清華大學
Tsinghua University

第二次作业分享

分工：

陈怀玉　必做题1,2(1)

张霁辰　必做题2(2)(3)

张秋迎　练习题

CONTENTS

清華大學 Tsinghua University

- BLAST数据库简介
- 什么是Bootstrap？
- 系统发育树构建算法原理
- PAM矩阵简介

CONTENTS

清華大學 Tsinghua University

- BLAST数据库简介
- 什么是Bootstrap？
- 系统发育树构建算法原理
- PAM矩阵简介

NCBI中BLAST 常见程序

- **BLASTP**：蛋白序列到蛋白库
- **BLASTN**：核酸序列到核酸库
- **BLASTX**：核酸序列到蛋白库
- **TBLASTN**：蛋白序列到核酸库
- **TBLASTX**：核酸序列到核酸库

BLASTP中常见数据库

- **Nr**：All non-redundant GenBank CDS translations + RefSeq Proteins + PDB + SwissProt + PRF
- **Refseq**
- **Swissprot**
- **Pat**
- **Pdb**
- **Month**
- **Evn nr**

BLASTN中常见数据库

- **Nr**：All GenBank + RefSeq Nucleotides + EMBL + DDBJ + PDB
- **Refseq_rna**
- **Refseq_genomic**
- **Est**：ex:est human, est mouse, est others
- **Htgs**
- **Pat、pdb、month**等常见核酸数据库

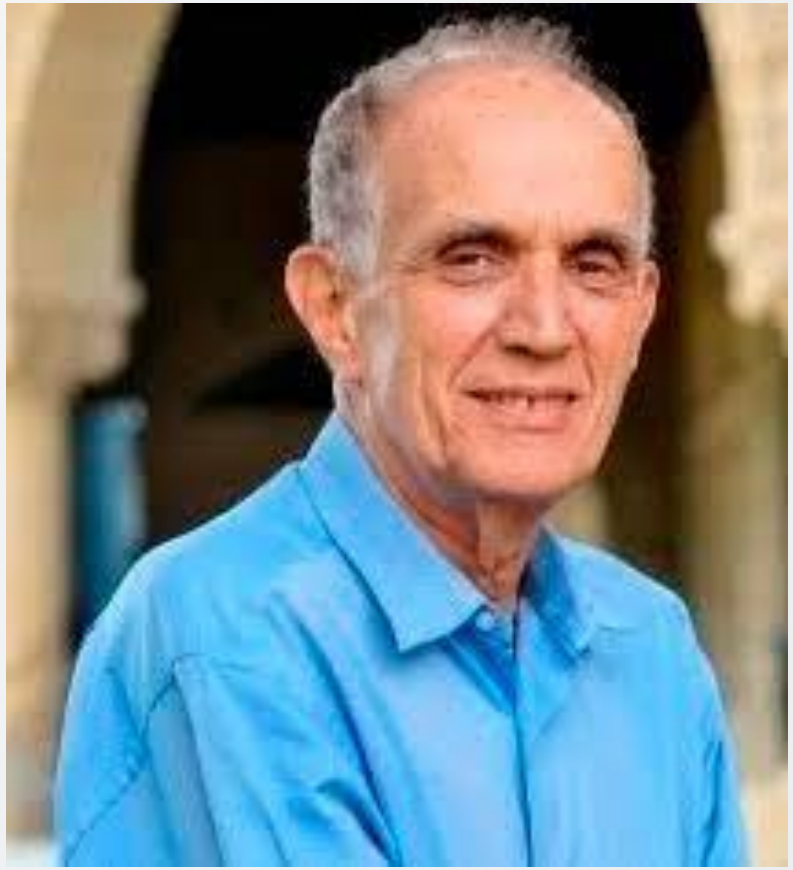
CONTENTS

清華大學 Tsinghua University

- BLAST数据库简介
- 什么是Bootstrap？
- 系统发育树构建算法原理
- PAM矩阵简介

Bootstrap Test

- 实质：对观测信息进行再抽样，进而对总体的分布特性进行统计推断
- 充分利用给定的观测信息
- 具稳健性和高效率
- 机器学习领域应用广泛



Bradley Efron，著名统计学家

Bootstrap Consensus Tree

- 多次bootstrap test得到的平均结果
- 无遗传距离信息
- 数字代表频率参数
- 频率参数反应进化树是否可靠

Bootstrap Consensus Tree vs. Original Tree

- **Original tree**
  - 最优系统树，树枝长短精确表示遗传距离数据，可显示频率参数；可确定树根
  - 是bootstrap test构建的 N次株树中的一株，未经过多棵树合并
- **Bootstrap consensus tree**
  - N次株树的该树枝的出现频率，反应该树枝的可信度

CONTENTS

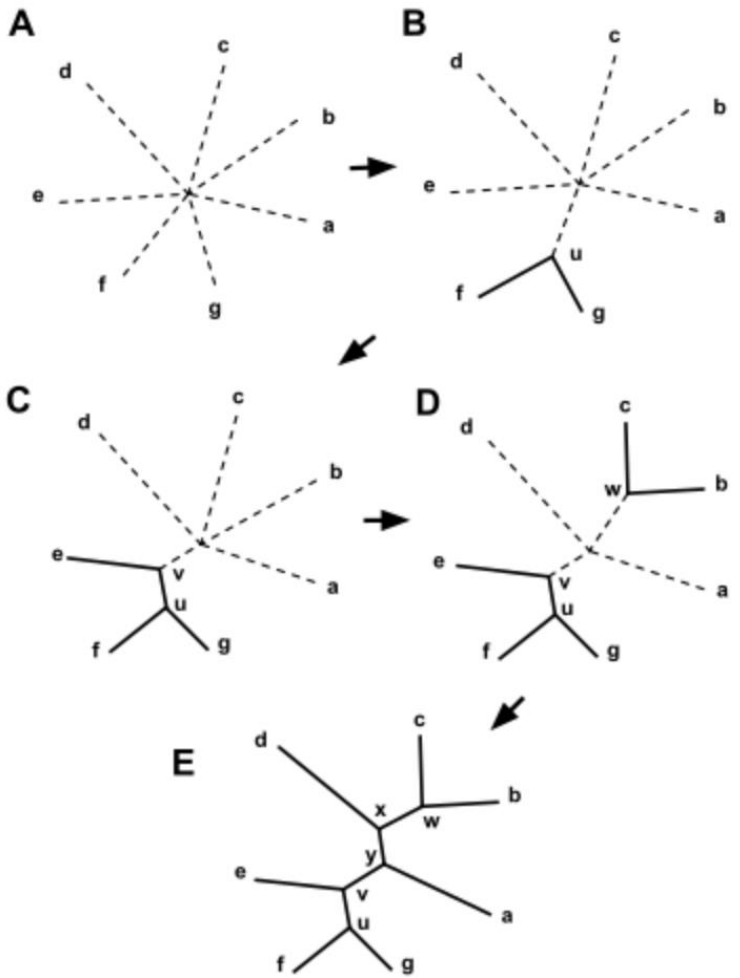
清華大學 Tsinghua University

- BLAST数据库简介
- 什么是Bootstrap？
- 系统发育树构建算法原理
- PAM矩阵简介

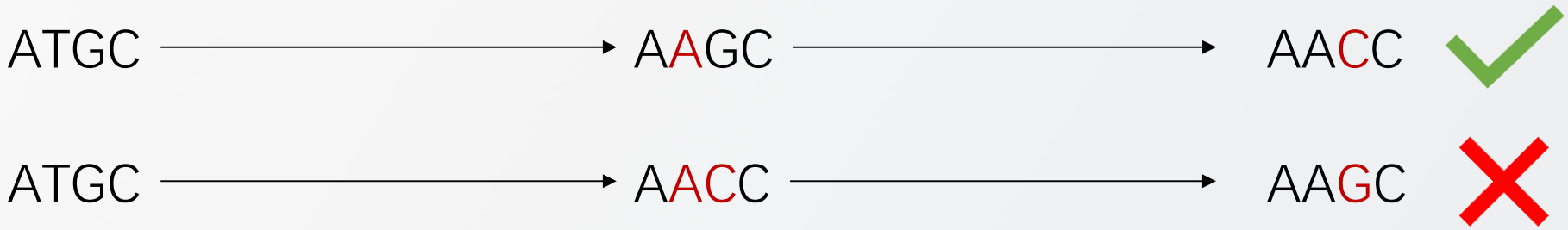
- 距离法

- 以邻接法为例（Neighbor-Joining，NJ）
  - 确定距离最近的成对分类单元从而使系统树的总距离达到最小

  - 优点：速度最快
  - 缺点：序列上的所有位点等同对待，且所分析的序列的进化距离不能太大



- 最大简约法（Maximum parsimony，MP）
  - “解释一个过程的最好理论是所需假设数目最少的那一个”
  - 依托于进化过程中所需核苷酸或氨基酸替代数目最少的假说，首先计算所有可能的拓扑结构，然后挑选出所需替代数最小的那个拓扑结构作为最优树。



  - 在分析序列上存在较多的回复突变或平行突变，而被检验的序列位点数又比较少的时候，最大简约法可能会给出一个不合理的或者错误的进化树推导结果。

- 最大似然法（Maximum likelihood，ML）
  - 基本思想：当从模型总体随机抽取n组样本观测值后，最合理的参数估计量应该使得从模型中抽取该n组样本观测值的概率最大。
  - 似然函数：参数给定时观测数据的概率

例子：抛硬币10次，得到：反正正正正反正正正反

假设：正面朝上的概率为p, 反面则为1-p

P(反正正正正反正正正反)=(1-p)*p*p*p*p*(1-p)*p*p*p*(1-p)= $p^7 \times (1-p)^3$

当p=0.7时，该函数取得最大值，即P(..)最有可能发生

- 最大似然法（Maximum likelihood，ML）
  - 将每个位点所有可能出现的残基替换概率进行累加，产生特定位点的似然值，对所有可能的树都计算似然函数，选取似然函数最大的那棵树
  - 假定所有序列都是从一条碱基进化而来，拥有共同祖先，给定一定的进化模型后，什么样的拓扑结构、多长的树枝、什么样的模型参数最有可能产出当前各序列

  - 优点：在进化模型确定的情况下，与进化事实吻合最好
  - 缺点：计算耗时，速度慢

不同算法的选择

- 一般情况，若有合适模型，ML的效果较好
- 近缘序列：一般使用MP（基于的假设少）
- 远缘序列：一般使用NJ或ML

以同样的方法分析同样的数据，所产生的树有可能存在不同吗?

- 有可能
- 最大简约法和最大似然法为了节约运算成本，会采用近似最优的启发式搜索等方法。如果算法是随机选取道路搜索起点的，则将有可能每次获得的近似最优解不同
- 在构建Bootstrap consensus tree的过程中，会随机产生1000次取样，因此用同样的方法分析同样的数据，也会产生不完全相同的树

CONTENTS

清華大學 Tsinghua University

- BLAST数据库简介
- 什么是Bootstrap？
- 系统发育树构建算法原理
- PAM矩阵简介

- Margaret Belle (Oakley) Dayhoff
- 1925-1983
- 生物信息学奠基人
  - PAM矩阵
  - 世界上第一个在线蛋白数据库
  - 氨基酸单字母代码
  - 用计算机构建系统发育树
  - …… ……
- 美国生物物理学协会前主席、秘书长

A Model of Evolutionary Change in Proteins

M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt

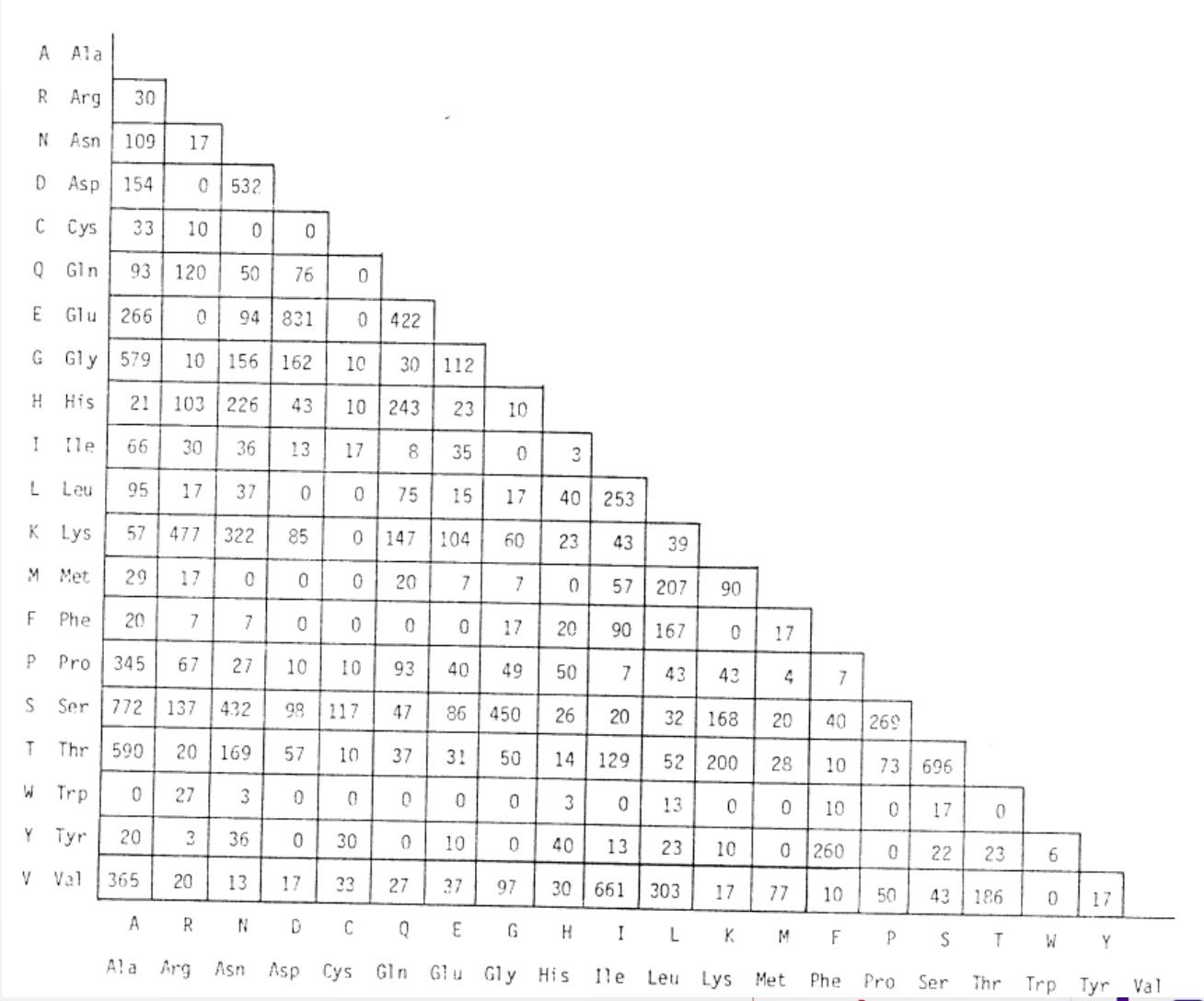
References

1. Dayhoff, M.O., Eck, R.V., and Park, C.M., *in* Atlas of Protein Sequence and Structure 1972, Vol.5, ed. Dayhoff, M.O., pp.89-99, Nat. Biomed. Res. Found., Washington, D.C., 1972
2. Schwartz, R.M., and Dayhoff, M.O., *in* Evolution of Protein Molecules, ed. Matsubara, H., and Yamanaka, T., pp.1-16, Japan Sci. Soc. Press, Tokyo, 1978
3. Schwartz, R.M., and Dayhoff, M.O., *in* Origin of Life, ed. Noda, H., pp.457-469, Center for Academic Pub. Japan/Japan Sci. Soc. Press, Tokyo, 1978
4. Dayhoff, M.O., and Eck, R.V., Atlas of Protein Sequence and Structure 1967-68, pp.33-45, Nat. Biomed. Res. Found., Silver Spring, Md., 1968

Dayhoff, M., Schwartz, R., & Orcutt, B. (1978). a model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 5, 345-352.

4.PAM矩阵简介

- ‘Accepted Point Mutation’
- 由大量观察数据得来
  - 71棵进化树
  - 差异小于15%
- i,j替换出现次数A(i,j)
- i氨基酸出现频率f(i)
- i氨基酸的相对突变性m(i)



		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
A	Ala																			
R	Arg	30																		
N	Asn	109	17																	
D	Asp	154	0	532																
C	Cys	33	10	0	0															
Q	Gln	93	120	50	76	0														
E	Glu	266	0	94	831	0	422													
G	Gly	579	10	156	162	10	30	112												
H	His	21	103	226	43	10	243	23	10											
I	Ile	66	30	36	13	17	8	35	0	3										
L	Leu	95	17	37	0	0	75	15	17	40	253									
K	Lys	57	477	322	85	0	147	104	60	23	43	39								
M	Met	29	17	0	0	0	20	7	7	0	57	207	90							
F	Phe	20	7	7	0	0	0	0	17	20	90	167	0	17						
P	Pro	345	67	27	10	10	93	40	49	50	7	43	43	4	7					
S	Ser	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269				
T	Thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696			
W	Trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0		
Y	Tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6	
V	Val	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val

- 突变概率矩阵M(i,j)
  - 非对角线元素：氨基酸j变为氨基酸i的概率
  - 对角线元素：氨基酸i保持不变的概率
    - 保持总概率为1
  - $\sum f(i)M(i,i)$ 的含义？
    - 在一段时间间隔内，某个未知氨基酸不发生突变的概率
    - 在1PAM的时间间隔内，这个概率被定义为0.99，由此求解出λ

4.PAM矩阵简介

ORIGINAL AMINO ACID

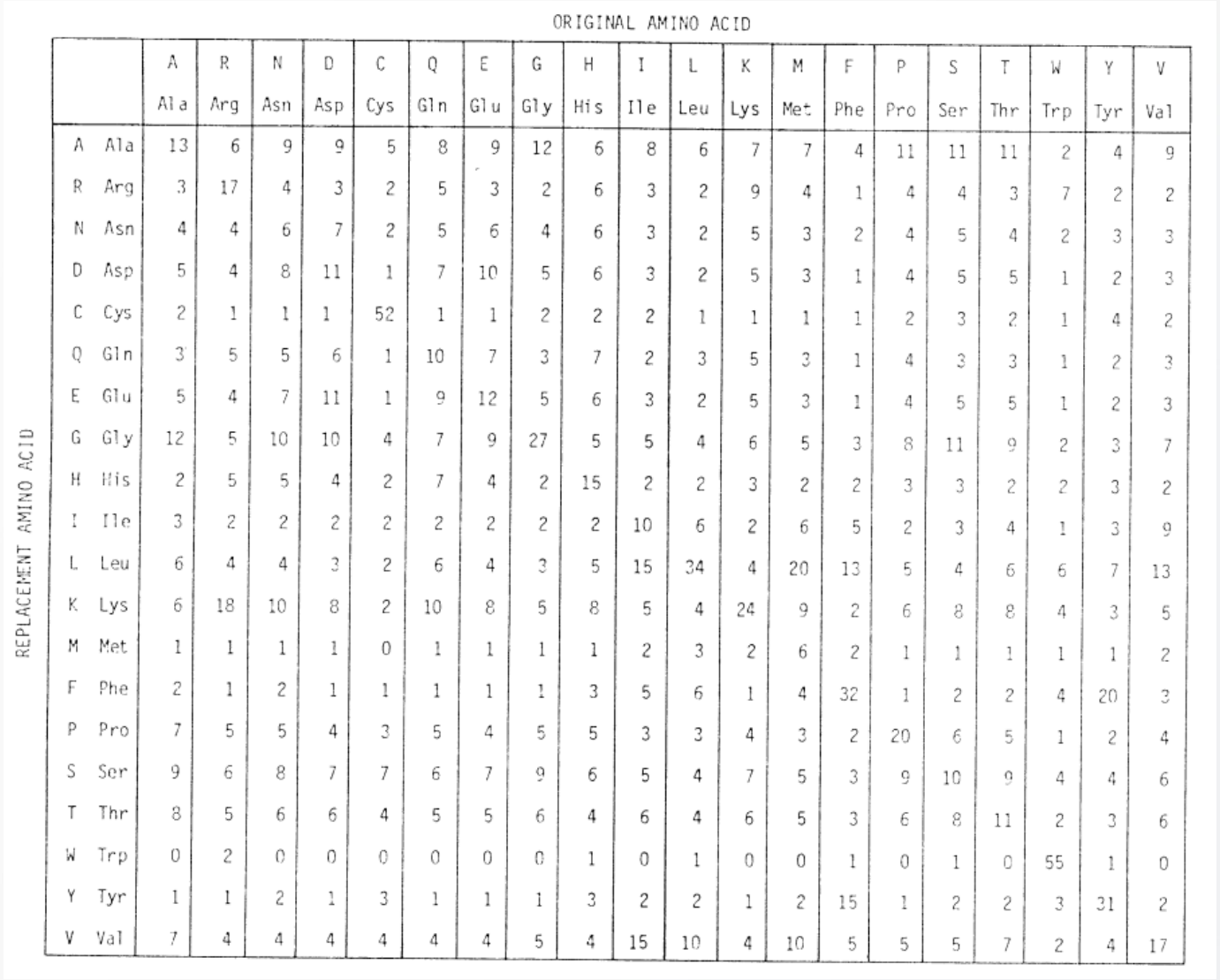
REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

1PAM=1%差异

Correspondence between Observed Differences and the Evolutionary Distance

Observed Percent Difference	Evolutionary Distance in PAMs
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

4.PAM矩阵简介



ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

非对称性的
PAM250矩阵

4.PAM矩阵简介

- 非对称性PAM250的缺点：
  - M(i,j)的值受到f(i)的影响
    - $R(i,j) = \frac{M(i,j)}{f(i)} = \frac{f(j)\cdot M(i,j)}{f(j)\cdot f(i)} = \frac{f(i)\cdot M(j,i}{f(i)\cdot f(j)}$
    - $= R(j,i)$
  - 概率乘法，难以计算
    - $PAM(i,j) = logR(i,j)$
- 由此得到对称性PAM250
  - 失去了数值实际意义
  - 方便了计算和储存

C	Cys	12																						
S	Ser	0	2																					
T	Thr	-2	1	3																				
P	Pro	-3	1	0	6																			
A	Ala	-2	1	1	1	2																		
G	Gly	-3	1	0	-1	1	5																	
N	Asn	-4	1	0	-1	0	0	2																
D	Asp	-5	0	0	-1	0	1	2	4															
E	Glu	-5	0	0	-1	0	0	1	3	4														
Q	Gln	-5	-1	-1	0	0	-1	1	2	2	4													
H	His	-3	-1	-1	0	-1	-2	2	1	1	3	6												
R	Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6											
K	Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5										
M	Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6									
I	Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5								
L	Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6							
V	Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4						
F	Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9					
Y	Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10				
W	Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17			
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W			
		Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp			

清華大學
Tsinghua University

谢谢大家！

Thank you for watching