生物信息学基础讲座

第三讲：美丽的数学与实用的数学

1. 常用数学公式的英语读法和意义

符号	英语读法	数学意义
逻辑符号	**Logic**	
$\exists$	there exists	存在
$\forall$	for any	对于任意值
$p \Rightarrow q$	p implies q / if p, then q	由 p 可以推导得到 q
$p \Leftrightarrow q$	p if and only if q / p is equivalent to q	p 等价于 q
集合	**Sets**	
$x \in A$	x belongs to A / x is an element of A	x 是集合 A 中的元素
$x \notin A$	x does not belong to A	x 不是集合 A 的元素
$A \subset B$	A is contained in B / A is a subset of B	A 是 B 的子集
$A \supset B$	A contains B / B is a subset of A	A 是 B 的父集
$A \cap B$	A cap B / A meet B / A intersect B	A 和 B 的交集
$A \cup B$	A cup B / A join B / A union B	A 和 B 的并集
$A \setminus B$	A minus B / the difference between A and B	A 和 B 的差集
$A \times B$	A cross B / the Cartesian product of A and B	A 和 B 的笛卡尔积
实数	**Real numbers**	
$x+1$	x plus one	
$x-1$	x minus one	
$x \pm 1$	x plus or minus one	
xy	xy / x multiplied by y	
$(x+y)(x-y)$	x plus y, x minus y	
$\frac{x}{y}$	x over y	
=	the equals sign	
$x = 5$	x equals 5 / x is equal to 5	
$x \neq 5$	x (is) not equal to 5	
$x \equiv y$	x is equivalent to (or identical with) y	
$x > y$	x is greater than y	
$x \geq y$	x is greater than or equal to y	
$x < y$	x is less than y	
$x \leq y$	x is less than or equal to y	
$0 < x < 1$	zero is less than x is less than 1	
$0 \leq x \leq 1$	zero is less than or equal to x is less than or equal to 1	
$\lvert x \rvert$	mod x / modulus x	
x^2	x squared / x (raised) to the power 2	

x^3	x cubed	
x^4	x to the fourth / x to the power four	
x^n	x to the nth / x to the power n	
x^{-n}	x to the (power) minus n	
$\sqrt{x}$	(square) root x / the square root of x	
$\sqrt[3]{x}$	cube root (of) x	
$\sqrt[4]{x}$	fourth root (of) x	
$\sqrt[n]{x}$	nth root of x	
$(x+y)^2$	x plus y all squared	
$\left(\frac{x}{y}\right)^2$	x over y all squared	
$n!$	n factorial	
$\hat{x}$	x hat	
$\overline{x}$	x bar	
$\tilde{x}$	x tilde	
x_i	xi / x subscript I / x suffix I / x sub i	
$\sum_{i=1}^{n} a_i$	the sum from i equal one to a_i / the sum as i runs from 1 to n of the a_i	
线性代数	**Linear algebra**	
$\|\| x \|\|$	the norm (or modulus) of x	向量 x 的范数
$\overrightarrow{OA}$	OA / vector OA	向量 OA
$\overline{OA}$	OA / the length of the segment OA	线段 OA 的长度
A^T	A transpose / the transpose of A	矩阵 A 的转秩
A^{-1}	A inverse	矩阵 A 的逆矩阵
函数	**Functions**	
$f(x)$	fx / f of x / the function f of x	x 的函数 f
$f: S \to T$	a function f from S to T	函数 f 是从集合 S 到 T 的映射
$x \mapsto y$	x maps to y / x is sent (or mapped) to y	x 映射到 y
$f'(x)$	f prime x / f dash x / the first derivative of f with respect to x	f 对 x 的一阶导数
$f''(x)$	f double-prime x / f double-dash x / the second derivative of f with respect to x	f 对 x 的二阶导数
$f'''(x)$	f triple-prime x / f triple-dash x / the third derivative of f with respect to x	f 对 x 的三阶导数
$f^{(4)}(x)$	f four x / the fourth derivative of f with respect to x	f 对 x 的四阶导数
$\frac{\partial f}{\partial x_1}$	the partial (derivative) of f with respect to x1	f 对 x1 的一阶偏导数
	the second partial (derivative) of f with	

$\frac{\partial^2 f}{\partial x_1^2}$	respect to x1	f 对 x1 的二阶偏导数
$\int_0^\infty$	the integral from zero to infinity	对某函数从零到无穷大积分
$\lim_{x \to 0}$	the limit as x approaches zero	x 逼近 0 时的极限
$\lim_{x \to +0}$	the limit as x approaches zero from above	x 从上逼近 0 时的极限
$\lim_{x \to -0}$	the limit as x approaches zero from below	
$\log_e y$	log y to the base e / log to the base e of y / natural log (of) y	y 的自然对数
$\ln y$	log y to the base e / log to the base e of y / natural log (of) y	

2. 微积分（Calculus）

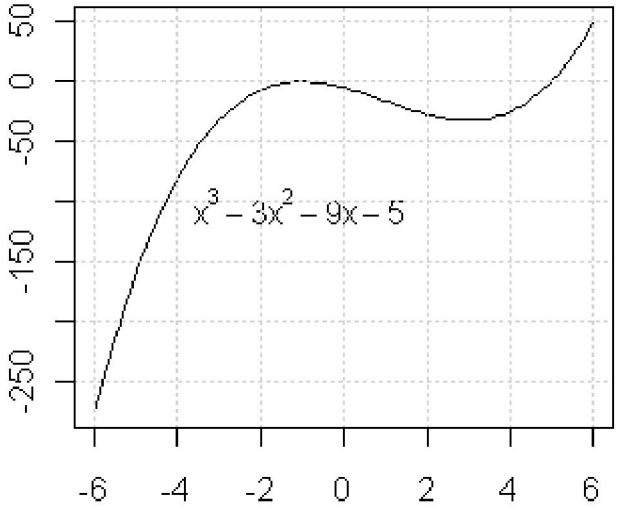
Calculus, which originally means a small stone used for counting, is a branch in mathematics focused on **limits**, **functions**, **derivatives**, **integrals**, and **infinite series**. It has two major branches, **differential calculus** and **integral calculus**, which are related by the fundamental theorem of calculus. Calculus is the study of change, in the same way that geometry is the study of shape and algebra is the study of operations and their applications to solving equations. A course in calculus is a gateway to other, more advanced courses in mathematics devoted to the study of functions and limits, broadly called mathematical analysis.

Goal:

(1) use linear, polynomial, rational, algebraic, exponential and logarithmic functions in applications
(2) determine the limits of functions graphically, numerically and analytically
(3) recognize and determine infinite limits and limits at infinity
(4) determine the continuity of functions at a point or on intervals
(5) understand the interpretation of the derivative as the slope of a line tangent to a graph and as the rate of a dependent variable with respect to an independent variable, and determine the derivative of a function using the limit definitions.
(6) Use differentials in approximation problems
(7) Determine derivatives using the power rule, sum & difference rules, product rule, quotient rule, and chain rule
(8) Determine derivatives of exponential and logarithmic functions
(9) Determine higher order derivatives of a function
(10) Understand velocity as the derivative of position and acceleration as the 2nd derivative of position
(11) Determine the absolute extrema of a continuous function on a closed interval.

(12)Use the first and 2nd derivatives to analyze and sketch the graph of a function, including determining intervals on which the graph is increasing, decreasing, constant, concave up, concave down, and finding relative extrema and infection point

(13)Apply differential calculus to application.



函数的定义

- → $f : A \to B$
- → Spoken: “f is a function from A to B”, or “f maps A to B”
- → 意义：A/B 为集合，集合 A 称为域（domain），而集合 B 则称为上域（codomain）或靶（target）

基本概念： 极限（limit）、连续（continuity）和可微（differentiability）

导数（derivative）：一阶导数（first-order derivative）、二阶导数（second-order derivatives）、高阶导数（higher-order derivatives）、偏导数（partial derivatives）

积分（integral）：左端点（left endpoint）,右端点（ right endpoint），梯形法（trapezoid Rule），中点法（midpoint rules），数值积分（numerical integration），辛普森法（Simpson’s rule），Accuracy of integration rules（积分方法的准确性），双重积分（Double integration）

数列（sequence）与级数（series），最大与最小值（maxima and minima）、收敛（convergence）、傅立叶级数（Fourier series）

3. 线性代数（Linear Algebra）

3.1 向量（Vector）的计算（加减、内积、外积）、范数（norm）

3.2 矩阵（Matrix）的运算（特别是乘法）、范数（norm）、行列式（determinant）、秩（rank）、逆矩阵（inverse matrix）、单位矩阵（identity matrix）、正交矩阵（orthogonal matrix）、对称矩阵（symmetric matrix）、正定矩阵（positive definite

matrix）、正半定矩阵（positive semi-definite）、上三角矩阵（upper triangular matrix）、下三角矩阵（lower triangular matrix）、对角矩阵（diagonal matrix）

3.3 矩阵分解（matrix decomposition）：LU 分解（LU-decomposition）、Cholesky 分解、奇异值分解（singular value decomposition，SVD）、QR 分解（QR-decomposition）、Schur 分解（Schur decomposition）

3.4 特征值（Eigenvalues）和特征向量（eigenvectors）

3.5 多项式（polynomials）和样条插值（Splines interpolation）：三次样条（cubic splines）、最小二乘插值（least squares interpolation）

矩阵分解在 microarray 数据分析中应用实例分析

奇异值分解（singular value decomposition, SVD）

Let $X \in \mathbb{R}^{m \times n}$ with rank r, $m \geq n \geq r$. In the case of microarray, x_{ij} is the expression level of the i[th] gene in the j[th] assay.

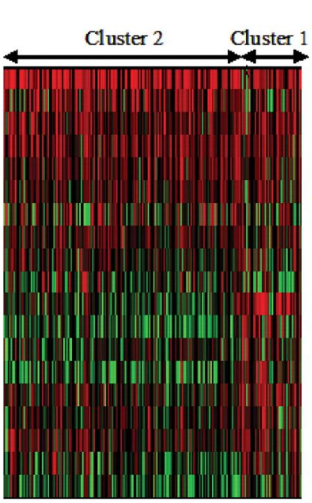
The singular value decomposition of X is as following:

$$X = USV^T$$

where $U \in R^{m \times n}$, $S \in R^{n \times n}$ is a diagonal matrix, and $V \in R^{n \times n}$. The columns of U are called the left singular vectors, $\{u_i\}$, which form an orthonormal basis for the assay expression profiles so that

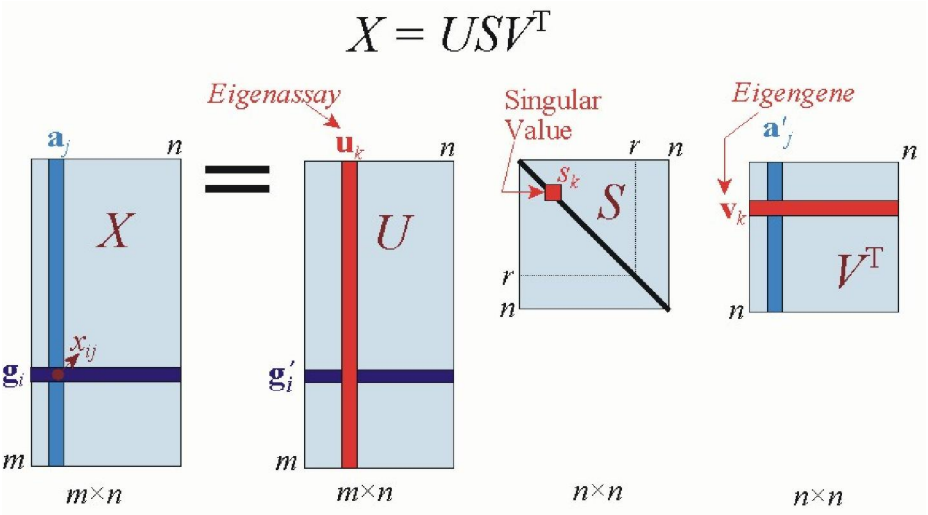
$$u_i \cdot u_j = \begin{cases} 1, & \text{when } i = j \\ 0, & \text{otherwise} \end{cases}$$

The columns of V contain the elements of the right singular vectors, $\{v_i\}$, which form an orthonormal basis for the gene transcriptional responses. The elements of S are only nonzero on the diagonal, and are called the singular values. Thus, $S = \text{diag}(s_1, s_2, ..., s_n)$. Furthermore, $s_k > 0$ for $k \in [1, r]$, and $s_k = 0$ for $k \in [r+1, n]$



Row labels

Row	Sample	Group
1	ALL_20185_B-cell	ALL_B2
2	ALL_21302_B-cell	ALL_B2
3	ALL_17281_B-cell	ALL_B2
4	ALL_19183_B-cell	ALL_B2
5	ALL_28373_B-cell	ALL_B2
6	ALL_5982_B-cell	ALL_B2
7	ALL_11103_B-cell	ALL_B2
8	ALL_20414_B-cell	ALL_B2
9	ALL_17929_B-cell	ALL_B1
10	ALL_549_B-cell	ALL_B1
11	ALL_7092_B-cell	ALL_B1
12	ALL_18239_B-cell	ALL_B1
13	ALL_9335_B-cell	ALL_B1
14	ALL_14749_B-cell	ALL_B1
15	ALL_R11_B-cell	ALL_B1
16	ALL_23953_B-cell	ALL_B1
17	ALL_19769_B-cell	ALL_B1
18	ALL_9692_B-cell	ALL_B1
19	ALL_R23_B-cell	ALL_B1



SVD application in microarray data analysis

As we mentioned above, the right singular vectors span the space of the gene transcriptional responses $\{g_i\}$ and the singular vectors span the space of the assay expression profiles $\{a_j\}$. And the left singular vectors $\{u_k\}$ are referred as eigenasays and the right singular vectors $\{v_k\}$ as eigengenes.

In systems biology applications, we wish to understand relations among genes. The signal of interest in this case is the gene transcriptional responses, $\{g_i\}$, and

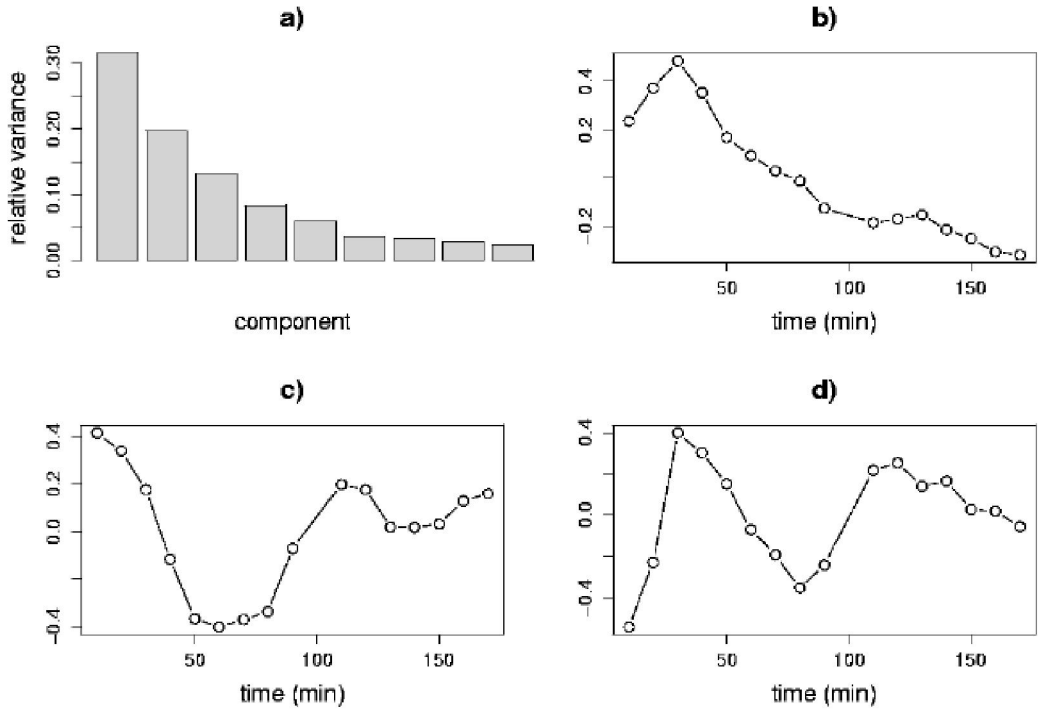
$$g_i = \sum_{k=1}^{r} u_{ik} s_k v_k$$

which is a linear combination of the eigengenes $\{v_k\}$. The i[th] row of U, g_i', contains the coordinates of the i[th] gene in the coordinate system (basis) of the scaled eigengenes, $s_k v_k$. If $r < n$, the transcriptional responses of the genes may be captured with fewer variables using g_i' rather than g_i. This property of the SVD is sometimes referred to as dimensionality reduction.

In diagnostic applications, we may wish to classify tissue samples from individuals with and without a disease. In this case, the signal of interest is the assay expression profile a_j, which is reducing the number of the variable for interpretation of the assay expression profiles.

$$a_j = \sum_{k=1}^{r} v_{jk} s_k u_k$$

Which is a linear combination of the eigenassays $\{u_k\}$



Visualization of the SVD of cell cycle data

4. 微分方程（Differential Equation）

(1) 线性和非线性一阶常微分方程（linear and nonlinear first order ODEs）

(2) 具有常系数的高阶常微分方程 (higher order ODEs with constant coefficients)

(3) 柯西方程（Cauchy's equation）和欧拉方程（Euler's equation）

(4) 拉普拉斯变换（Laplace transforms）

(5) 偏微分方程（partial differential equation）

(6) 随机微分方程（stochastic differential equation）

A differential equation is an equation relating a function to one or more of its derivatives. An initial value problem is a differential equation

$$\frac{dx}{dt} = f(t, x)$$

Where the initial condition, $x(t_0) = x_0$, is specified.

1) How to rigorously define a differential equation or system of differential equations?

2) Feasible solution or the solution is unique?

3) How to find the unique or non-unique solutions?

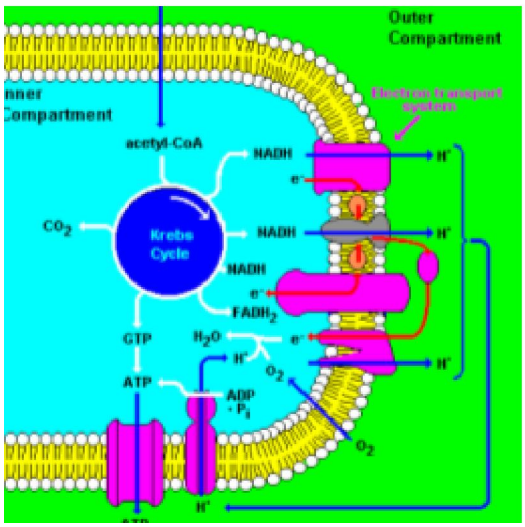
Existence and Uniqueness of Solutions

Theorem 1 If

$$y'+p(t)y=g(t)$$

is a differential equation such that $y(t_0)=y_0$, and $p(t)$ and $g(t)$ are continuous on the open interval $I=(\alpha,\beta)$, then there exists a unique function $y=\phi(t)$ satisfying the differential equation and the initial condition on I.

微分方程在系统生物学中的应用



$$\frac{dx_1}{dt} = k_{11}x_1 + k_{21}x_2 + k_{31}x_3 + \ldots$$

$$\frac{dx_2}{dt} = k_{12}x_1 + k_{22}x_2 + k_{32}x_3 + \ldots$$

$$\frac{dx_3}{dt} = k_{13}x_1 + k_{23}x_2 + k_{33}x_3 + \ldots$$

$$\frac{dx_4}{dt} = k_{13}x_1 + k_{24}x_2 + k_{34}x_3 + \ldots$$

5. 数值计算和最优化理论（Numeric Analysis & Optimization）

(1) 线性和非线性代数方程的解法（Solution of linear and nonlinear algebraic equations）

(2) 梯形积分（integration of trapezoidal）和辛普森法（Simpson's rule）

(3) 微分方程的单步和多步法（single and multistep methods）

Numerical methods, as an efficient supplement to analytical methods, are playing a central role in bioinformatics data analysis and algorithmic development. In this course you should learn to differentiate numerical methods from analytical methods.

5.1 Finding the roots for $f(x)=0$

①. Bisection method: based on the continuity of the function $f(x)$

②. Newton method: based on approximating the function by tangent lines

③. Secant method: based on approximating the function by secant lines

5.2 Convex optimization

①. 单纯形法（simplex method）

②. 线性规划（linear programming）

③. 二次规划（quadratic programming）

④. Optimization problem without constraints

⑤. Optimization problem with equality constraints

⑥. Optimization problem with inequality constraints

⑦. 动态规划（dynamic programming）

a) Shortest path problem

b) Best pairwise sequence alignment

6. 概率（Probability）

概率理论的所有认识论价值在于：大规模随机现象的集体行动产生严格的、非随机规律。也就是说，随机性源于人们对现实世界的不完全认识，源于对准确预测味蕾世界所需信息的匮乏；随机性还起源于复杂性。近年来，科学已经抛弃了 Laplace 的确定性思维，完全接受了解释随机性和发明足够的工具来描述其任务。例如，轮盘赌和掷骰子的机理明显是相同的，人们可以预测所有的可能结果不随时间变化——虽然每一次的单个结果是随机的。

6.1 概念（Concepts）

A probability of an event A: $\Pr(A) \in [0,1]$

The probability of the **complement event** of A: $\mathrm{P}(\bar{A}) = 1 - \mathrm{P}(A)$

The **joint probability** or **intersection** of two independent events A and B:

$$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$$

The probability of the **union** of two **mutually exclusive events** A and B:

$$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$$

If the events are not exclusive then

$$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

The probability of A given B (**conditional probability**): $\mathrm{P}(A \mid B) = \dfrac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$, thus

$$\mathrm{P}(A \cap B) = \mathrm{P}(A \mid B)\,\mathrm{P}(B)$$

Bayesian theorem: $\mathrm{P}(A \mid B) = \dfrac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \dfrac{\mathrm{P}(B \mid A)\,\mathrm{P}(A)}{\mathrm{P}(B \mid A)\,\mathrm{P}(A) + \mathrm{P}(B \mid \bar{A})\,\mathrm{P}(\bar{A})}$

6.2 概率分布模型（Probability distributions）

6.2.1 离散概率分布（discrete probability distribution）

(1) 二项式分布（binomial distribution）

进行 n 次 success/failure 贝努立实验（Bernoulli experiment）或贝努立试验（Bernoulli trial）得到 x 次 success 的概率。当 n=1 时，二项式分布也称为贝努立分布（Bernoulli distribution）。

Notation: $X \sim B(n,p)$, where $n \in N^+$ represents the number of trials, and $p \in [0,1]$ is the success probability in each trial.
Possible outcomes: $x \in \{0,\ldots,n\}$

Probability density function (pdf): $f(x;n;p) = \Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

Cumulative density function (cdf):

$$F(x;n;p) = \Pr(X \le x) = \sum_{i=1}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$

$$= I_{1-p}(n-x, x+1) = (n-x)\binom{n}{x}\int_0^{1-p} t^{n-x-1}(1-t)^x dt$$

For $x \le np$, we can derive the upper bounds for the lower tail of the distribution function:

霍夫丁不等式（**Hoeffding's inequality**）： $F(x;n,p) \le \exp\left(-2\frac{(np-x)^2}{n}\right)$

切诺夫不等式（**Chernoff's inequality**）： $F(x;n,p) \le \exp\left(-2\frac{(np-x)^2}{n}\right)$

Mean: $\hat{\mu} = np$

Variance: $\hat{\sigma}^2 = np(1-p)$

Skewness: $\frac{1-2p}{\sqrt{np(1-p)}}$

Kurtosis: $\frac{1-6p(1-p)}{np(1-p)}$

- If $X \sim B(n,p)$ and $Y \sim B(m,p)$ are independent binomial variables, then $X+Y$ is also a binomial variable: $X+Y \sim B(n+m,p)$;
- When n=1, $X \sim B(1,p) \equiv X \sim Bern(p)$;
- When n is large enough, and p is not near to 0 or 1, then an excellent approximation to $B(n,p)$ is given by normal distribution: $N(np, np(1-p))$;

- As n is sufficiently large while the product p is sufficiently small, another approximation to $B(n,p)$ is Poisson distribution with parameter $\lambda = np$;
- As n approaches infinity while p remains fixed, the distribution of $\frac{X - np}{\sqrt{np(1-p)}}$ approaches the standard normal distribution.

(2) 超几何分布（hypergeometric distribution）

超几何分布（hypergeometric distribution）是从一个有限样本中连续 n 次无放回（without replacement）取样成功 X 次的概率；粗看跟二项式分布相似，但二项式分布是有放回的取样（with replacement），也就是说每次取样成功的概率完全相同，在实际中这是不存在的，除非样本无穷大。

随机变量 $X \sim Hypergeometric(N,m,n)$

概率密度函数 $f(x;N,m,n) = \mathrm{P}(X = x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{m}}$

均值 $\mu = \frac{nm}{N}$

方差 $\sigma^2 = \frac{nm(N-n)(N-m)}{N^2(N-1)}$

偏度（skewness）为 $\frac{(N-2m)(N-1)^{1/2}(N-2n)}{[nm(N-m)(N-n)]^{1/2}(N-2)}$

峭度（kurtosis）为…

Let $X \sim Hypergeometric(N,m,n)$ and $p = m/N$

- If n=1 then $X \sim Bern(p)$;
- Let $Y \sim B(n,p)$; If N and M are large compared to n and p is not close to 0 or 1, then X and Y have similar distributions, i.e., $\mathrm{P}(X \le k) \approx \mathrm{P}(Y \le k)$;
- If n is large, N and m are large compared to n and p is not close to 0 or 1, then $\mathrm{P}(X \le k) \approx \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right)$ (standard normal distribution function).

(3) Poisson 分布

The Poisson distribution (or Poisson law of large numbers) is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time (or in other specified intervals such as distance, area or volume) if these events occur with a known average rate and independently of the time since the last event.

If the expected number of occurrences in the interval is λ, then the probability that there are exactly n occurrences ($n = 0,1,2,\ldots$) is equal to

$$f(n;\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

(4) 负二项式分布（Negative binomial distribution）

(5) 几何分布（Geometric distribution）

6.2.2 连续概率分布（continuous probability distribution）

(1) 均匀分布（Uniform distribution）

(2) 正态分布（Normal distribution）

(3) F 分布

(4) 卡方分布（Chi-square distribution）

(5) 指数分布（exponential distribution）

(6) Gamma 分布

(7) 拉普拉斯分布（Laplace distribution）

(8) 对数正态分布（Lognormal distribution）

(9) 学生 t 分布（student's t distribution）

(10) 威布尔分布（Weibull distribution）

6.3 概率分布的模拟（Simulation of random probability distribution）

在数据分析中我们常常需要对概率分布进行模拟，有时候是为了说明真实数据与对应随机数据之间某个统计值之间的差异性是否显著，有时候是为了用广谱数据说明我们设计算法的合理性和有效性。

6.4 马尔科夫链（Markov Chain）

马尔科夫链，也称为离散时间有限马尔科夫链（discrete-time finite Markov chain）或离散时间有限马尔科夫模型（discrete-time finite Markov model），简称为马尔科夫链或马尔科夫模型（Markov model）。考虑一个有限的集合 X，

N 个可能的**状态（states）** $G_1,...,G_N$。在每个时刻 $t=1,2,...$ 一条马尔科夫链只能处于某一状态。从时刻 t 到时刻 $t+1$，状态可在集合 X 内部按照固定的概率转换，时刻 $t+1$ 的状态 G_j 仅依赖于时刻 t 的状态 G_i，而跟 t 无关，也就是说概率 $p_{ij}=\Pr\left(G_j \mid G_i\right), i,j=1,...,N$ 称为马尔科夫链的**状态转换概率**，并可以写成矩阵 $P\in\mathbb{R}^{N\times N}$ 的形式

$$P=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{pmatrix}$$

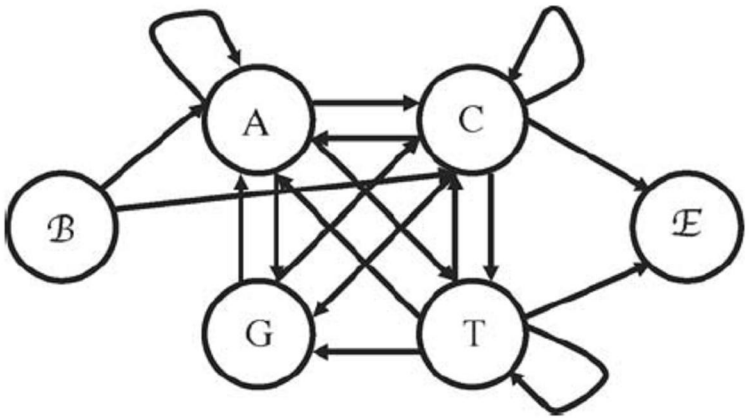
状态转换概率矩阵（matrix of transition probabilities）P 中的元素 p_{ij} 表示从状态 i 到状态 j 的概率，因此 $\forall i\in X, \sum_{j\in X} p_{ij}=1$。

原核基因去掉起始和终止密码子后的开放阅读框（Open Reading Frame，ORF）的例子

$x1$: **ATG***CTATTGATT***TAA**
$x2$: **GTG***AAAGACTTC***TAA**
$x3$: **ATG***CCCGATGAACGC***TAG**
$x4$: **ATG***AAGCATGAT***TAA**

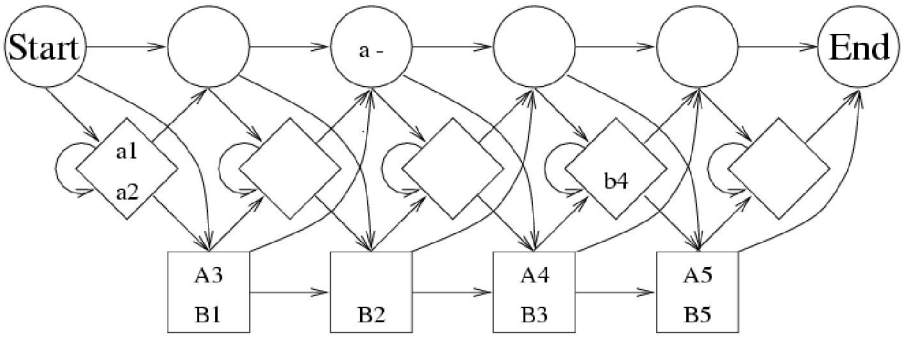
由以上序列可得到下面的状态转换矩阵和马尔科夫模型示意图。

	A	C	G	T	0
0	1/2	1/2	0	0	0
A	4/13	2/13	2/13	5/13	0
C	1/9	2/9	2/9	2/9	2/9
G	5/7	2/7	0	0	0
T	1/10	1/10	3/10	3/10	1/5



6.5 隐马尔科夫模型（Hidden Markov Model）

A hidden Markov model is a Markov model where the rules for producing the chain are unknown or “hidden”. The rules include two probabilities: ① that there will be a certain observation and ② that there will be a certain state transition, given the state of the model at a certain time.



The hidden Markov Model (HMM) method is a mathematical approach to solving certain types of problems:

(1) Given the model, find the probability of the observations;

(2) Given the model and the observations, find the most likely state transition trajectory;

(3) Maximize either (1) or (2) by adjusting the model parameters.

The corresponding solutions include the following algorithms:

(1) Forward-backward algorithm

(2) Viterbi algorithm

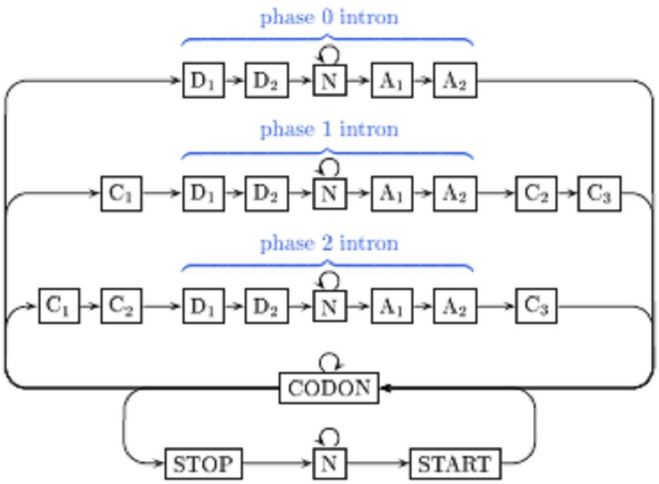
(3) Baum-Welch algorithm

The HMM method is generally used in pattern recognition problems, anywhere there may be a model producing a sequence of observations. It is widely used in bioinformatics including sequence alignment, gene prediction, structure prediction and other data mining approaches.

Applications

(1) Finding multiple sequence patterns, e.g., motif, conserved domain

(2) Gene prediction, including the gene structures



Architecture of Eukaryotic EHMM

Pedersen J.S., Hein J. **Gene finding with a hidden Markov model of genome structure and evolution** (2003) *Bioinformatics*, 19 (2), pp. 219-227.

7. 统计学（Statistics）

7.1 Descriptive statistics of continuous data

(1) Location: mean (arithmetic, geometric, harmonic), median

(2) Dispersion: standard deviation (sd), coefficient of variation, percentile

(3) Shape: absolute deviation, variance, semivariance, skewness, kurtosis, moments, L-moments

7.2 Descriptive statistics of categorical data

(1) Frequency, contingency table

7.3 Statistical graphics

(1) Barplot, biplot, boxplot, correlogram, forest plot, histogram, Q-Q plot, stemplot, scatterplot

7.4 Exploratory data analysis

(1) Inference: Confidence Interval (CI for frequentist), credible interval (CI for Bayesian)

(2) Experiment design: Case-control study, cohort study, observation study, perspective study, replication, sensitivity and specificity

(3) Sample size determination: statistical power, effect size, standard error

7.5 Parameter estimation

(1) Bayesian estimator, maximum likelihood estimator (MLE), method of moments, minimum distance

7.6 Hypothesis testing

(1) Z-test (normal), student's t-test (paired or unpaired), F-test, Chi-square test, Pearson's Chi-square test, Wald test, Mann-Whitney test, Shapiro-Wilk test, Fisher's exact test, Wilcoxon signed-rank test

7.7 Analysis of variance

7.8 Correlation analysis

(1) Pearson product-moment correlation, rank correlation (Spearman's rho, Kendall's tau), partial correlation, confounding variable

7.9 Regression analysis

(1) Linear regression: simple linear regression, ordinary least squares, general linear model, analysis of variance, analysis of covariance

(2) Non-standard: nonlinear regression, nonparametric, semiparametric, robust regression

(3) Non-normal errors: Generalized linear model (GLM), binomial regression, Poisson regression, Logistic regression

7.10 Multivariate data analysis

(1) Multivariate regression, principal components regression, factor analysis, cluster analysis, copulas

7.11 Survival analysis

(1) Survival function, Kaplan-Meier, Logrank test, Failure rate, Proportional Hazards model, Accelerated failure time model

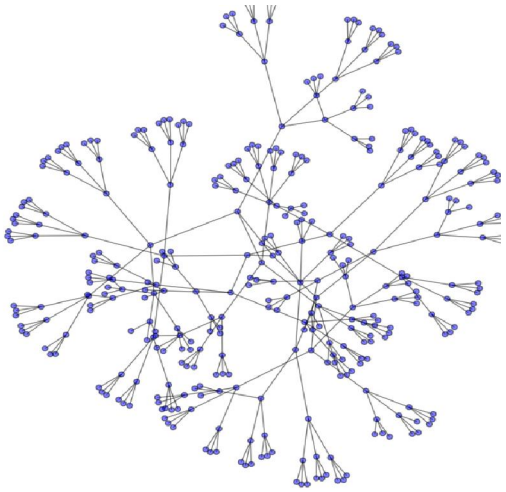
7.12 Time series analysis

(1) Decomposition, trend estimation, Box-Jenkins, ARMA models, spectral density estimation

8. 图论（Graph Theory）

As shown above, a **graph** is a pair of sets (**V**, **E**), where

- ✓ **V** is a nonempty set whose elements are called **vertices**.
- ✓ **E** is a collection of two-element subsets of **V** called **edges**.



The vertices correspond to the dots in the above network, and the edges correspond to the lines. Thus, the dots-and-lines diagram above is a pictorial representation of the graph (V, E)

8.1 Topological properties of a network

（1） node connectivity，also called degree of the node

（2） clustering coefficient

（3） the shortest path

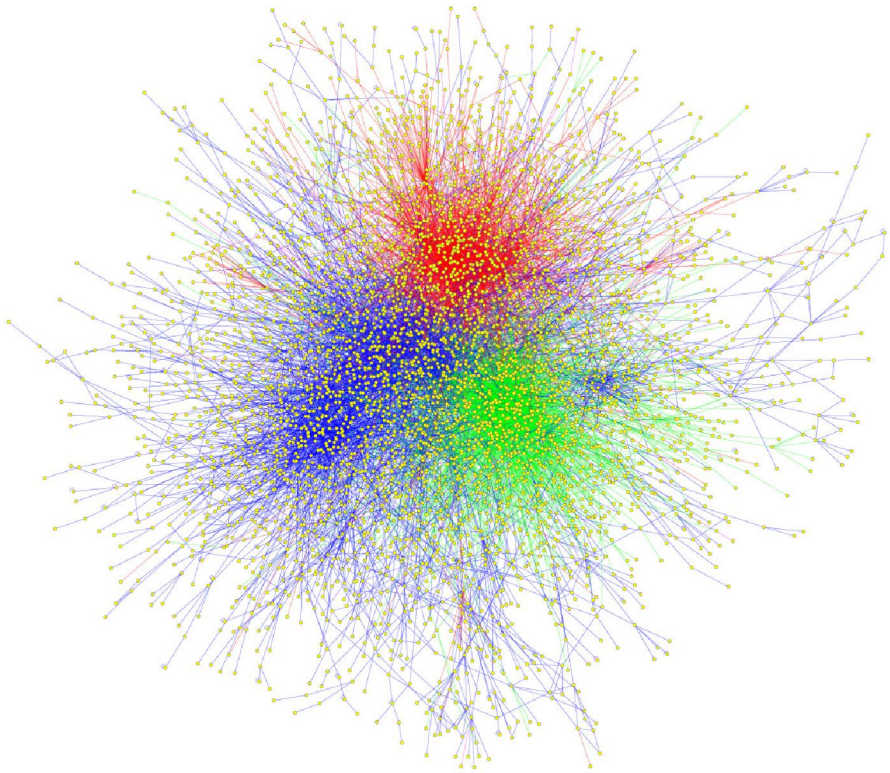
（4） Motif discovery

（5） Module identification

8.2 Dynamic properties of a graph

（1） Flux balance analysis (FBA)

（2） ...



人蛋白相互作用网络图

上面蛋白相互作用网络图中，蛋白表示为黄色的节点，来自 CCSB-HI1（Rual *et al.*, *Nature* 2005, **437:**1173–1178）和（Stelzl *et al.*, *Cell* 2005, **122:**957–968）的相互作用分别用红色和绿色的边表示。从数据库提取得到的具有相关文献支持的相互作用（LCI）则用蓝色边表示。三种来源数据同时支持的相互作用用黑色边表示，Rual 和 Stelzl 重叠的用黄色边表示，Rual 和 LCI 用紫红色表示，Stelzl 和 LCI 则用青色表示，后面几种颜色都是复合色（RGB 三原色）。

9. 课后练习

A. 利用 $\frac{de^x}{dx} = e^x$ 和 $\lim_{x\to 0} e^x = 1$，写出 e^x 的幂级数展开。

B. 将边分别为 **v** 和 **w** 的平行四边形的面积用点积（dot product）的形式表示。

注意：$\cos\theta = \frac{\langle v, w\rangle}{|v||w|}$，其中 θ 是向量 **v** 和 **w** 的夹角。

C. There are three commonly used studied metrics for the set $\mathbb{R}^N$.

i. Euclidean distance: $d(x,y) = (\sum_{i=1}^{N}(x_i - y_i)^2)^{1/2}$

ii. Manhattan distance: $D(x,y) = \sum_{i=1}^{N}|x_i - y_i|$

iii. Chebyshev distance: $d_{\infty}(x,y)=\max(|x_i-y_i|)$

iv. The more general Minkowski distance: $d_p=\left(\sum_{i=1}^{N}|x_i-y_i|^p\right)^{1/p}$

Graph each of the following on Cartesian coordinate systems

v. $A=\left\{x\in \mathrm{R}^2 : d(0,x)\le 1\right\}$

vi. $A=\left\{x\in \mathrm{R}^2 : D(0,x)\le 1\right\}$

vii. $A=\left\{x\in \mathrm{R}^2 : d_{\infty}(0,x)\le 1\right\}$

D. 什么是汉明距离（Hamming distance）？曼哈顿距离（Manhattan distance）？最大组分距离（maximum component distance）？欧氏距离（Euclidean distance）？明可夫斯基距离（Minkowski distance）？

E. 如果微分方程的形式是$\frac{dy}{dx}=axy$，其中 a 为常数，请问其可能的通解是什么？对于$\frac{dy}{dx}=\frac{ax}{y}$呢？

F. Use calculus to find the absolute maximum and minimum, if either exists for the function $f(x)=2x^3-3x^2-12x+24$ on these three intervals:

a) [-3, 4]

b) [-2, 3]

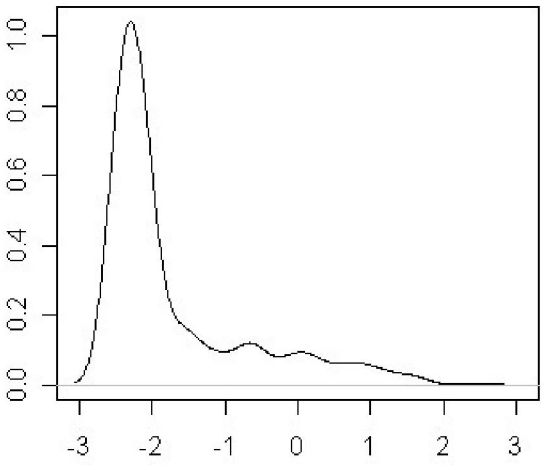
c) [-2, 1]

G. 请将蛋白相互作用网络写成邻接矩阵（adjacency matrix）的形式，并根据该矩阵计算度的分布情况。

H. 请问 $f(x)=\sqrt{x^2+x}$ 是凸函数（convex function）还是凹函数（concave function）？可以就不同区间进行说明

I. 请说明方程组 feasible solution 和 optimal solution 的联系和区别。

J. 试标出下面分布图中中位值（median）和均值（mean）的相对位置。



K. What is the negation of the statement "$\forall y > 0, \exists c$ that $f(c) = y$"?

L. If T: $\mathrm{R}^3 \to \mathrm{R}^3$ is a linear transformation for which

$$T\begin{bmatrix}1\\0\\2\end{bmatrix} = \begin{bmatrix}1\\1\\1\end{bmatrix},\ T\begin{bmatrix}1\\1\\0\end{bmatrix} = \begin{bmatrix}2\\1\\0\end{bmatrix},$$

then $T\begin{bmatrix}3\\2\\2\end{bmatrix} = ?$

M. Calculate the angle between the vectors $\begin{bmatrix}3\\1\end{bmatrix}$ and $\begin{bmatrix}-6\\2\end{bmatrix}$

N. The base composition of a certain microbial genome is $p_G = p_C = 0.3$ and $p_A = p_T = 0.2$. We are interested in the di-nucleotides where the letters are assumed to be independent. There are 4 × 4=16 dinucleotides.

i. Present these 16 probabilities in a table. Do your table sum to 1.0?

ii. Purine bases are defined by $R = \{A, G\}$ and pyrimidine bases by $Y = \{C, T\}$. Let E be the event that the first letter is a pyrimidine, and F the event that the second letter is A or C or T. Find $\mathrm{P}(E)$, $\mathrm{P}(F)$, $\mathrm{P}(E \cup F)$, $\mathrm{P}(E \cap F)$ and $\mathrm{P}(F^c)$.

iii. Set $G = \{CA, CC\}$. Calculate $\mathrm{P}(G \mid E)$, $\mathrm{P}(F \mid G \cup E)$, $\mathrm{P}(F \cup G \mid E)$.

O. Here is the transition matrix P for a first-order Markov Chain with four states $\{A, C, G, T\}$. Find the stationary distribution of the chain; that is, solve the equations $\pi = \pi P$ subject to the elements of π begin positive and summing to 1.

	A	C	G	T
A	0.423	0.151	0.168	0.258
C	0.399	0.184	0.063	0.354
G	0.314	0.189	0.176	0.321
T	0.258	0.138	0.187	0.415

(初稿，2010-4-9，待完善)

Examples of applications of optimization in systems biology, classified by type of optimization problem (note that several types overlap)

Problem type or application	Description	Examples with references
Linear programming (LP)	linear objective and constraints	maximal possible yield of a fermentation [83]; metabolic flux balancing [18,83]; review of flux balance analysis in [30]; use of LP with genome scale models reviewed in [27]; inference of regulatory networks [40,42]
Nonlinear programming (NLP)	some of the constraints or the objective function are nonlinear	applications to metabolic engineering and parameter estimation in pathways [69]; substrate metabolism in cardiomyocytes using ^{13}C data [84]; analysis of energy metabolism [85]
Semidefinite programming (SDP)	problems over symmetric positive semidefinite matrix variables with linear cost function and linear constraints	partitioning the parameter space of a model into feasible and infeasible regions [86]
Bilevel optimization (BLO)	objective subject to constraints which arise from solving an inner optimization problem	framework for identifying gene knockout strategies [87]; optimization of metabolic pathways under stability considerations [88]; optimal profiles of genetic alterations in metabolic engineering [89]
Mixed integer linear programming (MILP)	linear problem with both discrete and continuous decision variables	finding all alternate optima in metabolic networks [90,91]; optimal intervention strategies for designing strains with enhanced capabilities [91]; framework for finding biological network topologies [47]; inferring gene regulatory networks [41]
Mixed integer nonlinear programming (MINLP)	nonlinear problem with both discrete and continuous decision variables	analysis and design of metabolic reaction networks and their regulatory architecture [92,93]; inference of regulatory interactions using time-course DNA microarray expression data [45]
Parameter estimation	model calibration minimizing differences between predicted and experimental values	tutorial focused in systems biology [53]; parameter estimation using global and hybrid methods [52,54,55,59,70]; parameter estimation in stochastic models [58]
Dynamic optimization (DO)	Optimization with differential equations as constraints (and possible time-dependent decision variables)	discovery of biological network design strategies [94]; dynamic flux balance analysis [29]; optimal control for modification of self-organized dynamics [95]; optimal experimental design [66]
Mixed-integer dynamic optimization (MIDO)	Optimization with differential equations as constraints and both discrete and continuous decision variables (possibly time-dependent)	computational design of genetic circuits [76]