Health Informatics Lab@USTC

生物信息学算法

李骜

中国科学技术大学信息学院

生物医学工程研究中心

email: aoli@ustc.edu.cn

- 学习工作经历
  - 本科：科大生物系
  - 博士：科大信息学院
  - 博士后：耶鲁大学
- 研究方向：
  - 生物信息学的统计建模和算法
  - 生物医学信息处理
  - 机器学习与数据挖掘
  - 人机交互

学生自我介绍

- 姓名
- 专业背景
- 研究方向
- 选课目的
  - 为什么选课？希望从中学到什么？
- 其他

Health Informatics Lab@USTC

课程面向对象

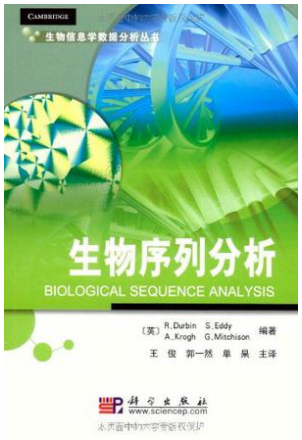
- 研究生/本科高年级学生：
  - 背景：信息、生物、计算机、数学、物理
  - 适合生物信息学领域的初学者
- 目标：
  - 算法中核心的理论原理和技术方法
  - 以算法为手段解决实际问题的思路和技巧
  - 开展生物信息学的科研工作

课程介绍

- 特点：
  - 强调生物信息学中的原理
  - 着重分析问题解决方法和思路
  - 介绍生物信息学研究中的实际例子
- 模式：
  - 从一个生物信息学的实际问题引出
  - 介绍相应的解决思路
  - 系统阐述智能信息处理的理论
  - 分析扩展问题
  - 该技术在其他问题中的应用

课程资料

- 教材：



- 其他资料：
  - 部分内容来源于网络公共资源及其他相关的课件资料，特此致谢！

选课要求

- 预修课程：
  - 高等数学
  - 概率论与数理统计
- 基本的编程能力：
  - Matlab
  - C/C++
  - Java
  - Perl
  - Python
  - R

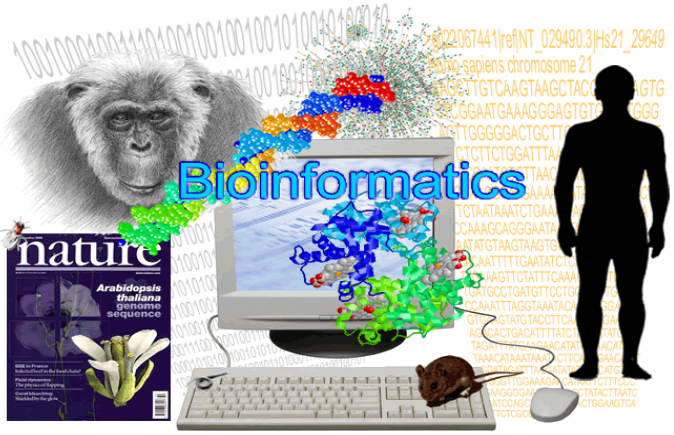
Health Informatics Lab@USTC

课程成绩

- 课程设计（60%）
  - 应用学到的算法知识，解决实际的生物信息学问题（蛋白质磷酸化预测、微阵列数据分析等）
  - 算法实现、程序编写、数据处理、结果分析
- 课程报告（40%）
  - 结合自己的工作内容或兴趣爱好，选择生物信息学算法研究方向，并调研、整理相关科研文献
    - 在课堂上介绍该方向的发展历史，现有方法的技术路线，未来的研究思路
    - 详细介绍一种算法在生物信息学中的应用
  - 时间安排：报告20-30分钟，提问5分钟

Health Informatics Lab@USTC

第一章
生物信息学介绍



生物信息学发展背景

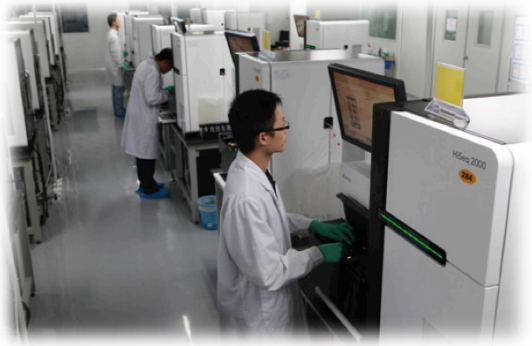
- 人类基因组计划(Human Genome Project, HGP)：
  - 由美国NIH和能源部提出和带头，美、英、德、法、日、中共同参与的国际合作项目
  - 重大国际研究项目：测定人类基因组全部DNA序列，构建人类基因组遗传图谱和物理图谱
  - 1990-2003年： 30亿美元
- 超过70个其他生物基因组图谱已经测序

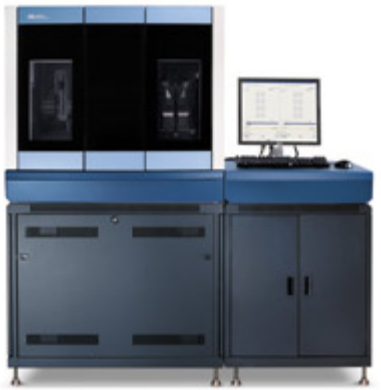
生物信息学发展背景

- 下一代测序（next-generation sequencing）技术

生物信息学发展背景

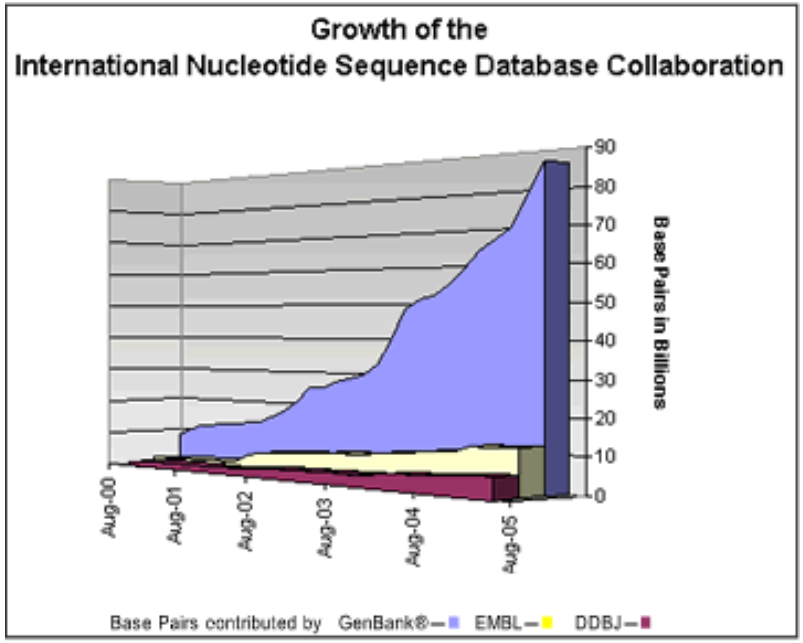
- 高通量生物实验技术
  - 微阵列技术（microarray）：
    - 基因表达微阵列：>4x10^{4}个转录本/样品
    - DNA微阵列：>10^{6}个单核苷酸多态性位点/样品
  - 下一代测序技术（Next-gen sequencing）：
    - SOLiD系统：> 3x10^{11}个碱基 ~= 80GB/样品

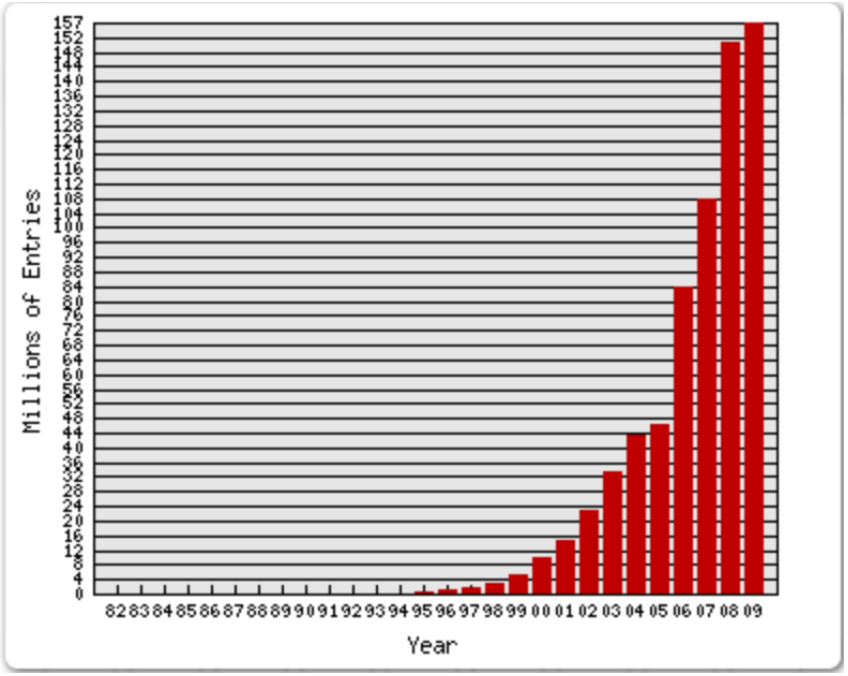




生物信息学发展背景

- 生物数据库增长情况





Health Informatics Lab@USTC

生物信息学发展背景

- 海量生物数据的迅速膨胀
  - DNA、RNA和蛋白质序列数据
  - 蛋白质结构数据
  - 基因功能数据
  - 蛋白功能数据
  - 蛋白质相互作用数据
  - 基因调控数据
  - 微阵列数据
  - 质谱数据
  - …

生物信息学发展背景

- 后基因组时代的“组”(-omics)学：
  - 功能基因组学
  - 功能蛋白质组学
  - 分子进化基因组学
  - 药物基因组学
  - 肿瘤蛋白质组学
  - 比较基因组学
  - 代谢组学
  - 转录组学
  - …

生物信息学的定义

- “21世纪的生物学正在从一门纯粹的实验科学转化成为一门信息科学”
  - 对大量生物数据的管理、分析和信息化需求促进了生物信息学的迅速发展
    - “生物信息学是一门交叉学科，它包含了生物信息的获取、处理、存储、分发、分析和解释等在内的所有方面，它综合运用数学、计算机科学和生物学的各种工具，来阐明和理解大量数据所包含的生物学意义。” –《人类基因组计划总结报告》
    - “生物信息学是一门以生物数据为研究对象，以信息、数学、物理等多学科的理论方法为研究手段，以算法、软件、数据库为研究工具的新兴交叉学科，主要的研究内容包括管理数量庞大且类型不同的生物数据，并对其进行有效的处理和整合；同时，通过对生物数据的深入挖掘和分析，从而最终阐释生命的奥秘。” –《教学大纲》

Health Informatics Lab@USTC

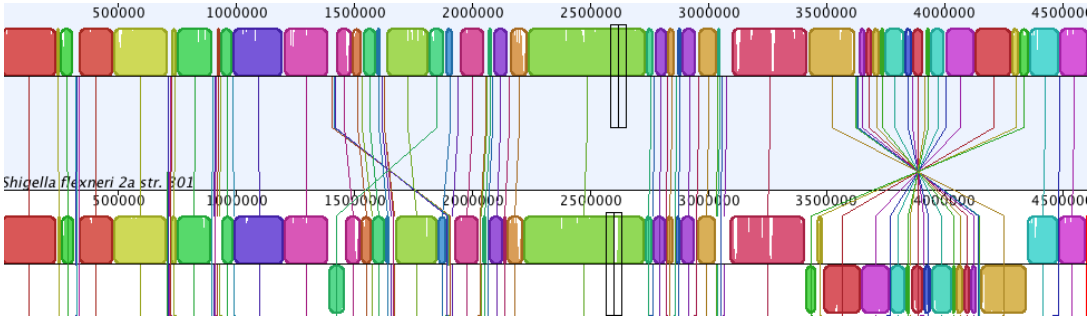
生物信息学的重要意义

- 二十世纪初，考古学家在克里特岛上发现了写有奇怪文字（线形文字B）的黏土板
- 之后50年中被认为是一种当地的“米诺斯”语言，Michael Ventris 花费了17年时间破译了这种语言
- 其实这是希腊语用另外一种字母表书写的

生物信息学的重要意义

- 对比的策略：

在进化存在联系的两种昆虫



生物信息学的主要研究内容

- 生物信息的存储与查询
- 序列比对
- 基因预测及基因组分析
- 分子进化与系统发育分析
- RNA结构预测
- 蛋白质结构、功能预测
- 分子设计与药物设计
- 生物芯片、测序
- 生物网络
- …

信息技术与生物信息学

- 通过公共的网络数据库，存储、检索这些信息的生物信息学技术已经趋于成熟
  - NCBI，USCS，…
- 分析、理解越来越多的海量生物信息
  - 借助信息技术
  - 生物信息学模型、算法、软件和工具

Health Informatics Lab@USTC

智能信息处理

- 智能信息处理是信息和计算机科学中的前沿交叉学科，其目标是处理海量和复杂信息，研究新的、先进的理论和技术
  - 信息和知识处理的理论
  - 复杂系统的算法设计和分析
  - 并行处理理论与算法
  - 机器学习理论和算法
  - 量子计算和生物计算
  - 生物信息学和系统生物学

生物信息学研究的一般步骤

- 1. 确立研究的生物学体系。例如：生物芯片数据分析；蛋白质结构与功能预测
- 2. 确定研究的问题。已有哪些计算方面的工作？是否需要实验的支持？
- 3. 构建生物学/数学模型，
- 4. 计算方法的选择或开发
- 5. 计算结果分析，与同类工具做比较。构建相应的数据库/软件/在线网站等
- 6. 扩展及应用：有哪些用处？

Health Informatics Lab@USTC

生物信息学研究的成功关键

- 研究方向：
  - 新颖
  - 重要的生物学意义
  - 发展前景良好，尚未或刚开始有人研究
- 研究思路/技术手段：
  - 提出新算法
    - 借鉴其他领域的成果
  - 现有方法的不足
    - 性能差
    - 假设不成立
    - 生物模型不完善
- 实验验证（follow-up experiments）

生物信息学的相关杂志

生物信息学相关期刊名称	网址
Bioinformatics	http://bioinformatics.oxfordjournals.org/
BMC Bioinformatics	http://www.biomedcentral.com/bmcbioinformatics/
Genome Biology	http://genomebiology.com/
Genome Research	http://www.genome.org/
Nucleic Acids Research	http://nar.oxfordjournals.org/
Briefings in Bioinformatics	http://www.henrystewart.com/briefings_in_bioinformatics/
FEBs letters	http://www.febsletters.org/
Biochemical and Biophysical Research Communications	http://www.sciencedirect.com/science/journal/0006291X
Molecular Systems Biology	http://www.nature.com/msb/index.html
Molecular Biology and Evolution	http://mbe.oxfordjournals.org/
PLoS Computational Biology	http://www.ploscompbiol.org/
PLoS ONE	http://www.plosone.org/
Protein Science	http://www.proteinscience.org/
Proteins	http://www3.interscience.wiley.com/cgi-bin/jhome/36176
Protein Engineering Design and Selection	http://peds.oxfordjournals.org/

Health Informatics Lab@USTC

小强读研记

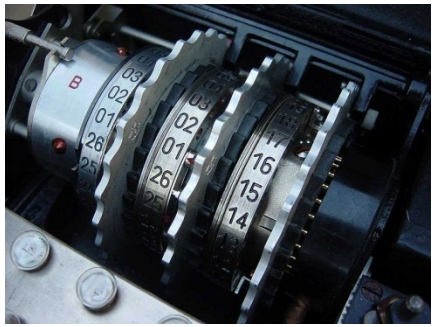
- 第一年：打基础
  - 学基础课：生物信息学、生物信息学算法、分子生物学、机器学习、统计建模
  - 看文献：CNS，生物信息学杂志
  - 寻找研究方向
- 第二年：迈出人生第一步
  - 确定研究方向和思路
  - 收集数据、follow现有工作
  - 编写算法、分析结果等
  - 文章投稿、修改 → BMC bioinformatics
- 第三年：跑步进入现代化
  - 深入研究 → NAR database server issue
  - 转换方向 → PLoS ONE
  - 开展合作 → NAR
- 总结：3年时间，4篇 SCI 影响因子>3（2篇影响因子>7）的生物信息学文章

成功的关键

Health Informatics Lab@USTC

什么是算法?

- 算法是为解决一个公式化表示的问题（任务）而必须执行的一系列步骤和方法（指令）。
  - 输入数据（初始状态）和输出数据（终止状态）
  - 公式化问题：清楚、明确
  - 解决问题：输入到输出的转化过程
- 算法不单单可以用计算机程序来实现，也可以在人工神经网络、电路或者机械设备上实现



算法的伪代码描述

- 适用于实现的算法：
  - 完全准确
  - 程序设计语言

- 伪代码语言：研究算法原理所使用近似编程代码
  - 描述算法的语言，忽略了许多程序设计语言所需要的细节，却能比这些语言更加准确无误地描述算法

伪代码举例

举例：MAX(a,b)

```
1 if a<b
2       return b
3 else
4       return a
```

结果：MAX(a,b) 计算 a 和 b 中的最大值。如 MAX(1,99) 返回值为 99

举例：DIST($x1$,$y1$,$x2$,$y2$)

```
1 dx←(x2−x1)²
2 dy←(y2−y1)²
3 return √(dx+dy)
```

结果：DIST($x1$,$y1$,$x2$,$y2$) 计算两点之间的欧几里德距离，两点坐标分别是 ($x1$,$y1$) 和 ($x2$,$y2$)。DISTANCE(0,0,3,4) 返回值为 5。

Health Informatics Lab@USTC

利用算法解决问题的步骤

- Step1：识别、分析遇到的问题
- Step2：设计算法
- Step3：算法评估
  - 算法正确性
  - 运行效率（复杂度）

Health Informatics Lab@USTC

算法实例：找钱问题

归纳问题

美国找钱问题（United States change problem）：

用最少的硬币个数转换一定金额的钱。

输入：金额总数 M，以分为单位。

输出：最少个数的 25 分硬币 q 个、10 分硬币 d 个、5 分硬币 n 个、1 分硬币 p 个相加，总金额等于 M（即 $25q+10d+5n+p=M$ 并且 $q+d+n+p$ 尽可能的小）。

算法实现

```
USCHANGE(M)
1 while M>0
2     c← 小于（或等于）M 的最大面值的硬币
3     把面值为 c 的硬币给顾客
4     M ←M−c
```

```
USCHANGE(M)
1 r←M
2 q←r/25
3 r←r−25 · q
4 d←r/10
5 r←r−10 · d
6 n←r/5
7 r←r−5 · n
8 p←r
9 return(q,d,n,p)
```

问题的推广

找钱问题（change problem）：

用可能最少数目的硬币，将一定金额的钱 M 兑换成等额的辅币。

输入：一定金额的钱 M，一个 d 种面值辅币的数组 $c=(c_1,c_2\cdots,c_d)$，数组中的元素按降序排列，即 $(c_1>c_2>\cdots>c_d)$。

输出：一列整数 i_1，i_2，…，i_d 使得 $c_1i_1+c_2i_2+\cdots+c_di_d=M$，且 $i_1+i_2+\cdots+i_d$ 尽可能小。

将算法直接推广?

```
BETTERCHANGE(M, c, d)
1 r←M
2 for k←1 to d
3       i_k←r/c_k
4       r←r−c_k • i_k
5 return(i_1, i_2, ⋯, i_d)
```

- 问题：M=40， c=(25, 20, 10, 5, 1)， d=5?

遍历（暴力）算法

```
BRUTEFORCECHANGE(M, c, d)
1  smallestNumberOfCoins ← ∞
2  for each (i_1, ..., i_d) from (0, ..., 0) to (M/c_1, ..., M/c_d)
3      valueOfCoins ← Σ_{k=1}^{d} i_k c_k
4      if valueOfCoins = M
5          numberOfCoins ← Σ_{k=1}^{d} i_k
6          if numberOfCoins < smallestNumberOfCoins
7              smallestNumberOfCoins ← numberOfCoins
8              bestChange ← (i_1, i_2, ..., i_d)
9  return (bestChange)
```

算法分析

- 算法评估：正确性？效率？
- 顺序取值

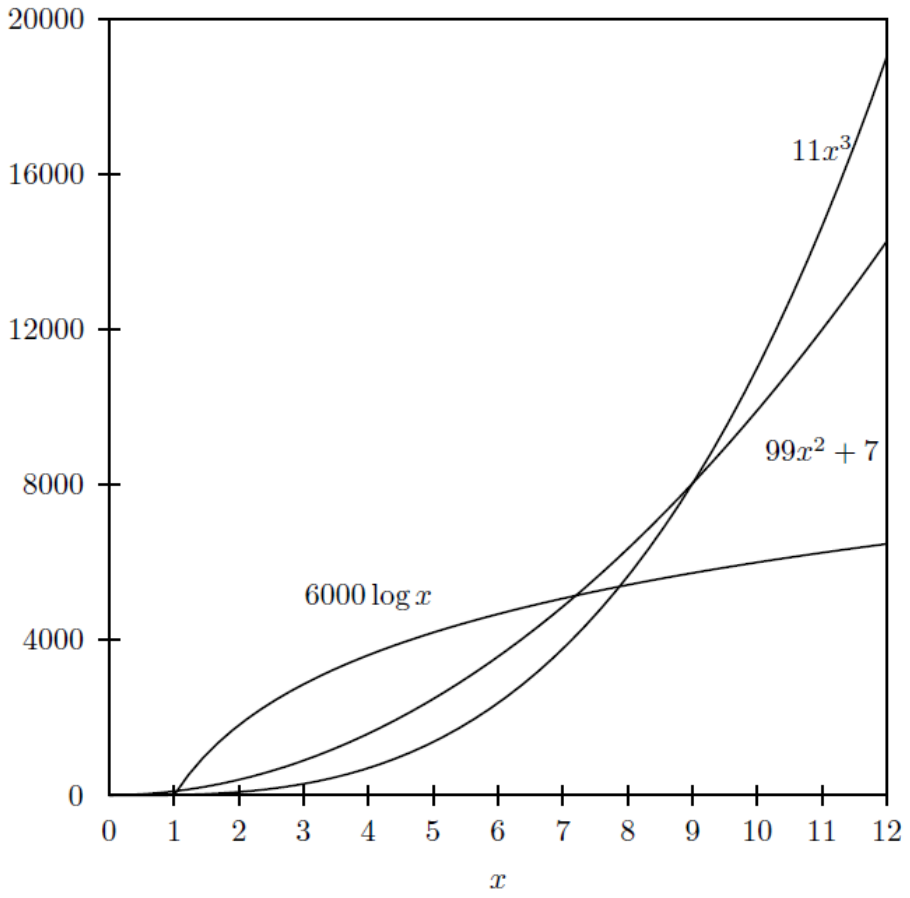
$$
\begin{array}{llll}
(0, & 0, & \ldots, & 0, \quad 0) \\
(0, & 0, & \ldots, & 0, \quad 1) \\
(0, & 0, & \ldots, & 0, \quad 2) \\
 & & \vdots & \\
(0, & 0, & \ldots, & 0, \quad \frac{M}{c_d}) \\
(0, & 0, & \ldots, & 1, \quad 0) \\
(0, & 0, & \ldots, & 1, \quad 1) \\
(0, & 0, & \ldots, & 1, \quad 2) \\
 & & \vdots & \\
(0, & 0, & \ldots, & 1, \quad \frac{M}{c_d}) \\
 & & \vdots &
\end{array}
$$

$$
\begin{array}{llll}
(\frac{M}{c_1}, & \frac{M}{c_2}, & \ldots, & \frac{M}{c_{d-1}} - 1, \quad 0) \\
(\frac{M}{c_1}, & \frac{M}{c_2}, & \ldots, & \frac{M}{c_{d-1}} - 1, \quad 1) \\
(\frac{M}{c_1}, & \frac{M}{c_2}, & \ldots, & \frac{M}{c_{d-1}} - 1, \quad 2) \\
 & & \vdots & \\
(\frac{M}{c_1}, & \frac{M}{c_2}, & \ldots, & \frac{M}{c_{d-1}} - 1, \quad \frac{M}{c_d}) \\
(\frac{M}{c_1}, & \frac{M}{c_2}, & \ldots, & \frac{M}{c_{d-1}}, \quad 0) \\
(\frac{M}{c_1}, & \frac{M}{c_2}, & \ldots, & \frac{M}{c_{d-1}}, \quad 1) \\
(\frac{M}{c_1}, & \frac{M}{c_2}, & \ldots, & \frac{M}{c_{d-1}}, \quad 2) \\
 & & \vdots & \\
(\frac{M}{c_1}, & \frac{M}{c_2}, & \ldots, & \frac{M}{c_{d-1}}, \quad \frac{M}{c_d})
\end{array}
$$

算法运行效率（速度）

- 评价指标
  - 运行时间：
    - 与计算机本身的性能有关
    - 算法A@超级计算机 = 9秒 vs 算法B@台式机 = 10秒
    - 算法A@计算机1 = 10秒 vs 算法B@计算机1 = 9秒
  - 算法运行的操作总数
    - 算法本身的属性：与计算机的属性无关
    - 输入规模的函数

算法运行速度的比较

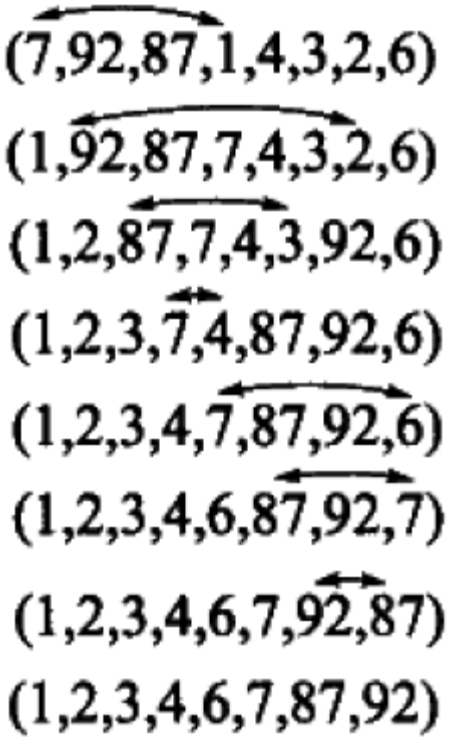

20000
16000
12000
8000
4000
0
0 1 2 3 4 5 6 7 8 9 10 11 12
x
$11x^3$
$99x^2 + 7$
$6000 \log x$

大O记号

- 多项式算法
  - 二项式算法 $O(n^2)$：
    - $99n^2$，$5n^2+3n+1000$
  - 三项式算法 $O(n^3)$：
    - $99n^3$，$n^3+5n^2+3n+1000$
- 指数算法 $O(M^d)$：
  - BRUTEFORCECHANGE：$d \cdot \frac{M^d}{c_1 \cdot c_2 \cdots c_d}$
- 算法效率是在最坏情况下的效率

算法复杂度：例子

- 排序问题



SELECTIONSORT(a, n)

1 **for** $i \leftarrow 1$ **to** $n-1$

2 　　$a_j \leftarrow a_i$，a_{i+1}，…，a_n中的最小元素

3 　　交换 a_i和 a_j

4 **return** a

算法复杂度：例子

```
SELECTIONSORT(a, n)
1 for i←1 to n−1
2     j←INDEXOFMIN(a, i, n)
3     交换 a_i 和 a_j
4 return a
```

```
INDEXOFMIN(array, first, last)
1 index←first
2 for k←first+1 to last
3     if array_k < array_index
4         index←k
5 return index
```

- 算法操作数~(n+1)n/2+3n，复杂度O(n^2)

Health Informatics Lab@USTC

算法设计技术

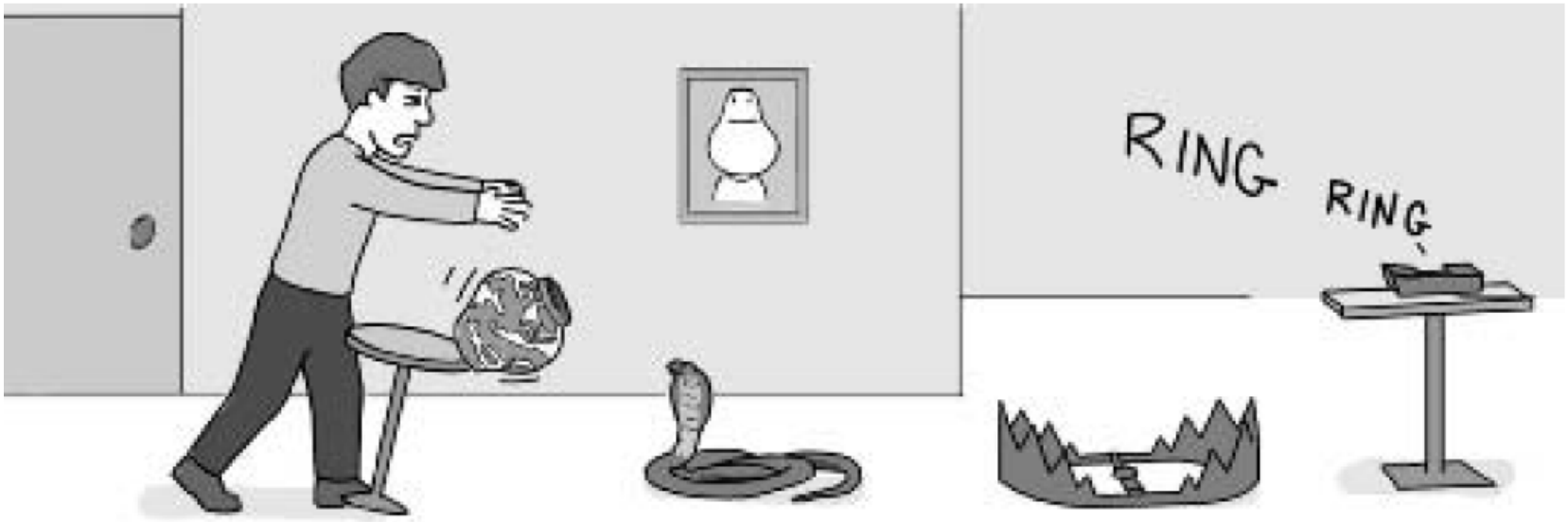
- 穷举（exhaustive search）/暴力（brute force）算法
- 分支定界（branch and bound）算法
- 贪婪（greedy）算法
- 动态规划（dynamic programming）算法
- 分而治之（divide-and-conquer）算法
- 随机化（randomized）算法
- 机器学习（machine learning）算法

一个有趣的例子

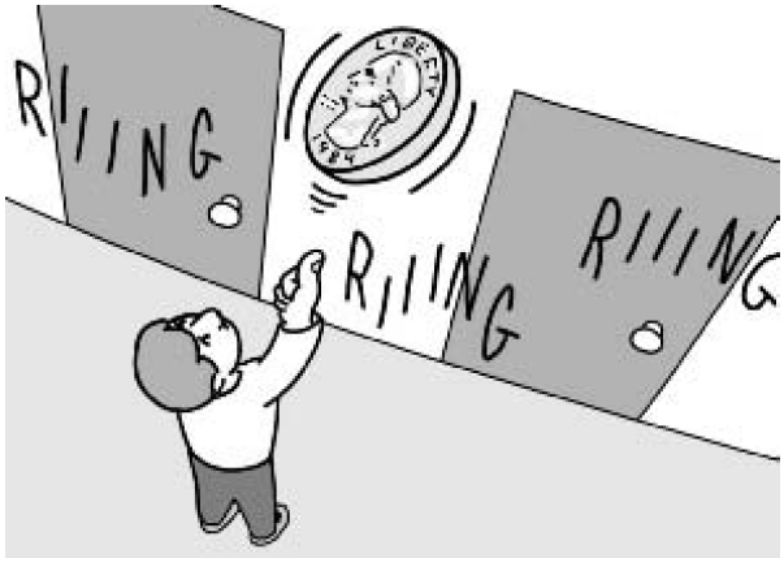
例子：遍历算法

例子：分支定界算法

例子：贪婪算法



例子：随机化算法



实验者	年代	投掷次数	相交次数	圆周率估计值
沃尔夫	1850	5000	2531	3.1596
史密斯	1855	3204	1219	3.1554
德摩根	1680	600	383	3.137
福克斯	1884	1030	489	3.1595
拉泽里尼	1901	3408	1808	3.1415929
赖纳	1925	2520	859	3.1795

投针问题的实验

Health Informatics Lab@USTC

例子：机器学习算法

	客厅	书房	厨房	卧室
周一	50	1	9	40
周二	0	90	0	10
周三	1	66	3	30
周四	2	57	2	39
周五	4	50	4	42
周六	78	1	1	20
周日	66	0	32	2

- p（电话在客厅|today=周六）？， p（电话在书房|today=周六）？
- 周六：客厅->卧室->书房->厨房
- 电话在书房，今天周几？