生物信息学

第三章 序列分析的数学基础

概率模型

- 概率模型:一个能够通过不同的概率产生不同结果的模型。概率模型可以模拟或者仿真某一类型的所有事件，并且对每个事件赋予一个概率
- 色子模型：一个色子存在6个概率值：p_1, p_2, …, p_6，其中掷出i的概率为p_i (i=1, 2, …, 6)。因此：
  - $pi \geq 0$，且 $\sum_{i=1}^{6} p_i = 1$
- 考虑三次连续的掷色子，结果为 [1，6，3]，则总概率为：$p_1 p_6 p_3$

概率分布

- 考虑连续变量x，例如：物体的重量。重量确切为1公斤时的概率为0
- 变量的区间：$P(x_0 \le x \le x_1)$
- 当区间无限小 -> 0时，上式：
  - $P(x - \delta x/2 \le x \le x + \delta x/2) = f(x)\delta x$
- $f(x)$称为概率密度函数
- 因此：$P(X_0 \le x \le X_1) = \int_{x_0}^{x_1} f(x)dx$ 且 $\int_{-\infty}^{\infty} f(x)dx = 1$

二项分布

- 事件只有两种可能出现的结果。例如掷硬币，正面记为“1”，反面记为“0”
- 则掷硬币*N*次，有*k*次是1的概率为：

$$P(k)=\binom{N}{k}p^{k}(1-p)^{N-k}$$

二项分布 (2)

平均数$E(x) = m$

$$m = \sum_{k=1}^{N} k \binom{N}{k} p^k (1-p)^{N-k} = Np$$

标准方差 $\mathbf{Var\ X=\sigma^2}$

$$\sigma^2 = \sum_{k=1}^{N} (k-m)^2 \binom{N}{k} p^k (1-p)^{N-k} = Np(1-p)$$

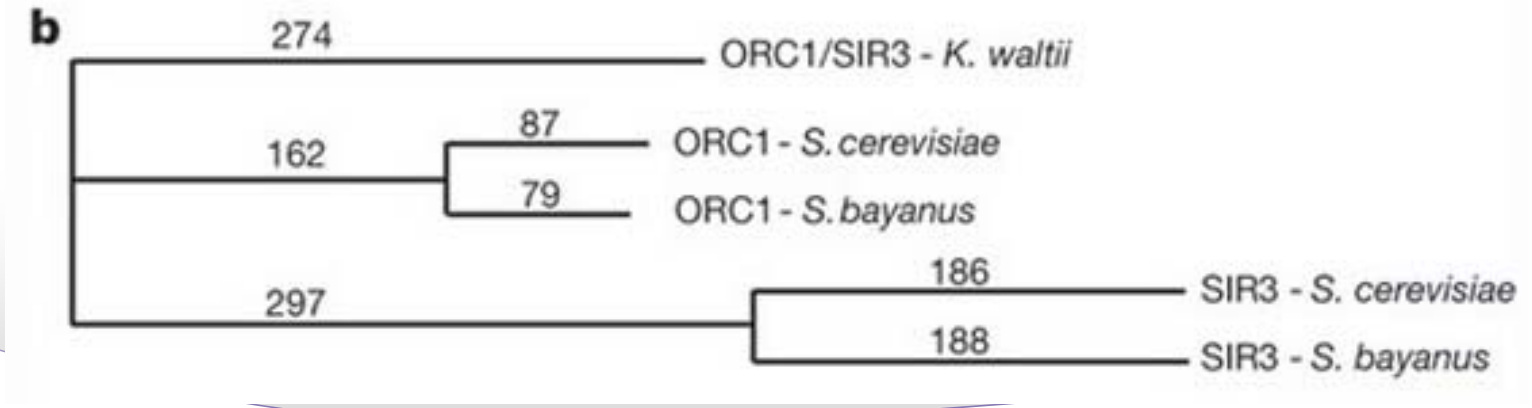
酵母的全基因组复制

- 基因数量的增加
  - 酵母~6000个基因，人类~21,000个基因
  - 单个基因复制、基因组复制、染色体片段复制
- 复制基因与已有基因的功能关系
  - “新功能形成”：Ohno one-gene-only speeds-up (OS) model，一个基因功能不变从而进化慢，另一个需要产生新功能从而进化快
  - “亚功能形成”：Both-genes speed-up (BS) model，两个基因都只保留原有基因的部分功能，因此进化速率都快

Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature. 2004 Apr 8;428(6983):617-24.

酵母的全基因组复制

- 复制基因分别的进化速率估算
  - 酵母属的两个种*S. cerevisiae* (酿酒酵母) 和*S. bayanus* (贝克酵母) ：由*K. waltii* (克鲁雄酵母)通过基因组复制后，分别进化形成
  - 克鲁雄酵母ORC1/SIR3：在酿酒酵母和贝克酵母中都有两个拷贝
  - OS模型：其中一个基因进化速率快
  - BS模型：两个基因进化速率都快



酵母的全基因组复制

- 作者鉴定了酵母中457对通过全基因组复制产生的复制基因对（总共914个基因）。在酿酒酵母中，其中76对有加速进化的现象。“加速进化”在文中的定义指的是酿酒酵母里氨基酸替代率要比克鲁雄酵母里高50%。在76对有加速进化的复制基因对里，其中只有4对是两个基因都加速进化。因此基因对里只有一个加速进化的为72个基因（72/76=95%）

- 问题：究竟应该怎样算*p*-value？

酵母的全基因组复制

- 统计模型
  - H_0为加速进化的基因随机成对，预期出现不少于4对加速进化
  - H_1为观察到4对加速进化
  - 457对复制基因共914个基因，其中72+4*2=80个基因存在加速进化，因此单个基因加速进化的概率=80/914=0.088
  - 一对基因同时加速进化的概率为0.088*0.088=0.0077
  - 考虑二项分布，总共457对，观察到4对加速进化
  - *p*-value=BINOMDIST(4,457,0.0077,TRUE) = 0.72

泊松分布

- 稀有事件发生的概率：在一个连续的时间或空间中，稀有离散变量出现的概率
- N -> ∞, E(x)=Variance=μ

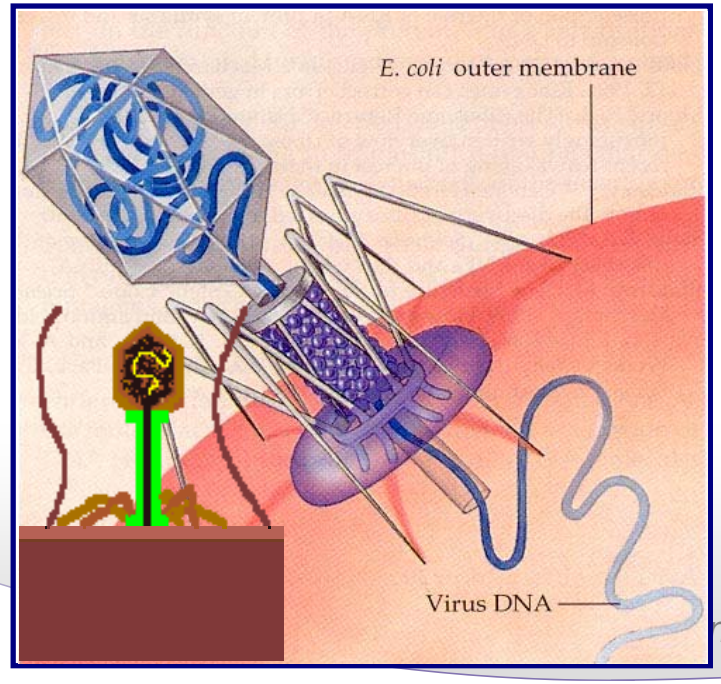
$$f(x)=\frac{e^{-\mu}(\mu)^{x}}{x!},x=0,1,2\ldots$$

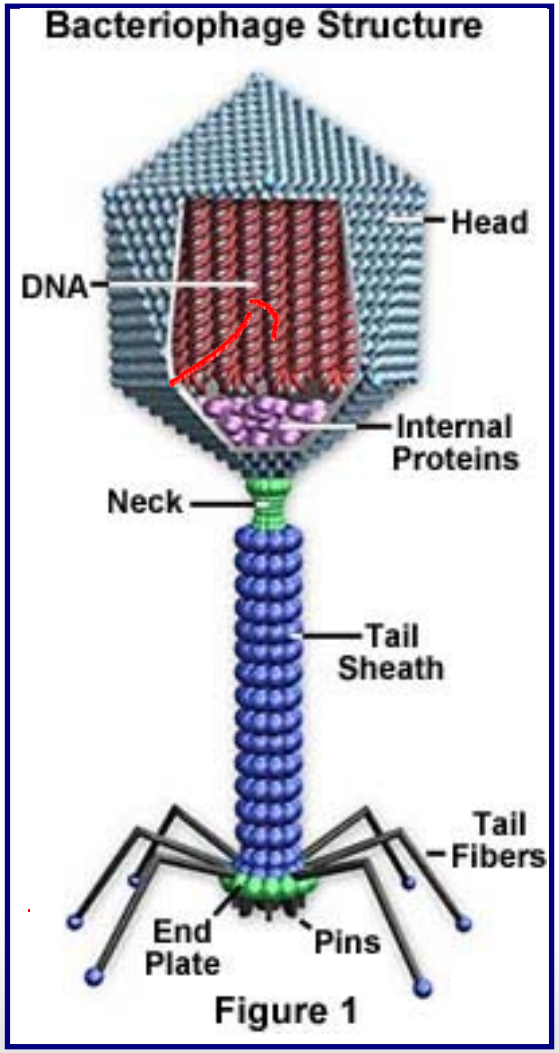
e = 2.71828…

方差等于均值

细菌 *vs.* 噬菌体


SCIENCEphotoLIBRARY


E. coli outer membrane
Virus DNA


Bacteriophage Structure
Head
DNA
Internal Proteins
Neck
Tail Sheath
Tail Fibers
End Plate
Pins
Figure 1

细菌对噬菌体的应答

- 数十亿细菌与噬菌体混合后，几乎所有的细菌将被杀死
- 仅有很少的细菌能够存活，生长成克隆，并且对噬菌体具有特异性抵抗能力
- 进化：细菌是否有基因？受到噬菌体攻击如何生存？
  - 拉马克机制：获得性遗传免疫 假说– 细菌在接触到噬菌体后，小概率产生抵抗，不需要基因或遗传物质
  - 孟德尔机制：突变假说

细菌生存的潜在机制

- 孟德尔 – 遗传变异
  - 细菌在噬菌体攻击之前已经具有抵抗能力，不需要与病毒相互作用，受到攻击时也不产生新的突变
- 拉马克 – 获得性遗传免疫
  - 细菌在受到攻击的时候才产生免疫能力

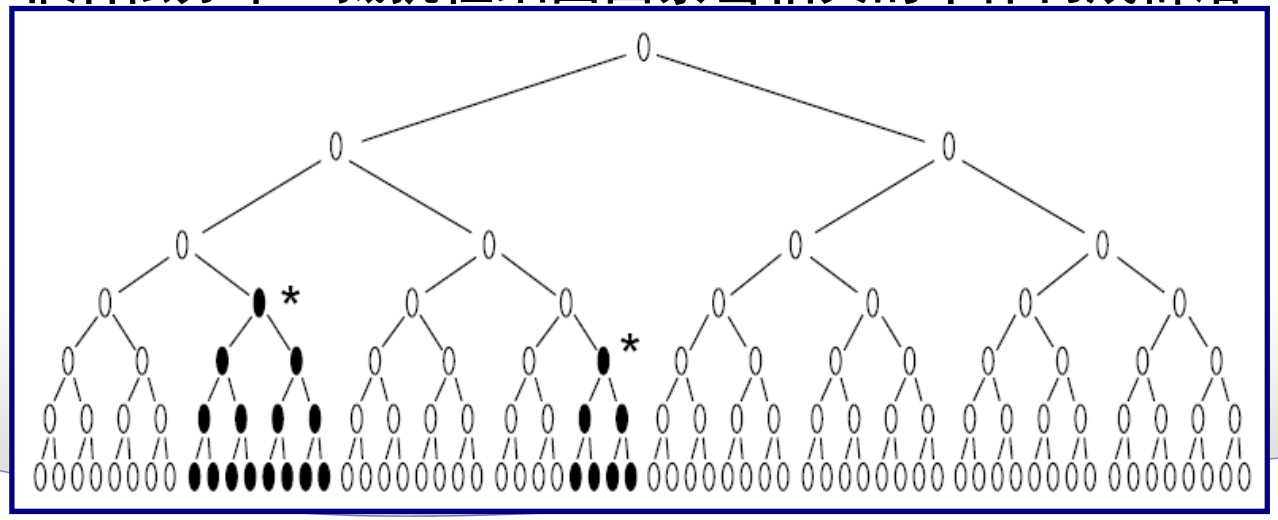
MUTATIONS OF BACTERIA FROM VIRUS SENSITIVITY TO VIRUS RESISTANCE[1,2]

S. E. LURIA[3] AND M. DELBRÜCK
Indiana University, Bloomington, Indiana, and Vanderbilt University, Nashville, Tennessee

Received May 29, 1943

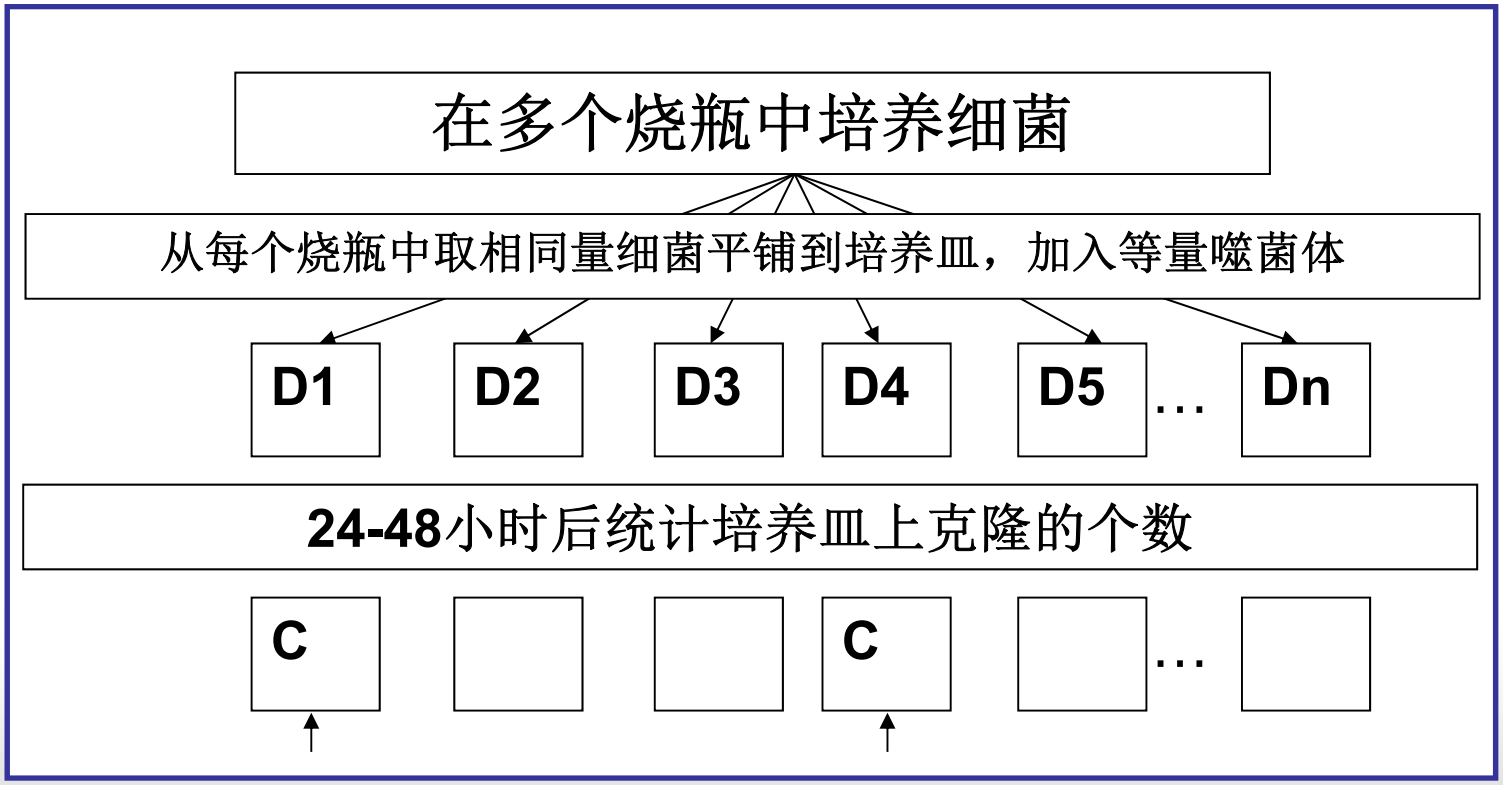
细菌生存的潜在机制

- 拉马克 – 获得性遗传免疫
  - 具有抵抗能力的细菌在受到攻击时的比例恒定
  - 泊松分布：每一个抵抗是一个独立的事件
  - 只有当与病毒接触时才产生免疫
- 孟德尔 – 遗传变异
  - 具有抵抗能力的细菌随时间比例增加
  - 非泊松分布：抵抗性细菌由紧密相关的个体构成群落



那种生存机制是正确的？

- 两类实验
  - 有抵抗力的细菌，比例是否随时间上升
  - 观察细菌克隆的个数，看抵抗是否与遗传突变相关



结果: 方差分析

❒ 相似培养条件下可抵抗细菌的克隆个数

Experiment No.	1	10	11	15	16	17	21a	21b
Number of Cultures	9	8	10	10	20	12	19	5
Volume of Cultures, cc	10.0	10.0	10.0	10.0	.2*	.2*	.2	10.0
Volume of Samples, cc	.05	.05	.05	.05	.08	.08	.05	.05
Culture No.								
1	10	29	30	6	1	1	0	38
2	18	41	10	5	0	0	0	28
3	125	17	40	10	3	0	0	35
4	10	20	45	8	0	7	0	107
5	14	31	183	24	0	0	8	13
6	27	30	12	13	5	303	1	
7	3	7	173	165	0	0	0	
8	17	17	23	15	5	0	1	
9			57	6	0	3	0	
10			51	10	6	48	15	
11					107	1	0	
12					0	4	0	
13					0		19	
14					0		0	
15					1		0	
16					0		17	
17					0		11	
18					64		0	
19					0		0	
20					33			
Average per sample	26.8	23.8	62	26.2	11.35	30	3.8	48.2
Variance (corrected for sampling)	1217	84	3498	2178	694	6620	40.8	1172
Average per culture	5360	4760	12400	5240	28.4	75	15.1	8440
Bacteria per culture	3.4×10^{10}	4×10^{10}	4×10^{10}	2.9×10^{10}	5.6×10^{8}	5×10^{8}	1.1×10^{8}	3.2×10^{10}
Mutation rate	1.8×10^{-8}	1.4×10^{-8}	4.1×10^{-8}	2.1×10^{-8}	1.1×10^{-8}	3.0×10^{-8}	3.3×10^{-8}	3.0×10^{-8}

将方差与均值进行比较

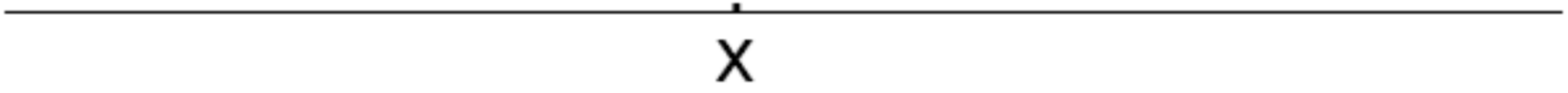
在每一个实验中，可抵抗细菌的波动（**fluctuation**）远比均值高，不能归因于采样误差，与获得性遗传免疫的假设冲突

Average per sample	26.8	23.8	62	26.2	11.35	30	3.8	48.2
Variance (corrected for sampling)	1217	84	3498	2178	694	6620	40.8	1172

例1：鸟枪法的覆盖率

Michael Waterman

- Lander-Waterman Model
- 近似符合泊松分布（Poisson distribution）
- 假设：需要测序的BAC长度200 kbp
  - 总共测序的序列数量：N
  - 每次测序：500 bp
  - 每次测序的覆盖率 p：500/200 kbp=0.0025
  - 因此：总覆盖率 C=Np（每个点平均覆盖到的次数）
- Y: 测序能够覆盖到点X的次数



鸟枪法：覆盖率

因此：点X被覆盖k次的概率：二项分布~泊松分布

$$P(Y=k) = (N!/(N-k)!k!)\, p^k(1-p)^{N-k} \quad \approx e^{-c}c^k / k!$$

当点X一次都不被覆盖时，k=0; 此时的概率为：

$$P(Y=0)= e^{-c}$$

覆盖率 *vs.* 准确性

Fold coverage	$P_0=e^{-c}$		P_0x100= % not sequence	% sequenced
0.25	$P_0=e^{-0.25}$	$=1/e^{0.25}=$	0.78	78%	22%
0.50	$P_0=e^{-0.50}$	$=1/e^{0.50}=$	0.61	61%	39%
0.75	$P_0=e^{-0.75}$	$=1/e^{0.75}=$	0.47	47%	53%
1	$P_0=e^{-1}=1/e^{1}=1/2.718=$		0.37	37%	63%
2	$P_0=e^{-2}=1/e^{2}=1/7.389=$		0.135	13.5%	87.5%
3	$P_0=e^{-3}=1/e^{3}=1/20.086=$		0.05	5%	95%
4	$P_0=e^{-4}=1/e^{4}=1/54.598=$		0.018	1.8%	98.2%
5	$P_0=e^{-5}=1/e^{5}=1/148.4=$		0.0067	0.6%	99.4%
6	$P_0=e^{-6}=1/e^{6}=1/403.4=$		0.0025	0.25%	99.75%
7	$P_0=e^{-7}=1/e^{7}=1/1096.6=$		0.0009	0.09%	99.91%
8	$P_0=e^{-8}=1/e^{8}=1/2980.95=$		0.0003	0.03%	99.97
9	$P_0=e^{-9}=1/e^{9}=1/8103.08=$		0.0001	0.01%	99.99%
10	$P_0=e^{-10}=1/e^{10}=1/22026.5=$		0.000045	0.005%	99.995%

泊松分布：例2

- 某种序列调控信号，在人类基因组上平均每500 kbp一个。随机给一条1 mbp的序列，在上面发现5个这样的信号，完全是随机产生的概率是多少？

- 本例中，N=3.0*10^9 bp $\rightarrow \infty$, E(x)=μ= 2 (1 mbp)

$$P(5) = f(5) = \frac{e^{-2}(2)5}{5!} = 0.036 < 0.05$$

- 统计性显著：*p*-value < 0.05

超几何分布

- 与二项式分布的区别：不放回抽样
- 例：有N个球，其中红球M个，白球N-M个，每次拿出一个球再放回，总共n次，其中有m个球是红球的概率为（二项式分布）：

$$P(m)=\binom{n}{m}p^{m}(1-p)^{n-m}$$

p=M/N

超几何分布

- 上例改为：有N个球，其中红球M个，白球N-M个，每次拿出一个球不放回，总共n次，其中有m个球是红球的概率为：

$$P(m)=\frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

并且，**0≤m≤M<N**

超几何分布

- 上例再改为：有N个球，其中红球M个，白球N-M个，每次拿出一个球不放回，总共n次，其中有至少有m个球是红球的概率为：

$$p-value = P(m' \geq m) = \sum_{m'=m}^{n} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}}$$

并且，**0≤m≤M<N**

超几何分布

- 上例再改为：有N个球，其中红球M个，白球N-M个，每次拿出一个球不放回，总共n次，其中有最多有m个球是红球的概率为：

$$p-value = P(m' \le m) = \sum_{m'=0}^{m} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}}$$

并且，**0≤m≤M<N**

超几何分布：例

- 研究者从26873个人类蛋白质中预测了2264个具有某种特定功能的底物，并进行进一步的分析。其中，有421个人类蛋白质具有某种功能结构域D，而在预测的2264个底物中，有94个蛋白质具有结构域D

- 问：结构域D在2264个底物中是显著出现，显著不出现，还是随机出现？

超几何分布：例 (2)

- N = 26873; n= 2264; M = 421; m = 94;
- (m/n)/(M/N) = 2.65
- 因此，问题转化：在26873个人的蛋白质中，抓出2264个蛋白质，其中至少有94个蛋白质具有功能结构域的概率是多少？

$$p-value = P(m' \geq m) = \sum_{m'=m}^{n} \frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}}$$

结果

命令提示符

```
C:\>hypergeometric.pl
N= 26873
n= 421
M= 2264
m= 94
This is Enrichment_ratio!
2.65024172632886
This is p-value!
1.15913702840128e-018

C:\>
```

统计显著性

- 考虑两个假设H_0（空假设）和H_1（备择假设）
  - H_0代表随机情况下事件出现的概率
  - H_1代表当前出现事件的概率
  - 如果$H_0/H_1 << 0.05$，则接受H_1而不接受H_0
- 统计显著：*p*-value < 0.05
- 超几何分布的*p*-value
  - “完全随机状态下”事件出现的概率，即*p*-value=H_0
  - H_1=1

Ronald Fisher

- 英国统计学家、进化生物学家、数学家、遗传学家和优生学家
- 数量遗传学的三个创始人之一
- 1918年批评孟德尔的数据过于完美
- Richard Dawkins：达尔文之后最伟大的生物学家

1890~1962

Fisher's Exact Test

- 超几何分布的精确概率计算：2X2表

	B1	B2	Totals
A1	a	b	a+b
A2	c	d	c+d
Totals	a+c	b+d	n

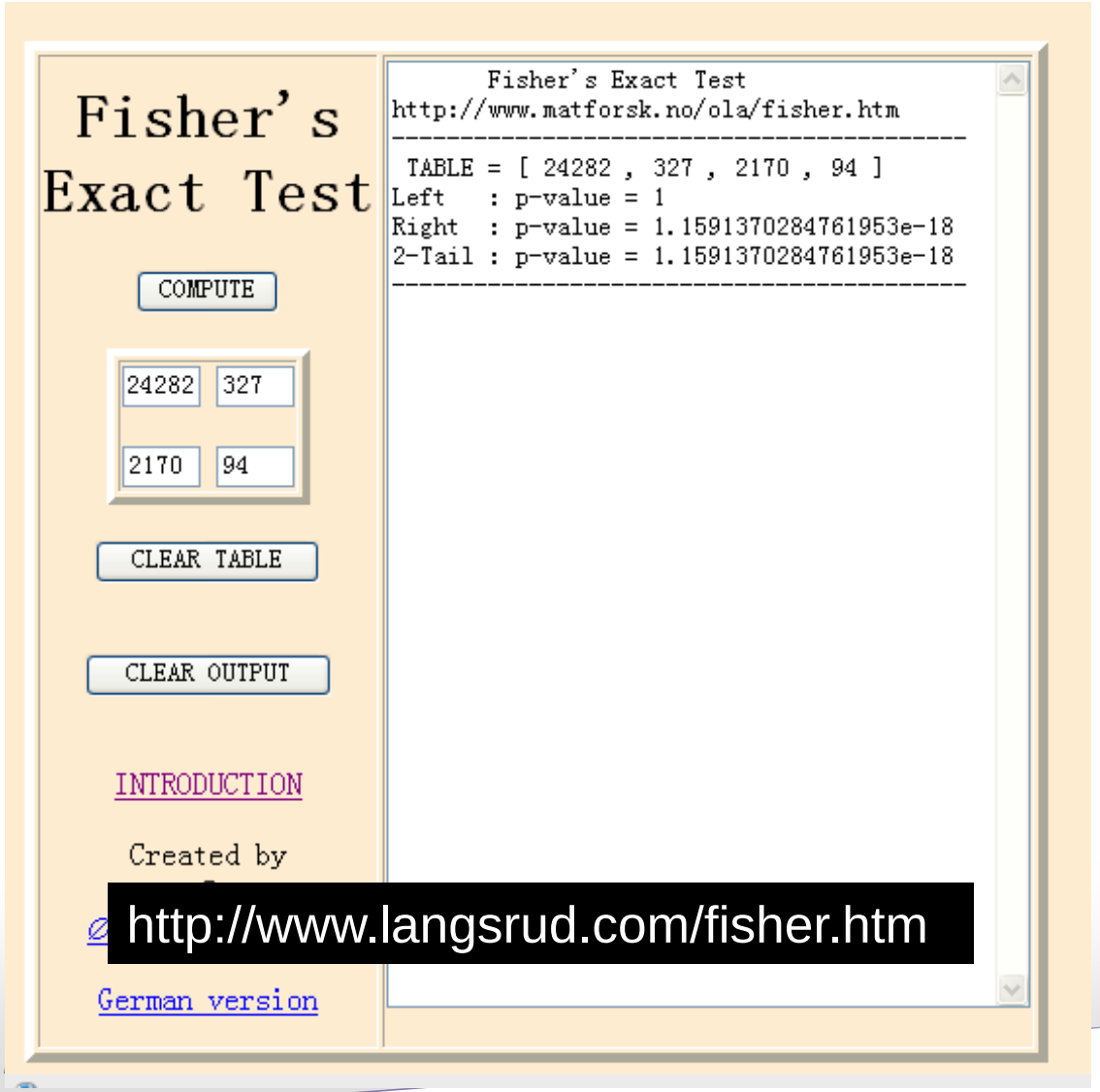
因此，超几何分布计算公式

$$p\text{-}value = \frac{\frac{(a+c)!}{a!c!} \times \frac{(b+d)!}{b!d!}}{\frac{n!}{(a+b)!(c+d)!}}$$

$$= \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

如上例

- a+b+c+d=26873,
- c+d=2264,
- b+d=421,
- d=94,
- 因此：



Fisher's Exact Test：再例

- 假设，我们调查了100个学生，比较是否男生比女生更喜欢玩电子游戏。数据统计如下：

	玩游戏	不玩游戏
男生	45	15
女生	27	13

***p*-value > 0.05**，统计性不显著！

随机序列模型

- 假设一个残基a随机出现的概率为q_a，并且该概率独立于其它残基而存在
- 则对于一段蛋白质或DNA序列：$x_1x_2...x_n$，整个序列出现的概率为：$q_{x1}q_{x2}...q_{xn}$:$\prod_{i=1}^{n} q_{xi}$

最大似然性估计

- 概率模型的参数通常是从大的可靠的数据集，即训练集中估算得到
- 例如：通过对Swissprot数据库分析，各个物种中，20种氨基酸出现的频率
- 估算的参数作为概率模型的参数，即最大似然性估计：充分使用了训练集的数据
- 一般的，给定一个模型，包括参数θ以及数据集D，则对于参数θ的最大似然性估计，要保证$P(D|\theta)$的最大化

几个主要真核物种中的氨基酸频率

AA	*S.cerevisiae*		*S.pombe*		*C.elegans*		*D.melanogaster*		*M.musculus*		*H.sapiens*	
	Num.	Per.	Num.	Per.	Num.	Per.	Num.	Per.	Num.	Per.	Num.	Per.
A	182589	5.51%	150066	6.24%	644995	6.37%	1018991	7.35%	1951767	6.90%	1917786	6.98%
C	43791	1.32%	35268	1.47%	204160	2.02%	274295	1.98%	646608	2.29%	613701	2.23%
D	189958	5.73%	128878	5.36%	542678	5.36%	714203	5.15%	1363975	4.82%	1291018	4.70%
E	213550	6.45%	156945	6.52%	669038	6.61%	880507	6.35%	1957893	6.92%	1913306	6.96%
F	149792	4.52%	110809	4.61%	476721	4.71%	484995	3.50%	1060957	3.75%	1014225	3.69%
G	165520	5.00%	118620	4.93%	541945	5.35%	849857	6.13%	1823069	6.45%	1805724	6.57%
H	71464	2.16%	54332	2.26%	234586	2.32%	372816	2.69%	737425	2.61%	725024	2.64%
I	217427	6.56%	147805	6.14%	617883	6.10%	678404	4.90%	1242781	4.39%	1207472	4.40%
K	240119	7.25%	154387	6.42%	642638	6.35%	778288	5.62%	1608966	5.69%	1527230	5.56%
L	316667	9.56%	237640	9.88%	872362	8.61%	1252315	9.04%	2835685	10.03%	2753451	10.02%
M	69484	2.10%	49557	2.06%	265730	2.62%	324098	2.34%	628623	2.22%	605750	2.20%
N	201584	6.08%	125243	5.21%	492995	4.87%	658568	4.75%	1013396	3.58%	985966	3.59%
P	145487	4.39%	113453	4.72%	497816	4.92%	765595	5.53%	1734018	6.13%	1712723	6.23%
Q	129461	3.91%	91663	3.81%	422211	4.17%	716329	5.17%	1339988	4.74%	1309438	4.77%
R	146367	4.42%	117272	4.87%	526718	5.20%	774601	5.59%	1583353	5.60%	1562613	5.69%
S	299056	9.03%	227040	9.44%	819366	8.09%	1158270	8.36%	2371524	8.38%	2270931	8.27%
T	197230	5.95%	132228	5.50%	594292	5.87%	794141	5.73%	1529233	5.41%	1505568	5.48%
V	185494	5.60%	145399	6.04%	630910	6.23%	815956	5.89%	1738580	6.15%	1636629	5.96%
W	35117	1.06%	26958	1.12%	111273	1.10%	140654	1.02%	343331	1.21%	362072	1.32%
Y	113063	3.41%	82252	3.42%	318131	3.14%	403645	2.91%	771962	2.73%	752485	2.74%
Total	3313220		2405815		10126448		13856528		28283134		27473112	

最大似然性的缺点

- 过训练（over-fitting）
- 例如：掷色子3次，得到 [6，6，6]，根据最大似然性的模型，则

$p_1=p_2=p_3=p_4=p_5=0$，$p_6=1$

条件、连接、边际的概率

- 考虑两个色子D_1和D_2
- 条件概率：用色子D_1掷出i的概率为$P(i|D_1)$;用色子D_2掷出i的概率为$P(i|D_2)$
- 连接概率：随机挑出一个色子的概率$P(D_j)$, j=1,2; 挑到第j色子且掷出一个i的概率（条件概率）为：$P(i,D_j)= P(D_j)P(i|D_j)$。一般定义为：
  - $P(X,Y) = P(X|Y)P(Y)$
- 边际概率：当条件或者连接概率已知的时候，可以计算边际概率并去掉一个变量：

$$P(X)=\sum_{Y}P(X,Y)=\sum_{Y}P(X\mid Y)P(Y)$$

故事及问题

- 某天，Prof. Gene来到拉斯维加斯去旅游，一时兴起，就去了一个赌场玩两把。游戏是掷色子。但是，据说这个赌场的荷官不老实，使用了两种色子，其中99%的色子是正常（fair）的，而1%的色子（loaded）则使得出现6的概率为50%
- 那么，$P(6|D_{loaded})$和$P(6|D_{fair})$是多少？而$P(6,D_{loaded})$和$P(6,D_{fair})$呢？随机拿到一个色子掷出6的概率是多少？

故事及问题

- 某天，Prof. Gene来到拉斯维加斯去旅游，一时兴起，就去了一个赌场玩两把。游戏是掷色子。但是，据说这个赌场的荷官不老实，使用了两种色子，其中99%的色子是正常（fair）的，而1%的色子（loaded）则使得出现6的概率为50%
- 那么，$P(6|D_{loaded})$和$P(6|D_{fair})$是多少？而$P(6,D_{loaded})$和$P(6,D_{fair})$呢？随机拿到一个色子掷出6的概率是多少？

Probability

- $P(6|D_{loaded})=0.5$
- $P(6|D_{fair})=1/6$
- $P(6,D_{loaded})=0.5*0.01=0.005$
- $P(6,D_{fair})=(1/6)*0.99$

- 随机拿到一个色子掷出6的概率:
- $P(6,D_{loaded}) + P(6,D_{fair})$

新问题

- Prof. Gene拿起一个色子，连续掷了三次，都是6，因此，他判断这个色子是loaded。他这样的判断可靠吗？如果不可靠，那么，怎样才能判断色子可能是loaded呢？

贝叶斯理论及模型比较

- 前向概率（prior probability）：
  - $P(D_{loaded})$=0.01 和$P(D_{fair})$=0.99
- 后向概率（posterior probability）：
  - $P(D_{loaded}|3个6)$
- 根据条件概率公式：
  - $P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$ =>

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

在本例中：

$$P(D_{loaded} \mid n个6) = \frac{P(n个6 \mid D_{loaded})P(D_{loaded})}{P(n个\times 6)}$$

其中，$n = 3$

结果：

$$P(D_{loaded} \mid 3个6) = \frac{(0.5^3)(0.01)}{(0.5^3)(0.01)+(\frac{1}{6})^3(0.99)} = 0.21$$

不能判断是否为**loaded**色子！**Prof. Gene**判断的不合理！

怎样才能认为是loaded色子？

- $P(D_{\text{loaded}}|n\text{个}6)$
- 四个6：$P=0.45$
- 五个6：$P=0.71>0.5$
- 当连续掷出5个6以上时，我们认为可能是loaded！
- …

例2：DNA序列模体

- 假设，基因组上存在一种未知的X-box的DNA序列（例如转录子结合位点、启动子或沉默子等），包含4个bp。Prof. Gene为了验证这种X-box序列，实验分析了1000个4 bp的DNA序列，他发现，其中100个4 bp的DNA序列为真实的、有功能的X-box序列。对这100个X-box的序列分析，他发现：

	第一位	第二位	第三位	第四位
A	70%	10%	1%	5%
T	10%	10%	97%	5%
C	10%	70%	1%	5%
G	10%	10%	1%	85%

预测：新的序列

- Prof. Gene拿到4条4 bp的序列：
  - ACTG
  - ATTT
  - AGTG
  - CCGA
- 计算预测这些序列是不是可能的X-box序列

对于给定4 bp的序列

$$P(X-box \mid X_1X_2X_3X_4) = \frac{P(X-box)\prod_i q_{xi}^{x-box}}{P(X-box)\prod_i q_{xi}^{x-box} + P(nonX-box)\prod_i q_{xi}^{nonX-box}}$$

其中：

P(X-box)=0.1

P(nonX-box)=0.9

$$q_{xi}^{nonX-box} = 0.25$$

对于ACTG序列

$$P(X-box \mid ACTG) = \frac{0.1*0.7*0.7*.097*0.85}{0.1*0.7*0.7*.097*0.85 + 0.9*(0.25)^4}$$

$$= 0.91$$

Perl编程：DNA模体的预测

```
my $DNA1="A";
my $DNA2="C";
my $DNA3="T";
my $DNA4="G";

my $Pp=0.01; my $Pn=0.99; my $Pn_all=0.99*0.25*0.25*0.25*.025;

if ($DNA1 eq "A")     { $Pp=$Pp*0.7;}
elsif ($DNA1 eq "T") { $Pp=$Pp*0.1;}
elsif ($DNA1 eq "C") { $Pp=$Pp*0.1;}
elsif ($DNA1 eq "G") { $Pp=$Pp*0.1;}

if ($DNA2 eq "A")     { $Pp=$Pp*0.1;}
elsif ($DNA2 eq "T") { $Pp=$Pp*0.1;}
elsif ($DNA2 eq "C") { $Pp=$Pp*0.7;}
elsif ($DNA2 eq "G") { $Pp=$Pp*0.1;}

if ($DNA3 eq "A")     { $Pp=$Pp*0.01;}
elsif ($DNA3 eq "T") { $Pp=$Pp*0.97;}
elsif ($DNA3 eq "C") { $Pp=$Pp*0.01;}
elsif ($DNA3 eq "G") { $Pp=$Pp*0.01;}

if ($DNA4 eq "A")     { $Pp=$Pp*0.05;}
elsif ($DNA4 eq "T") { $Pp=$Pp*0.05;}
elsif ($DNA4 eq "C") { $Pp=$Pp*0.05;}
elsif ($DNA4 eq "G") { $Pp=$Pp*0.85;}

my $P=$Pp/($Pp+$Pn_all);
print $P,"\n";
```

预测结果！

- P(X-box|ACTG)=0.91！
- *P(X-box|ATTT)=0.08*
- P(X-box|AGTG)=0.60
- *P(X-box|CCGA)=0.0009!*

Homework 2#: 蕾丝短裤之谜


豆瓣小组 精选 文化 行摄 娱乐 时尚 生活 科技

在男友家发现一条不是自己的内裤。。。

来自: 野樱桃(我们是一只蚂蚁) 2013-10-15 11:01:29

有一个大抽屉是专门放我的衣物的 里面原来遗留他的一两件衣服 我也没检查过
那天打开抽屉惊现一条内裤不是我的 问他他说就是我的。。。。
其实我是很相信他的
问了我身边的姐们 她们说肯定是我的 我容易迷糊肯定忘记了
但是那真的不是我的啊。。。。
这问题一直没解决 然后就这样了
总觉得有个结
也不知道怎么办


博文

CY呼唤肖子：蕾丝短裤之谜 精选

已有 2042 次阅读 2014-2-24 16:47 | 个人分类:课件科普 | 系统分类:教学心得

在 求教：贝叶斯定理（乳腺癌例） 评论4，王春艳 说：“我都回来了，肖子还端着不下来。很感兴趣老邪的问题，可惜手头上没书，以后也弄本翻翻，要是肖子能赏脸耗力帮大家把问题整理出来讨论，那是要非常感谢滴”。

感谢肖子，及时解答了老邪的疑问。希望响应CY妹妹的呼唤，撺着回答老邪的疑问：怎么样讲贝叶斯定理，最容易被同学理解无误。我想，就从这个乳腺癌例讲起。Silver取4个用历史先验案例表达的概率，和4个导出的边际概率，分别为：

	真有癌	无癌	总计	边际概率
阳性：	11	99	110	P(阳)
阴性：	3	887	890	P(阴)
总计：	14	986	1000	
边际概率：	P(有)	P(无)		

注意这里，人们习惯是说真阳性、假阴性表示真有癌。但如果按真假排列两列，边际概率就便成对角线的和了。折衷表

李小文
解除好友 给我留言
打个招呼 发送消息

作者的精选博文
- 贝叶斯定理：走桃花运和遭桃
- 费希尔对贝叶斯的批判和后者
- 蕾丝短裤之谜--揭晓
- 围怨：男生不宜
- 关于国家重点实验室开放基金

作者的其他最新博文 全部

- ❒ 李小文院士博文：发现男友衣柜里有一条蕾丝内裤！男友出轨了吗？
- ❒ 先验概率：
  - ❁ P(出)=0.04　　来自国家统计局数据
  - ❁ P(裤|出)=0.5　　来自对男友细心程度的估计
  - ❁ P(裤|未)=0.05　　来自对男友各种可能辩护合理性的估计