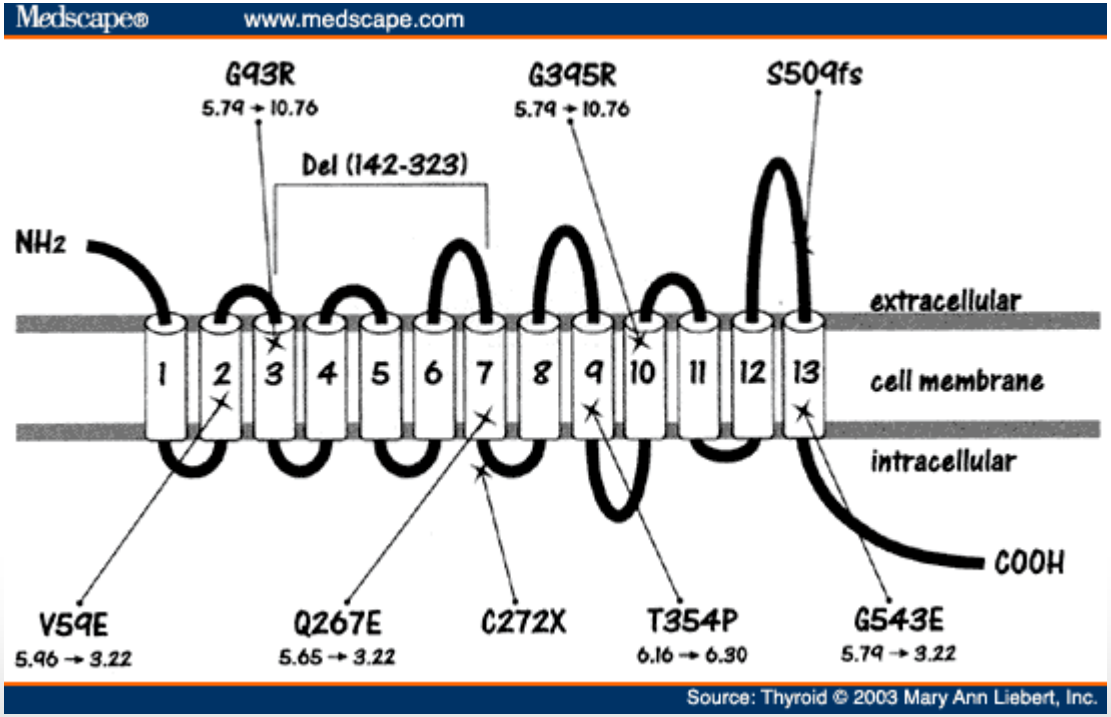
- 长度不确定！
- 起始位置不知！
- Markov models & PSSM: Not work!!!

CpG岛的预测

- CpG岛：在人的基因组中，如果双碱基对CG出现，则C通常被甲基化。并且，甲基化的C很快会突变成T。因此基因组中CpG岛非常少。然而，在基因的起始位置，例如promotor区域，因为功能的保守性，其序列很少突变，CpG的含量能够保持在40~60%
- How to predict? PSSM & Markov are not work at all!

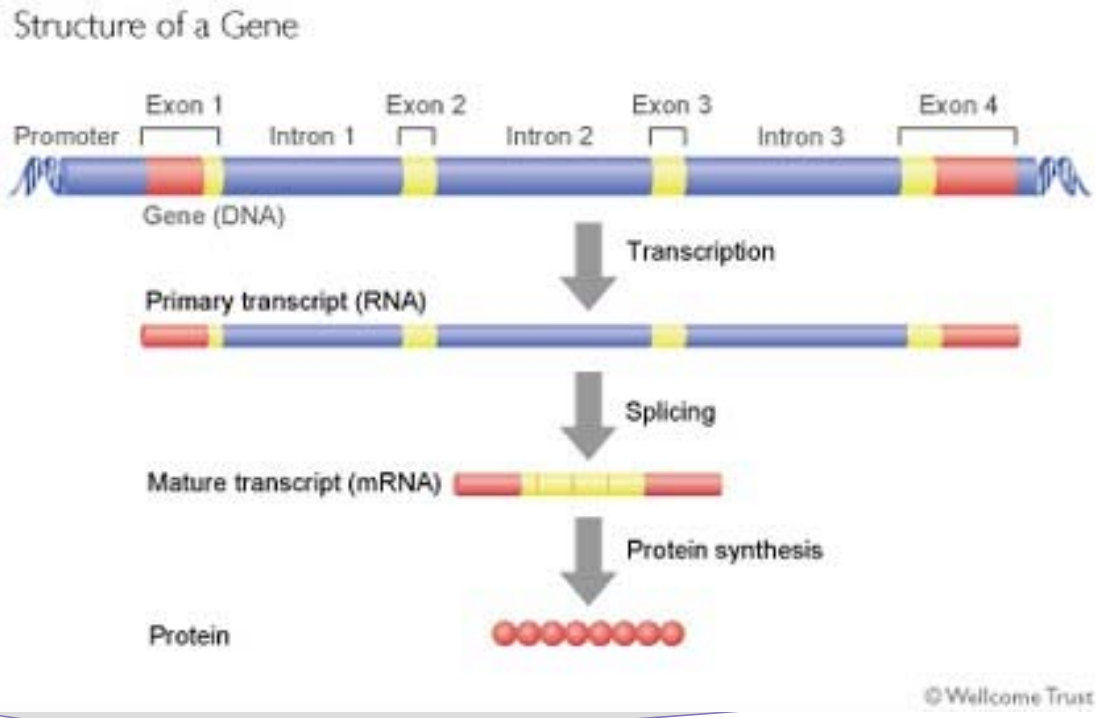
蛋白质跨膜结构域

- 细胞内，有些蛋白质部分在细胞膜外，部分在细胞膜内，负责细胞外信号向细胞内的转导等其他功能
- 跨膜结构域：长度不定，具有强的疏水性



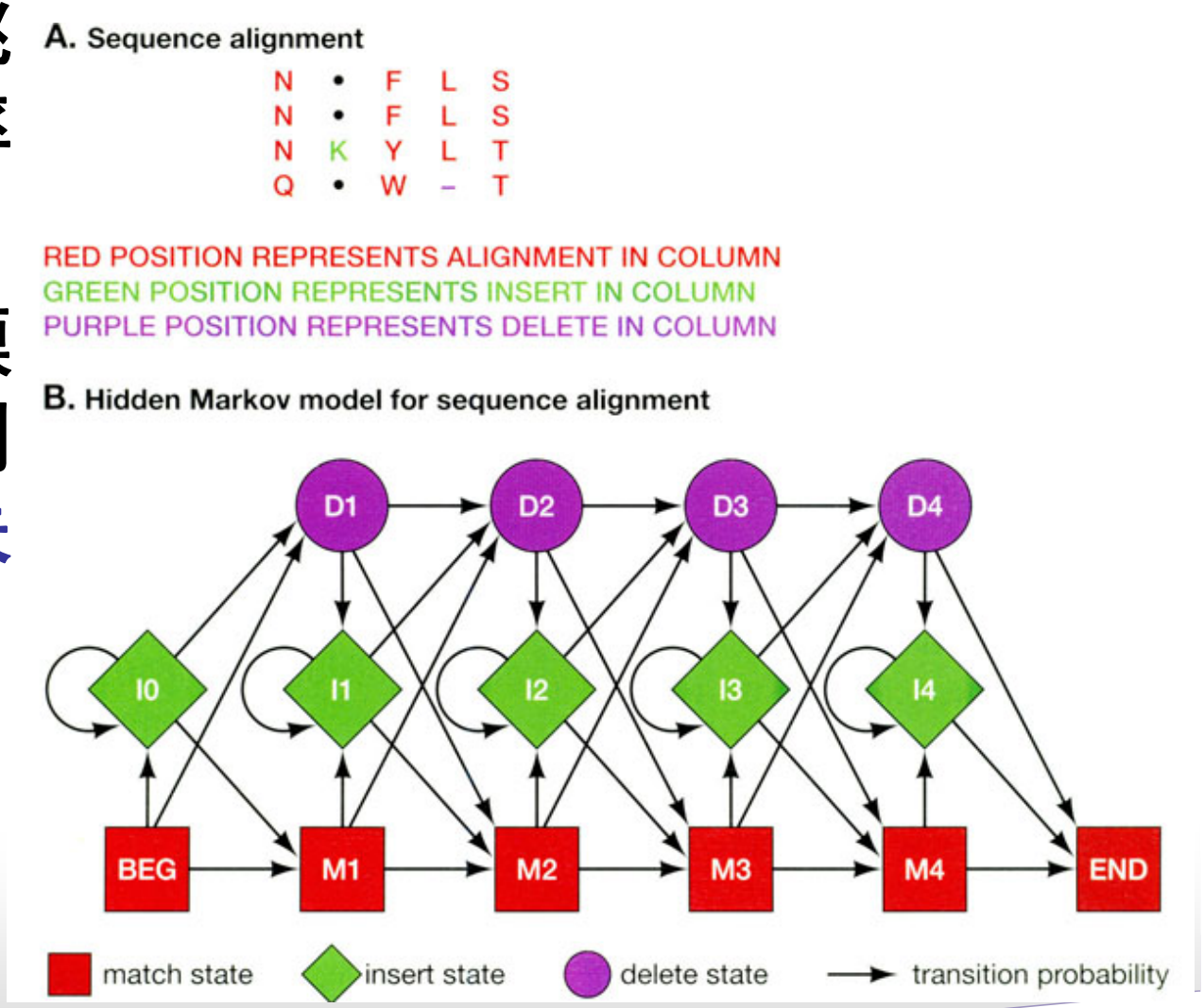
基因预测

- 基因预测：给定一段序列，能否预测是否包含基因
- 基因结构预测：真核生物的基因，包括启动子，外显子，内含子，剪切子，ESE，沉默子……能够正确预测基因的结构？



隐马尔科夫模型

- 存在状态的跳转：转移概率->未知！
- 隐马尔科夫模型：状态之间的转移概率未知

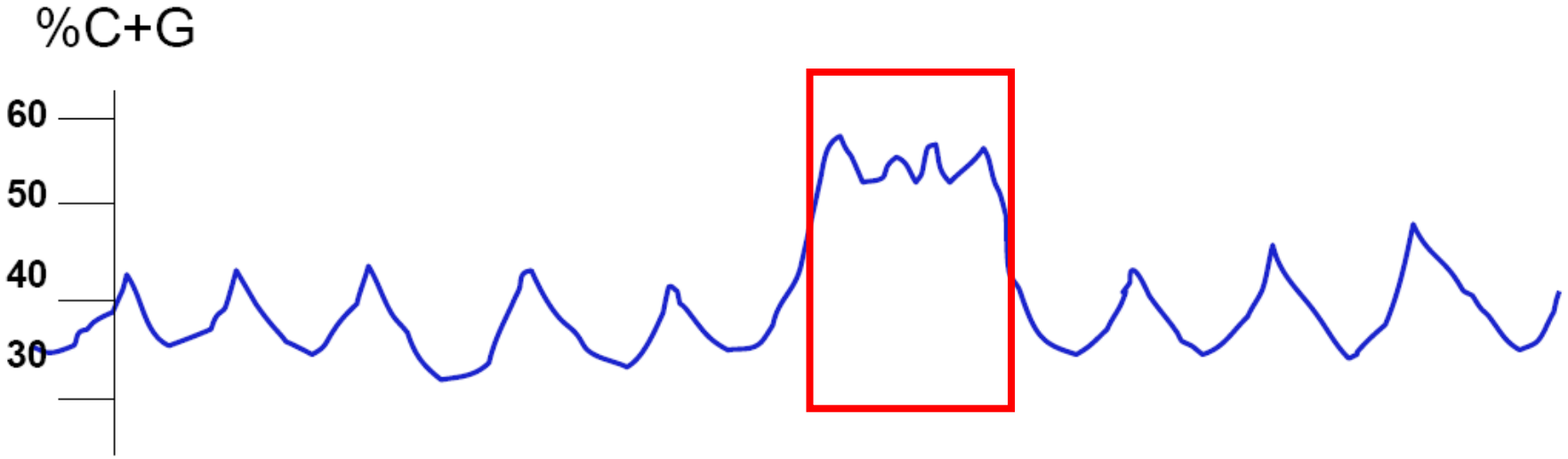


Profile HMM

- 多序列比对的结果中，氨基酸之间存在三种状态：
  - 匹配（M），插入 I）和缺失（D）
- HMM：三种状态之间的转换关系未知 -> hidden -> 转移概率
- 每个位置上的氨基酸/碱基以及插入、缺失的频率/概率可以通过观测求得 -> not hidden
- 模型训练：通过训练，估算转移概率

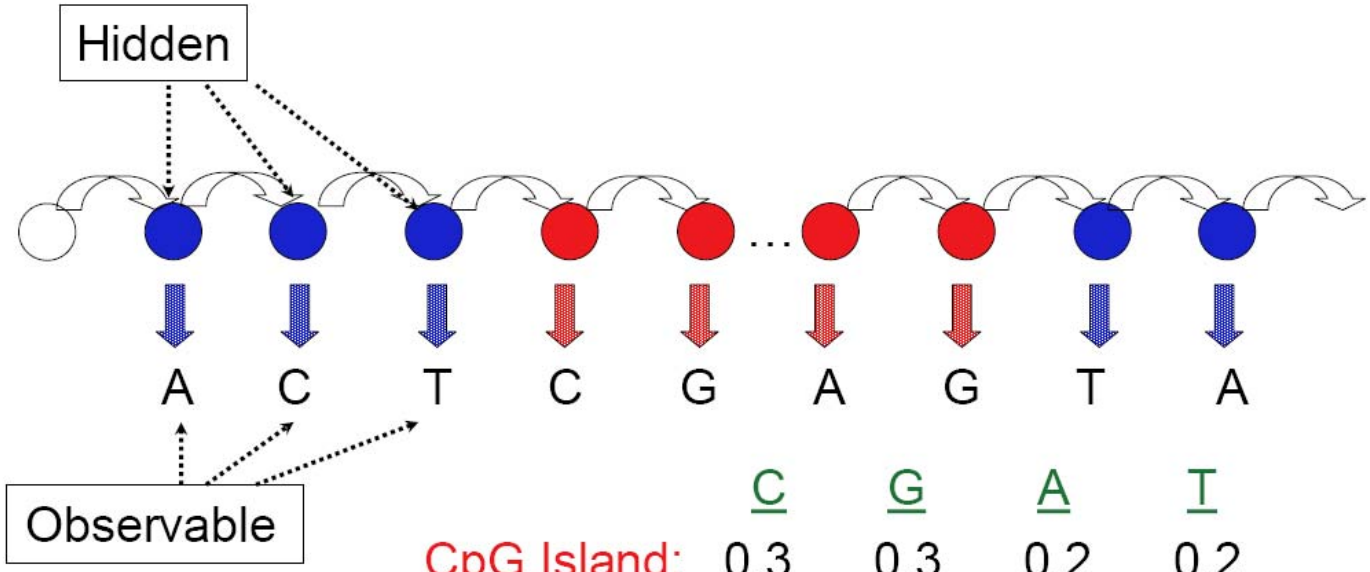
例：CpG岛的HMM

- 存在两种状态：是CpG岛 (CpG Island, I)，不是CpG岛 (Genome, G)



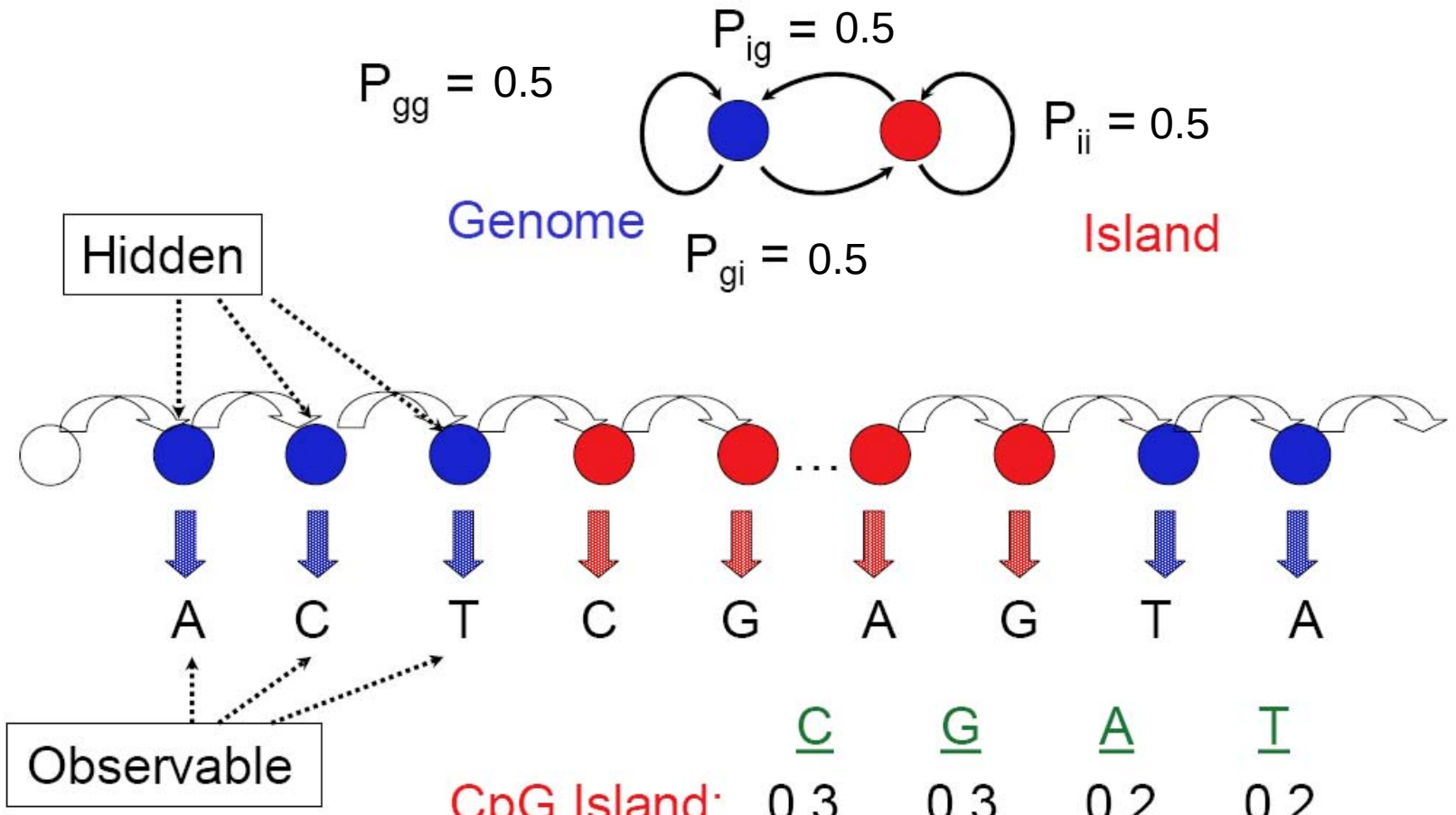
CpG岛：HMM

- **Hidden: 对当前未知的碱基，跳转到下一个位置，究竟是I还是G的概率，未知**
- **Observable: I和G中的四种碱基分布的概率能够通过实际数据的观测进行计算**



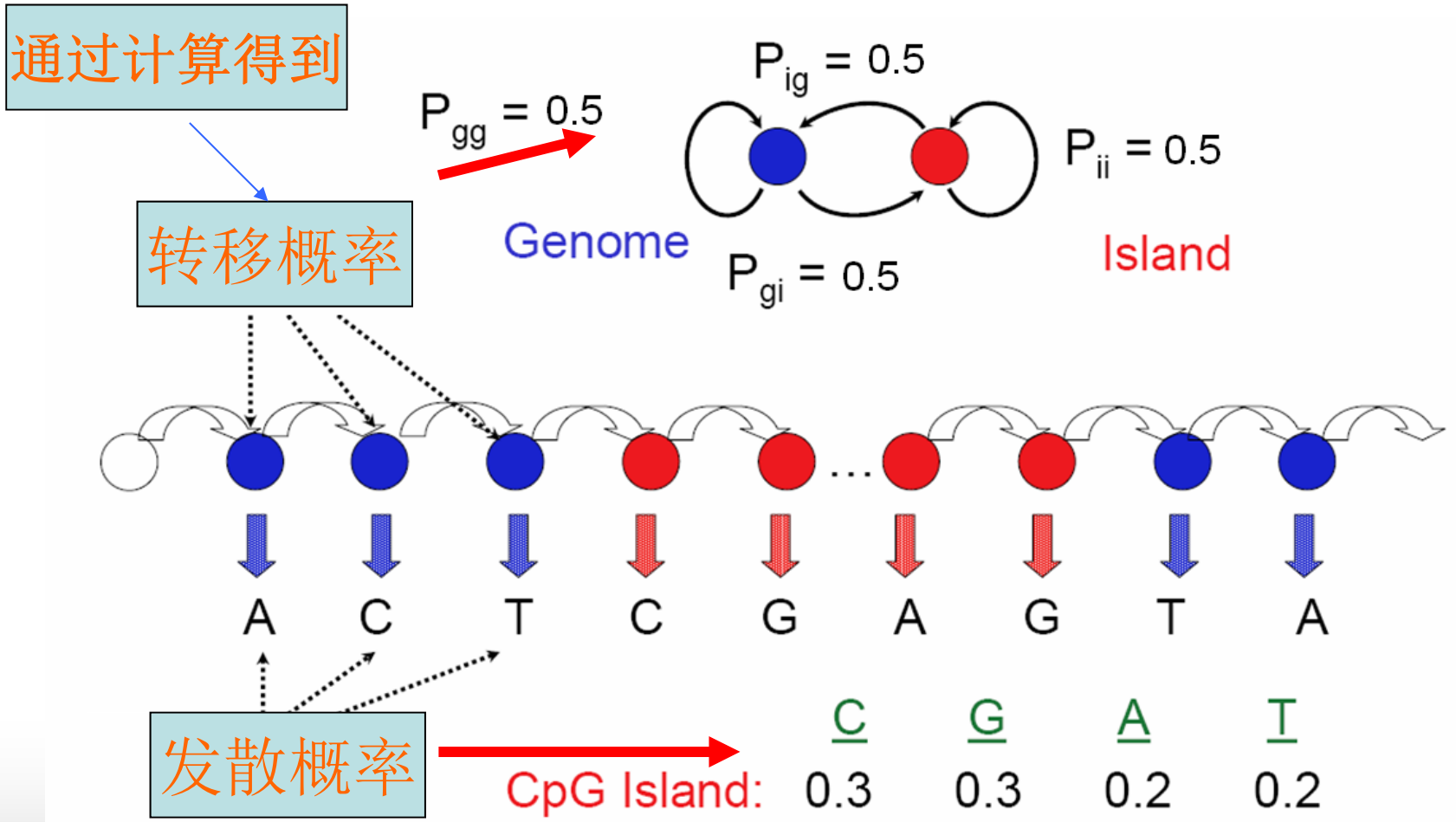
	C	G	A	T
CpG Island:	0.3	0.3	0.2	0.2
Genome:	0.2	0.2	0.3	0.3

CpG岛：HMM (2)



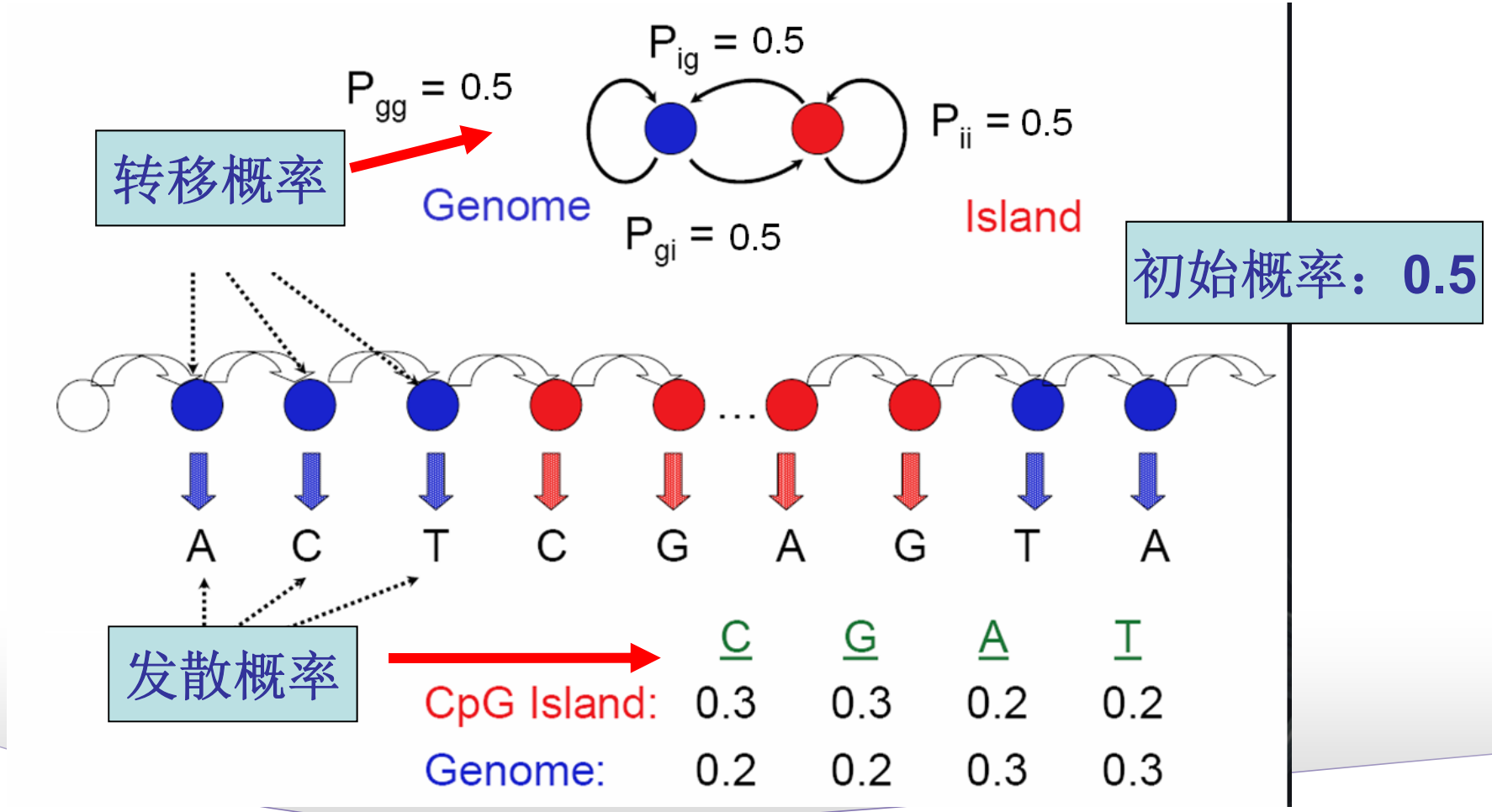
	C	G	A	T
CpG Island:	0.3	0.3	0.2	0.2
Genome:	0.2	0.2	0.3	0.3

CpG岛：HMM (3)


通过计算得到
转移概率
发散概率
P_{gg} = 0.5
P_{ig} = 0.5
P_{ii} = 0.5
P_{gi} = 0.5
Genome
Island
A C T C G A G T A
C G A T
CpG Island: 0.3 0.3 0.2 0.2
Genome: 0.2 0.2 0.3 0.3

预测CpG Island: Viterbi算法

- 给定序列：ATCGCA,预测CpG的位置？



CpG Island: Viterbi算法 (1)

v		A	T	C	G	C	A
β	1						
C^+							
G^+							
A^+		0.1					
T^+							
C^-							
G^-							
A^-		0.15					
T^-							

0.5*0.2

0.5*0.3

CpG Island: Viterbi算法 (2)

v		A	T	C	G	C	A
β	1						
C^+							
G^+							
A^+		0.1					
T^+			0.015				
C^-							
G^-							
A^-		0.15					
T^-			0.0225				

0.1*0.5*0.2

0.15*0.5*0.2

0.1*0.5*0.3

0.15*0.5*0.3

CpG Island: Viterbi算法 (3)

v		A	T	C	G	C	A
β	1						
C^+				0.0034			
G^+							
A^+		0.1					
T^+			0.015				
C^-				0.00225			
G^-							
A^-		0.15					
T^-			0.0225				

0.015*0.5*0.3

0.0225*0.5*0.3

0.015*0.5*0.2

0.0225*0.5*0.2

CpG Island: Viterbi算法 (4)

v		A	T	C	G	C	A
β	1						
C^+				0.0034			
G^+					0.0005		
A^+		0.1					
T^+			0.015				
C^-				0.00225			
G^-					0.00034		
A^-		0.15					
T^-			0.0225				

0.0034*0.5*0.3

0.00225*0.5*0.3

0.0034*0.5*0.2

0.00225*0.5*0.2

CpG Island: Viterbi算法 (4)

v		A	T	C	G	C	A
β	1						
C^+				0.0034		0.000075	
G^+					0.0005		
A^+		0.1					
T^+			0.015				
C^-				0.00225		0.00005	
G^-					0.00034		
A^-		0.15					
T^-			0.0225				

0.0005*0.5*0.3

0.0034*0.5*0.3

0.0005*0.5*0.2

0.0034*0.5*0.2

CpG Island: Viterbi算法 (5)

v		A	T	C	G	C	A
β	1						
C^+				0.0034		0.000075	
G^+					0.0005		
A^+		0.1					0.0000075
T^+			0.015				
C^-				0.00225			
G^-					0.00034		
A^-		0.15					0.0000112
T^-			0.0225				

0.000075*0.5*0.2

0.000075*0.5*0.3

0.00005*0.5*0.3

0.00005*0.5*0.3

CpG Island: Viterbi算法 (6)

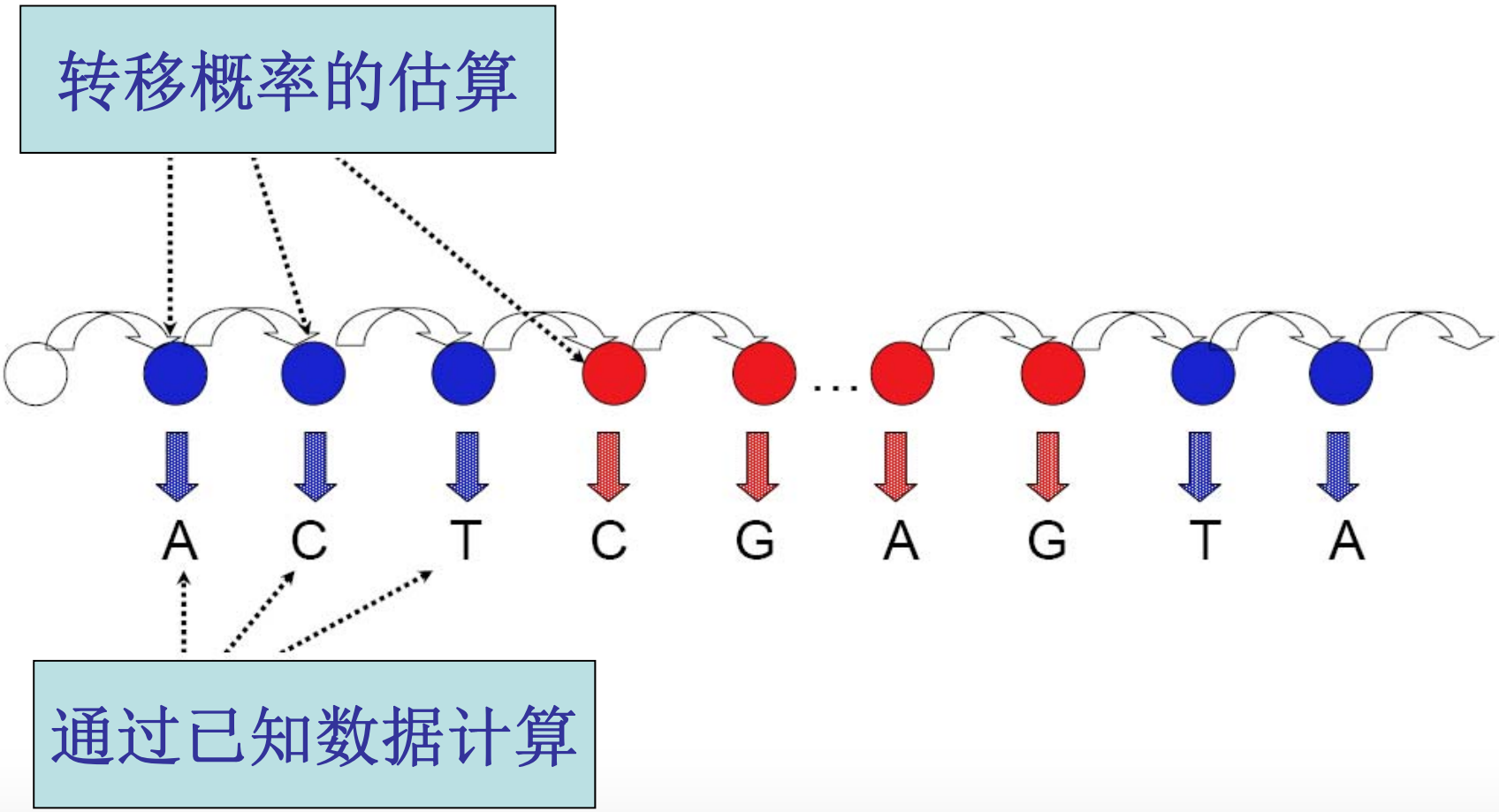
v		A	T	C	G	C	A
β	1						
C^+				0.0034		0.000075	
G^+					0.0005		
A^+		0.1					0.0000075
T^+			0.015				
C^-				0.00225		0.00005	
G^-					0.00034		
A^-		0.15					0.0000112
T^-			0.0225				

CpG Island: 预测结果

- ATCGCA: 其中，CGC被预测为CpG Island

  ATCGCA
- Viterbi算法：求出在当前结果最大的概率值，以及保存相应的路线
- 递归算法：动态规划的算法
- 该例中，我们假设状态转移概率矩阵已知
- 如何推算状态的概率矩阵？

HMM:参数估计



通过实际数据来估算参数：贝叶斯理论

给定一系列的训练序列O1,O2,…,On, 使得最终的转移概率矩阵概率最大：参数最大化

$$P(H = h_1, h_2, \ldots, h_n \mid O = o_1, o_2, \ldots, o_n)$$

$$= \frac{P(H = h_1, \ldots, h_n, O = o_1, \ldots, o_n)}{P(O = o_1, \ldots, o_n)}$$

$$= \frac{P(H = h_1, \ldots, h_n) P(O = o_1, \ldots, o_n \mid H = h_1, \ldots, h_n)}{P(O = o_1, \ldots, o_n)}$$

Conditional Prob:
P(A|B) = P(A,B)/P(B)

给定转移概率矩阵，实际序列的计算概率

$P(O = o_1, \ldots, o_n)$

最终的转移概率矩阵的概率

实际数据的隐含概率

给定转移矩阵时，实际数据的概率

通过实际数据来估算参数：贝叶斯理论 (2)

$P(O = o_1, \ldots, o_n)$

- 真实数据的实际概率，不随模型、参数的变化而改变。未知，也无法估计

- 对于最终的模型，即参数最大化的情况下，存在：

$$P(H = h_1, \ldots, h_n, O = o_1, \ldots, o_n) > P(H = h'_1, \ldots, h'_n, O = o_1, \ldots, o_n)$$

$$P(H = h_1, \ldots, h_n \mid O = o_1, \ldots, o_n) > P(H = h'_1, \ldots, h'_n \mid O = o_1, \ldots, o_n)$$

- 因此，可以将 $P(O = o_1, \ldots, o_n)$ 看成常数

参数估计：Baum-Welch算法

- 目的：给定观察值序列O，通过计算确定一个模型H， 使得P(O|H)最大
- 算法步骤：

  初始模型（待训练模型） H_0

  基于H_0 以及观察值序列O，训练新模型 H

  如果 log P(O|H) - log(P(O|H_0) < Delta，说明训练结果已经收敛， 算法结束

  否则，令H_0 ＝ H ，继续第2步工作

Baum-Welch算法: 操作流程

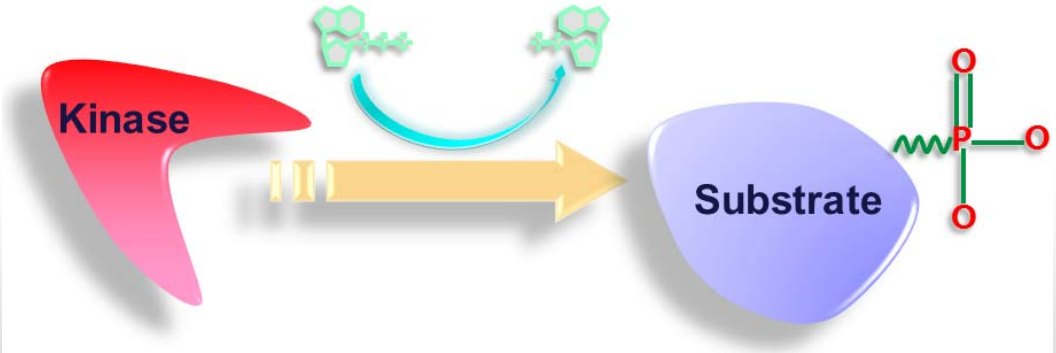
- 以CpG Island为例
- 需要估计的转移概率矩阵有四个值：P_{ii}, P_{ig}, P_{gg}, P_{gi}
- 初始化转移概率矩阵H_0，例如，都设为0.5
- 用Viterbi算法算出所有给定数据的结果及路径
- 根据所有的路径，可以得到N_{ii}, N_{ig}, N_{gg}, N_{gi}
- 计算新的P_{ii}, P_{ig}, P_{gg}, P_{gi}
- 如果结果收敛，停止；否则，重复4-6

马尔科夫 & 隐马尔科夫：总结

- 马尔科夫：所有状态已知或者可以通过实际数据的观测得到
- 隐马尔科夫：发散概率可以通过数据观测得到，状态之间的转移概率矩阵未知
- 给定转移概率矩阵，使用Viterbi算法，求出使给定序列概率值最大的路径
- 参数估计：Baum-Welch算法，递归算法，直到结果收敛

蛋白质翻译后修饰

- **POST**: DNA转录为RNA，RNA翻译成蛋白质之后发生
- 生化反应：产生或破坏共价键
- 发生在氨基酸的主链或侧链上
- 通常由特定酶催化
- 可发生在15种非疏水的氨基酸上
- 目前已发现>680 PTMs
  - 磷酸化、乙酰化、泛素化、SUMO化



http://www.uniprot.org/docs/ptmlist

第一个修饰底物

- 酪蛋白（Casein）
  - 牛奶中高丰度的磷酸化蛋白质
  - 1883年，瑞典生化学家Olof Hammarsten，首次发现包含磷酸根的蛋白质

Year	Event
1883	Casein is a phosphoprotein
1900	Vitellinic acid/phosvitin is a phosphoprotein
1932	Phosphoserine in vitellinic acid/phosvitin
1933	Phosphoserine in casein
1954	Casein is used as a model substrate for the detection of the first protein kinase activity
1960	Casein/phosvitin kinases are distinct from phosphorylase kinase
1969	Two distinct casein kinases, CK1 and CK2
1971	Complete amino acid sequence of α_{s1} casein is determined
1972	Casein kinase activity detected in Golgi fractions of lactating rat mammary gland
1988–1989	'Genuine/Golgi' casein kinase (G-CK) consensus is different from those of CK2 and CK1
1995	Casein kinase-1 and casein kinase-2 renamed protein kinase CK1 and CK2
1996	Specificity determinants for G-CK and the development of a synthetic peptide to detect G-CK activity with absolute specificity
2012	G-CK is identified as Fam20C

Olof Hammarsten
(1841–1932)

Tagliabracci et al., Science, 2012, 336, 1150-1153
Tagliabracci et al., Trends Biochem Sci, 2013, 38, 121-130

蛋白质磷酸化

- 蛋白激酶通过ATP水解所释放的能量，将磷酸基团转移到底物蛋白的丝氨酸/苏氨酸，或者酪氨酸的残基上
- 可逆性
  - 蛋白激酶：磷酸化 - *writer*
  - 蛋白磷酸酶: 去磷酸化 - *eraser*
  - 磷酸化结合结构域：识别磷酸化位点 - *reader*

Edmond H. Fischer

Edwin G. Krebs

The Nobel Prize in Physiology or Medicine 1992

Fischer, E.H. and Krebs, E.G. J Biol Chem, 1955, 216, 121-132

蛋白激酶PKC与信号转导

- 细胞内信号转导级联（intracellular signal transduction cascade）
- PKC：被脂质水解通路激活
- PKC促进肿瘤发生：Albert Lasker 1989
- 1995, the President of Kobe University

1932~2005

Yasutomi Nishizuka

西冢泰美

J. Biochem. 2010;148(2):125–130 doi:10.1093/jb/mvq066

JB THE JOURNAL OF BIOCHEMISTRY

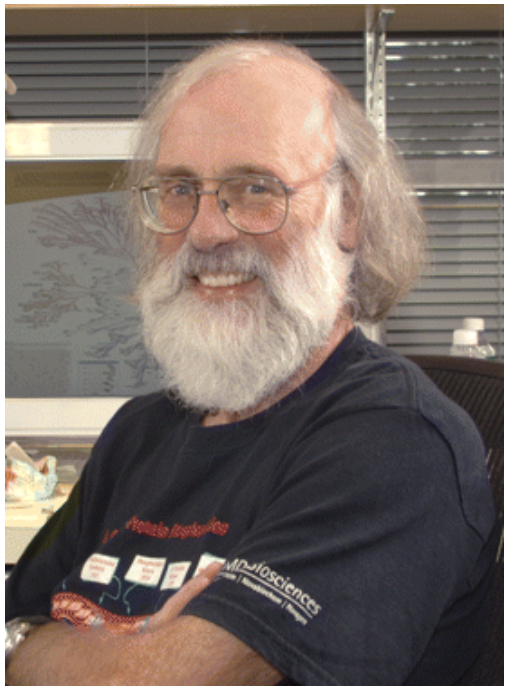
JB Reflections and Perspectives

Yasutomi Nishizuka: Father of protein kinase C

Inoue et al., J Biol Chem, 1977, 252, 7610-6

酪氨酸磷酸化

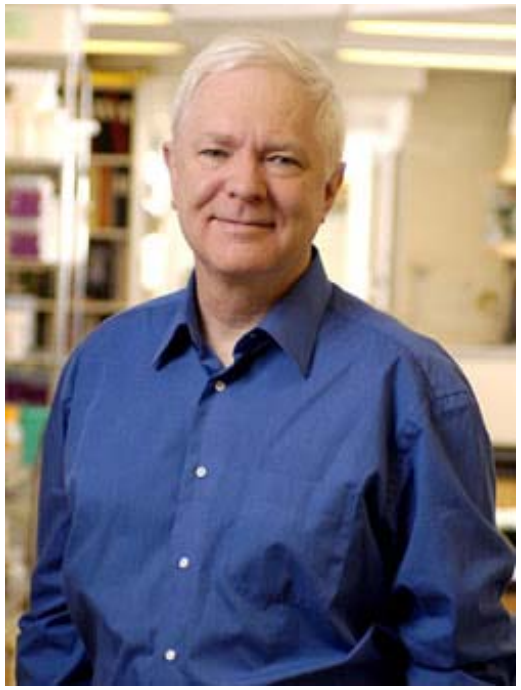
Tony Hunter: kinase king (1943~)



Discovery of tyrosine kinase

Hunter et al., PNAS, 1980, 77, 1311-1315

Tony Pawson, 1952~2013



Discovery of SH2 domain that binds with phosphotyrosine

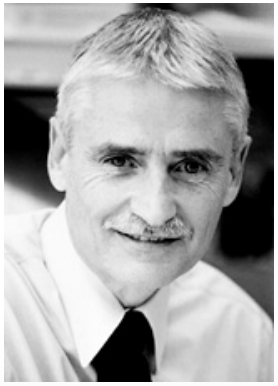
Sadowski et al., Mol Cell Biol, 1986, 6, 4396-4408

Hunter, J Cell Biol., 2008 181:572-3

细胞周期（Cell cycle）

- 1970, cdc1, *cdc2*, cdc3
- 1981, cdc2 is required
- 1984, cdc2/cdk1 activity is required
- How about Tim Hunt?

Marc W. Kirschner



Leland H. Hartwell

Tim Hunt

Paul M. Nurse

The Nobel Prize in Physiology or Medicine 2001

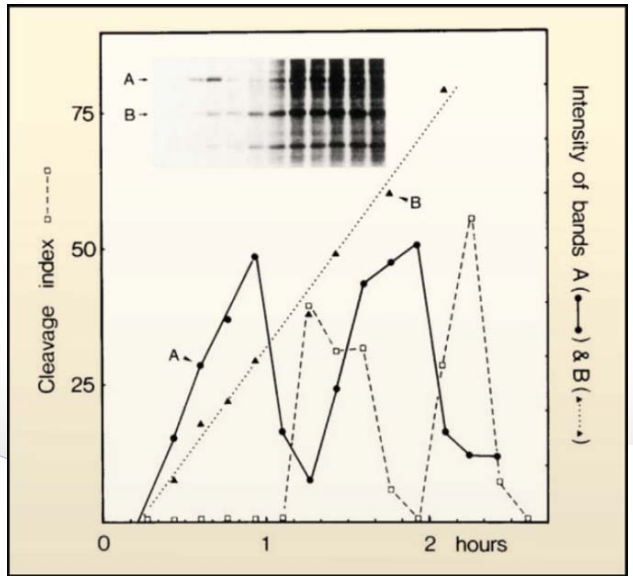
Hartwell et al., PNAS, 1970, 66:352-9
Nurse et al., Nature, 1981, 292, 558-60
Newport et al., Cell, 1984, 37, 731-42

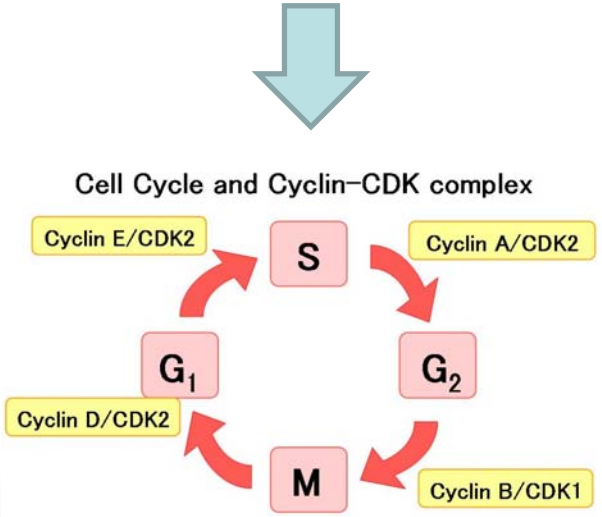
周期素Cyclin的发现

- July 22nd, 1982, Cyclin A & B
- 1989, Cyclins control cdks
- Why Tim Hunt?
  - *Sea urchin*
  - SDS-Page
  - Marc W. Kirschner

I was very lucky.

Marc W. Kirschner





Evans et al., Cell, 1983, 33, 389-96
Murray et al., Nature. 1989, 339, 275-80
Murray et al., Nature, 1989, 339, 280-6

磷酸化位点鉴定 & 预测

- 磷酸化调控了细胞生命中绝大部分的分子、细胞和生理过程，在细胞内，约1/3的蛋白质被磷酸化
- 在哺乳动物的基因组中，编码了约~500个蛋白激酶
- 磷酸化位点的鉴定，是理解整个细胞内的信号通路、网络以及进一步的系统生物学研究的基础
- Phosphoproteomics：磷酸化底物、位点的高通量检测
  - 质谱
  - 蛋白质、肽段和激酶芯片
- 目前进展: >1,600,000个磷酸化位点
- 挑战: 哪些激酶磷酸化这些位点？

GPS算法

- 问题：给定蛋白质序列，能否准确预测究竟哪些位点可能发生磷酸化？
- 基本假设：
  - 相似肽段，相似功能
  - 相似激酶，相似底物
- 打分策略：
  - Phosphorylation site peptide: PSP(3, 3): XXX-[S/T/Y]-XXX
  - 通过BLOSUM62矩阵的打分来评估两个PSP(3, 3)之间的相似性。
  - 两个7肽之间相似性的打分定义如下：

$$S(\mathrm{A},\mathrm{B}) = \sum_{1 \leqslant i \leqslant 7} \mathrm{Score}(\mathrm{A}[i],\mathrm{B}[i])$$

打分策略

- 给定两条肽段：

  AQE<u>S</u>ILR (磷酸肽)

  IQESLIR (未知)

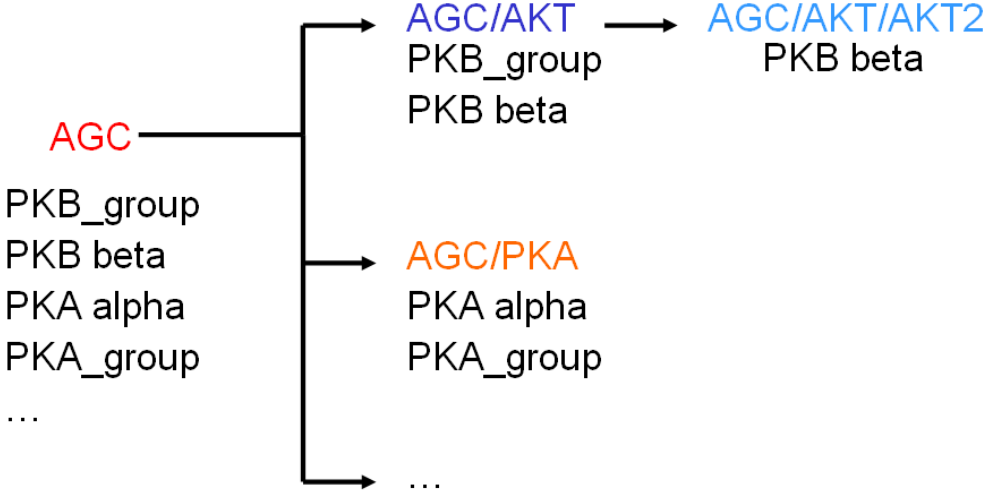
- 相似性分值：
  - **-1+5+5+6+2+2+5=24**
- 算法流程：
  - **(1) 给定肽段与每一条已知磷酸肽比较，计算相似性分值**
  - **(2) 分值<0则设为0**
  - **(3) 最终分值：平均值**

BLOSUM62 (部分)

	A	R	N	S	T	Q	E	G	H	I	L
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3
S	-2	-2	1	6	-3	0	2	-1	-1	-3	-4
T	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4

激酶的层次分类

- GPS 1.0 & 1.0：根据序列相似性分类激酶
- GPS 2.0：激酶分类的4个层次，包括组、家族、亚家族和单个激酶
- 将已知磷酸化位点根据激酶信息进行归类



为什么是BLOSUM62矩阵？

- 给定两条肽段：

  AQE<u>S</u>ILR (磷酸肽)

  IQESLIR (未知)

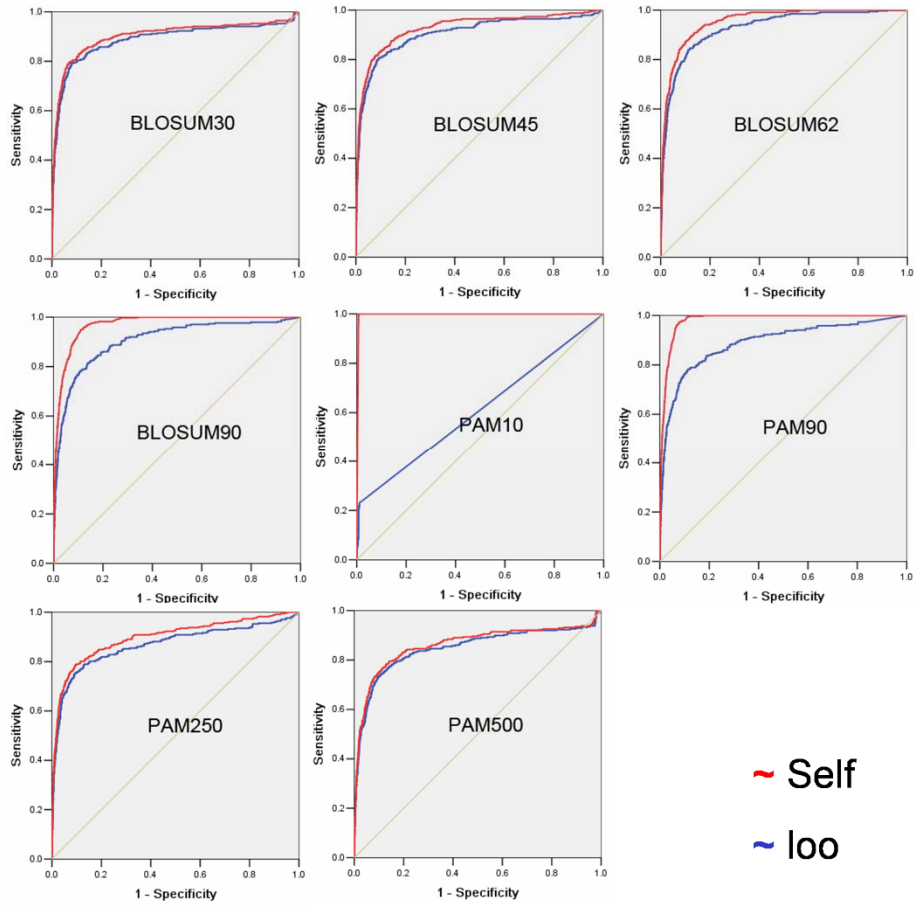
- 相似性分值：
  - **-1+5+5+6+2+2+5=24**
- 算法流程：
  - **(1) 给定肽段与每一条已知磷酸肽比较，计算相似性分值**
  - **(2) 分值<0则设为0**
  - **(3) 最终分值：平均值**

BLOSUM62 (部分)

	A	R	N	S	T	Q	E	G	H	I	L
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3
S	-2	-2	1	6	-3	0	2	-1	-1	-3	-4
T	0	-3	-3	-3	9	-1	-4	-3	-3	-1	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4

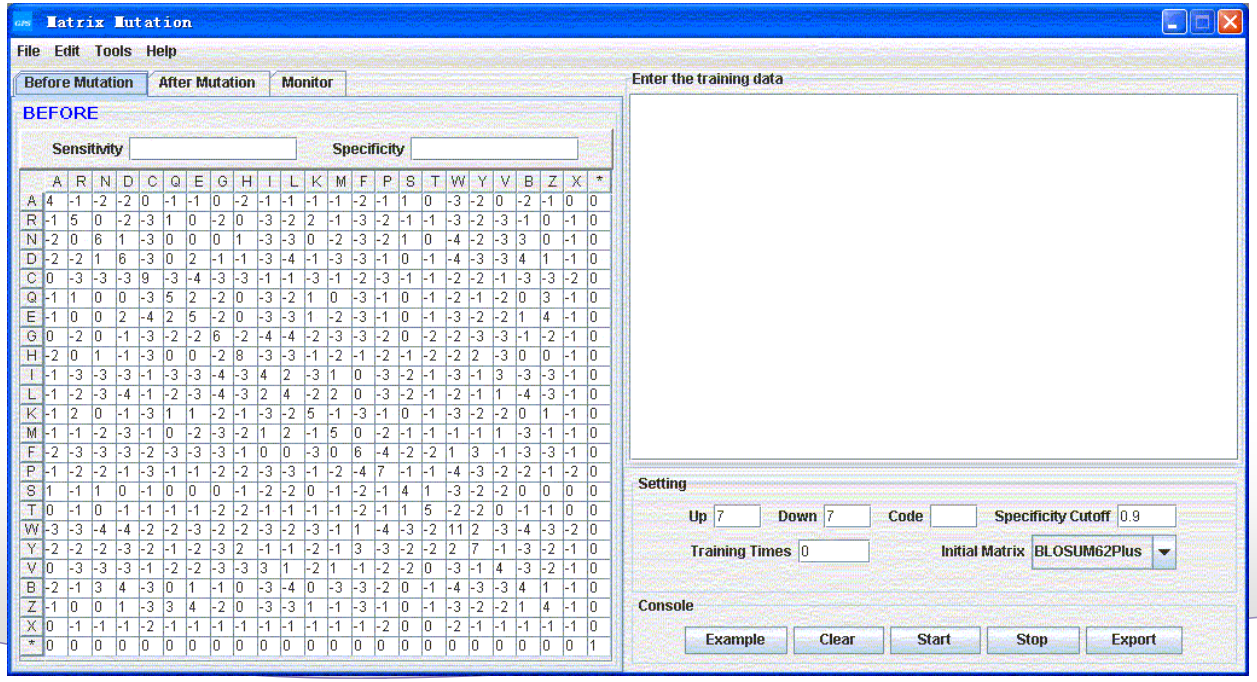
打分矩阵

- 利用自检法和除一法评估性能
- 不同矩阵的性能不同
- 问题：如何找到“最优”的矩阵？



GPS2.0：矩阵突变

- 以BLOSUM62为初始矩阵，利用除一法计算*Sn* & *Sp*值
- 将*Sp* 固定为90%，从矩阵中随机挑选任意值，进行+1或-1操作，重新计算*Sn*值
- 若*Sn*值提高，则该操作保留
- 重复该过程10,000次



为什么是PSP(3,3)

- 给定两条肽段：

  AQESLLR (磷酸肽)

  IQESLIR (未知)

- 相似性分值：
  - -1+5+5+6+2+2+5=24
- 算法流程：
  - (1) 给定肽段与每一条已知磷酸肽比较，计算相似性分值
  - (2) 分值<0则设为0
  - (3) 最终分值：平均值

BLOSUM62 (部分)

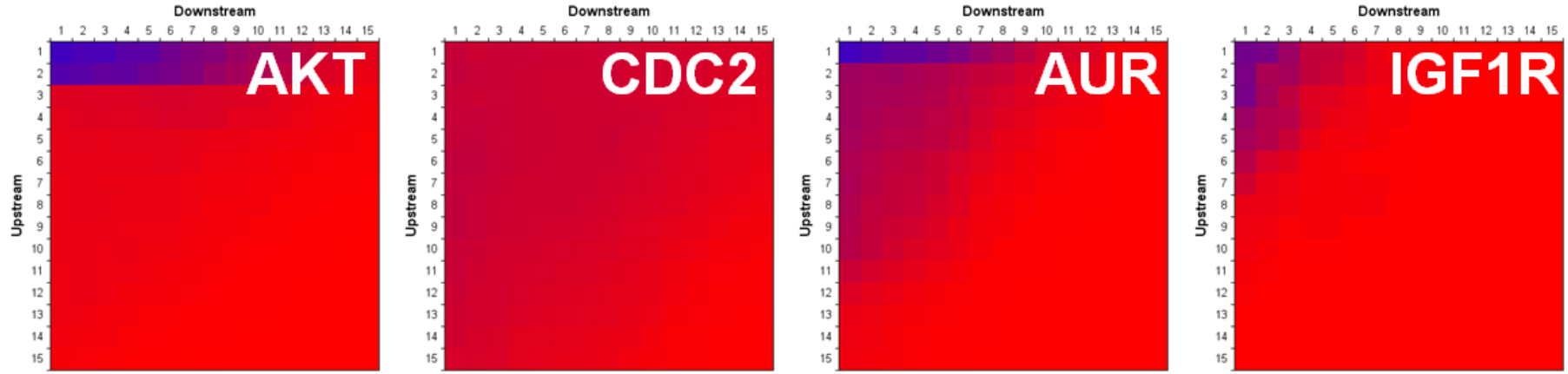
	A	R	N	S	T	Q	E	G	H	I	L
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3
S	-2	-2	1	6	-3	0	2	-1	-1	-3	-4
T	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4

GPS 2.1: 肽段选择

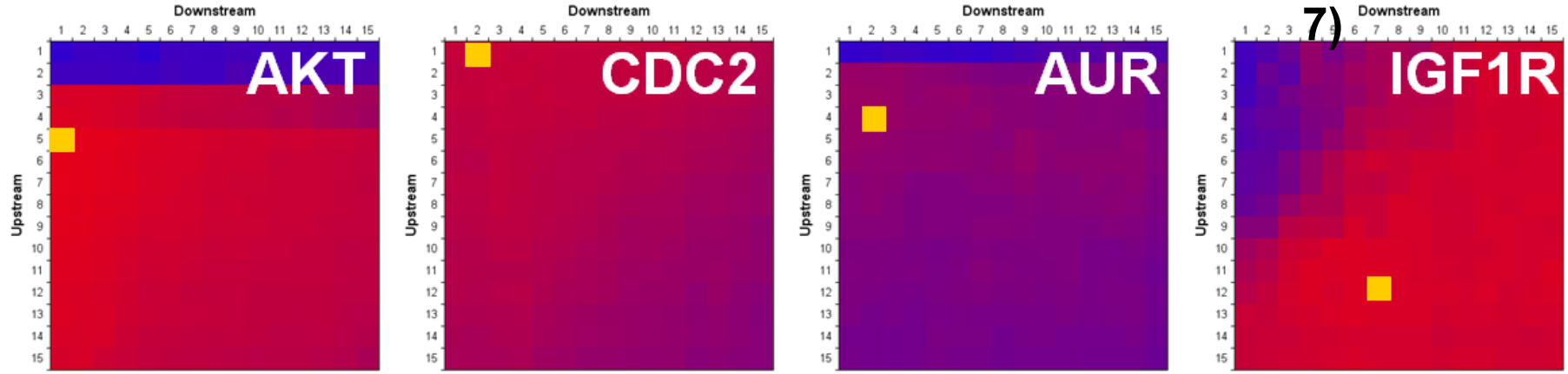
- 当PSP(*m*, *n*)中*m*, *n*值增加时
  - 自检法↑–> *Sn* = *Sp* = 100% (过拟合)
  - 除一法先↑，再↓–> 存在峰值
- 肽段选择
  - PSP(*m*, *n*) (*m*=1, …,15; *n*=1, …, 15)，*Sp*为85%, 90%和95%时，计算平均*Sn*值
  - 最优PSP(*m*, *n*): Max(*Sn*)
- 与GPS 2.0相比
  - 除一法*Sn*值：15.62%↑
  - 自检法*Sn*值：2.28%↓

Peptide Selection (PS)

A.



B. PSP(5, 1) PSP(1, 2) PSP(4, 2) PSP(12, 7)



蛋白质序列编码

- 将蛋白质序列转化为向量或矩阵
- GPS算法的打分函数：
  - 给定PSP(*m*, *n*)的*a*残基与已知磷酸肽*b*残基的相似性分值

$$S_{ab} = \frac{1}{\sum_{j=1}^{L} C_j} \sum_{j=1}^{L} C_j \times M[a,b] \times W_j$$

L：PSP(*m*, *n*)长度；C_j: 第*j*位*ab*氨基酸对的数量；W_j: 第*j*位权重；*M*：打分矩阵

  - 20种常见氨基酸 + N或C端的空位符“*”
  - S_{ab}值共有[21*(21+1)]/2=231种 ($S_{ab} = S_{ba}$)
  - 因此，PSP(*m*, *n*)可表征为：

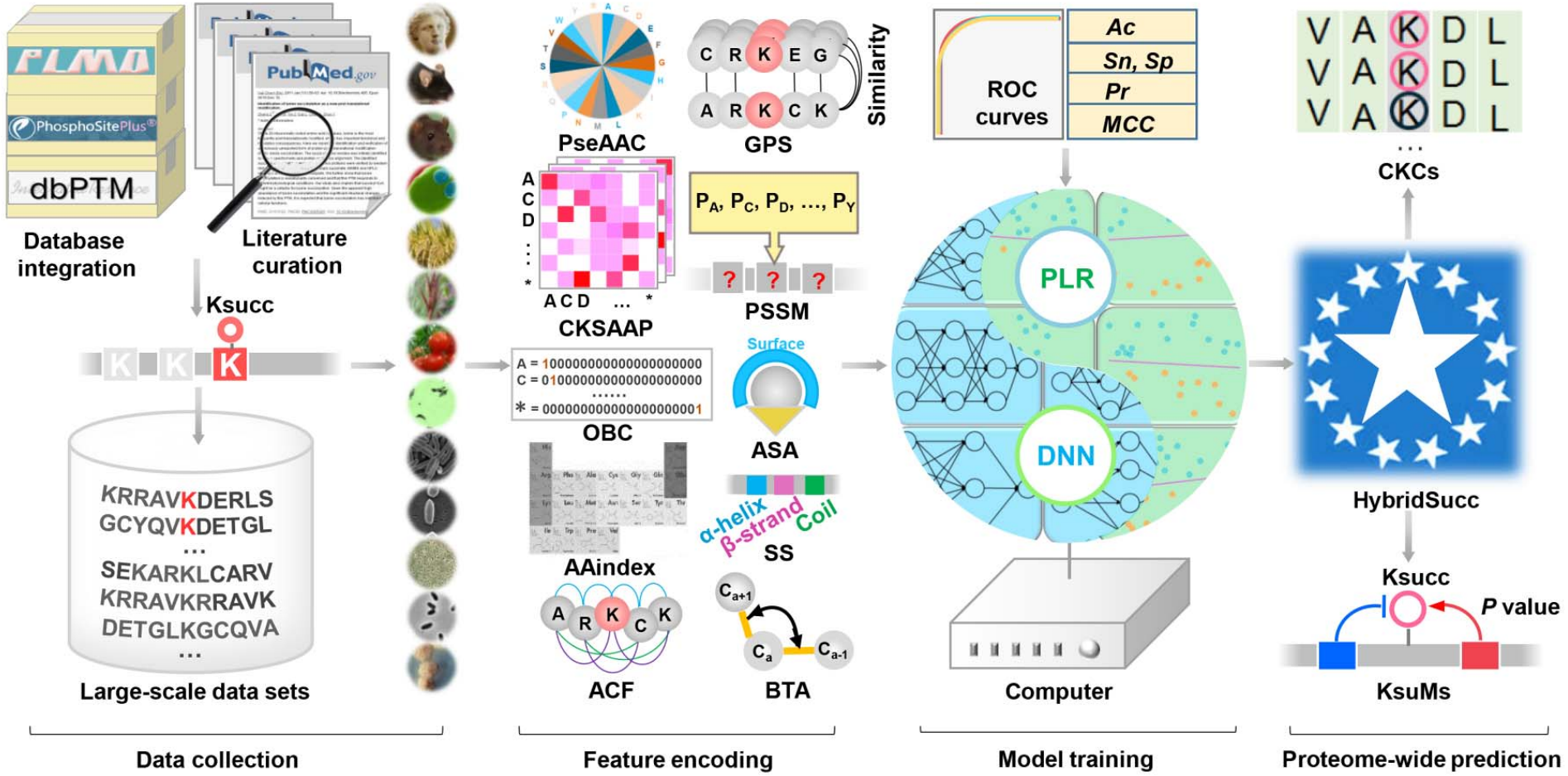
$$V = (S_{AA}, S_{AC}, S_{AD}, \cdots, S_{**})_{231}$$

其他编码方法

- 赝氨基酸组成（Pseudo amino acid composition）
  - 统计出现频率：$V_i = (F_A, F_C, F_D, \cdots, F_Y, F_*)_{21}$
- *k*间隔氨基酸对组成（Composition of k-spaced amino acid pairs）
  - 统计*k*个空位间隔的氨基酸对出现频率：21*21*(k_{max}+1)
  - $V_i = [(C_{AA}, C_{AC}, \cdots, C_{**})_{441}]_{K=0}, \cdots, [(C_{AA}, C_{AC}, \cdots, C_{**})_{441}]_{K=K_{max}}$
- 正交二进制编码（Orthogonal binary coding/one-hot coding）
  - Alanine：[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
  - Cysteine：[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

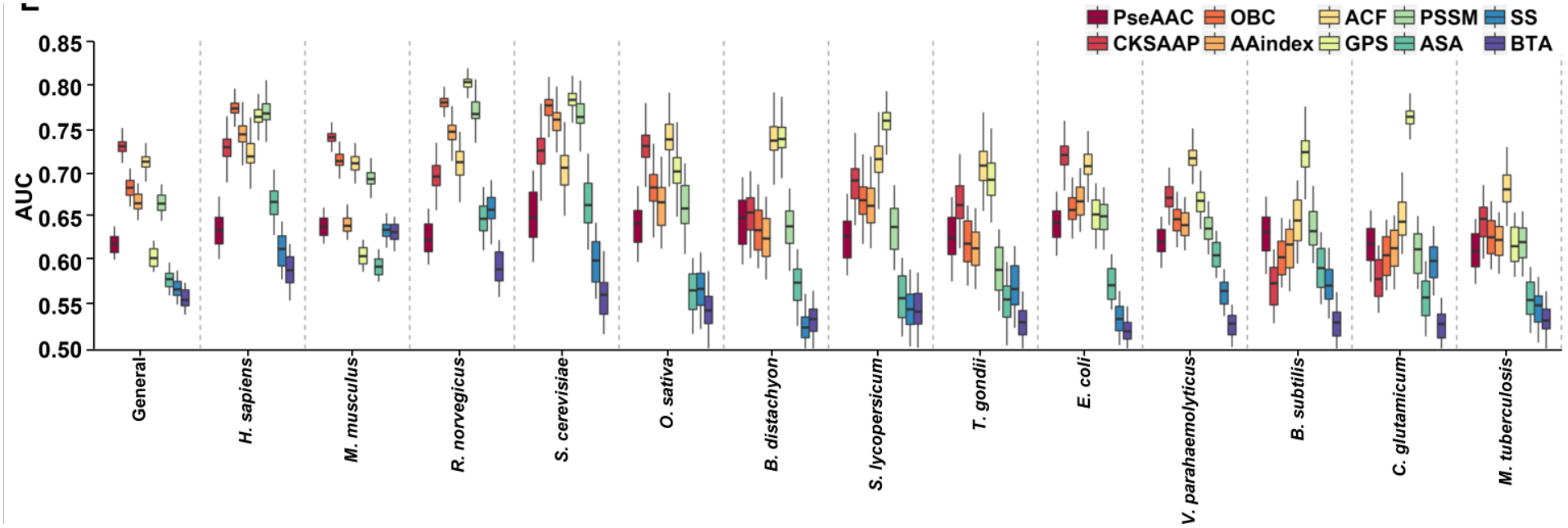
混合学习框架

- http://hybridsucc.biocuckoo.org/

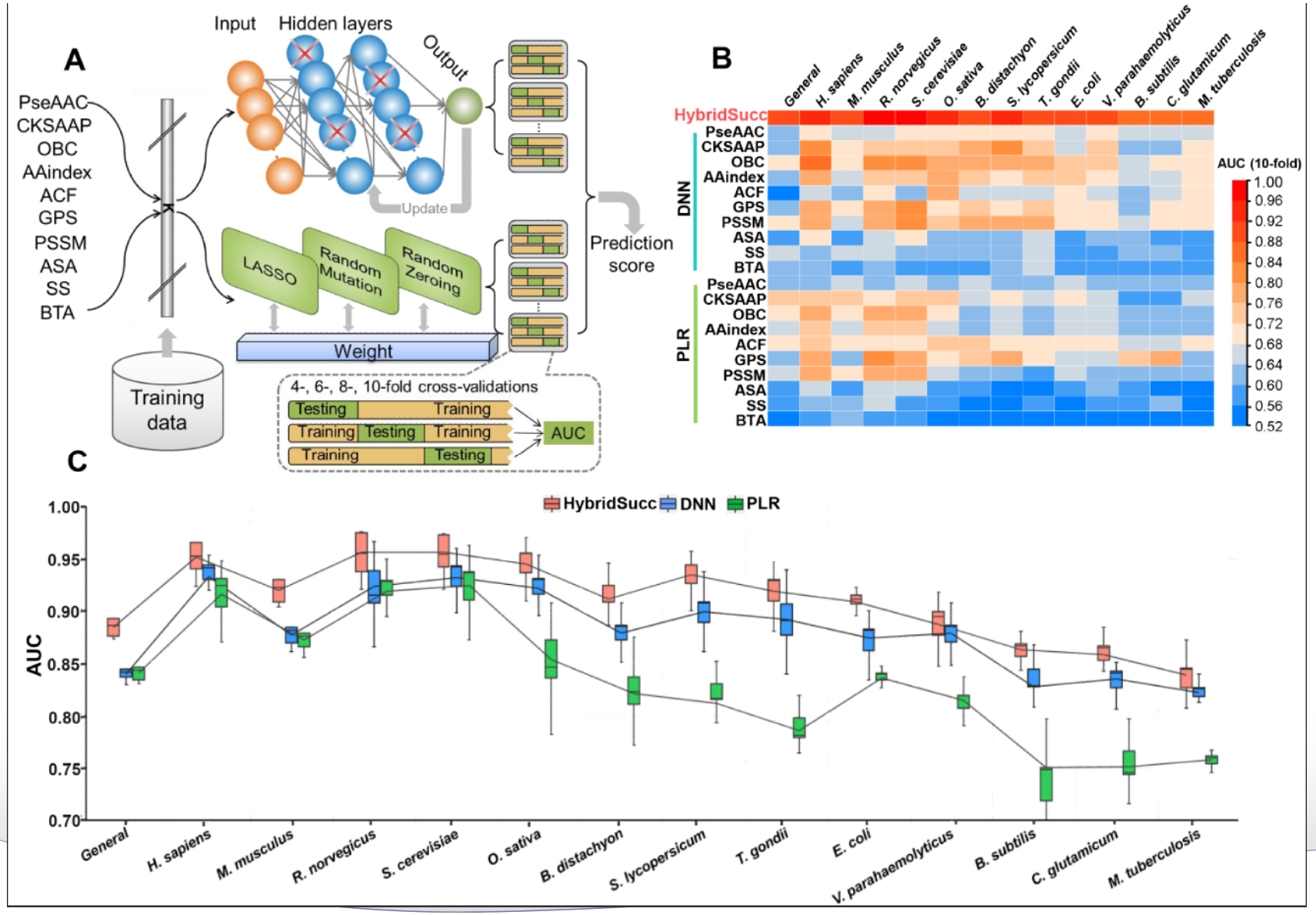


序列、结构特征

- 7种序列特征，3种结构特征



传统机器学习 *vs.* 深度学习



Graphic Presentation System

❒ GPS-Palm: parallel CNNs + PLR, 10 feature types