11.串

（b1）串匹配

邓俊辉

deng@tsinghua.edu.cn

串匹配

```
% grep <pattern> <text>
```

文本 T = now is the time for all good people to come

模式 P = people

❖记 n = |T| 和 m = |P|，通常有 n >> m >> 2　　//比如，100,000 >> 100 >> 2

❖Pattern matching

detection：P是否出现？

location：首次在哪里出现？　　//本章主要讨论的问题

counting：共有几次出现？　　//find /c "2013" students.txt

enumeration：各出现在哪里？　　//find "2013" students.txt

串匹配

❖ 歧义： T = " 1 0 0 1 1 0 1 1 0 1 0 1 1 0 1 1 0 0 1 "

P = " 1 0 1 1 "

❖ 应用： 文本编辑器、数据库检索、C++模板匹配、模式识别、搜索引擎、...

❖ 应用： 生物序列分析（biological sequence analysis）

通常不能完全匹配

——alignment：最接近的匹配在什么位置？

HBA_HUMAN vs. HBB_HUMAN

G S A Q V K G H G K K V A D A L T N A V A H V D D M P N A L S A L S D L H A H

G N P K V K A H G K K V L G A F S D G L A H L D N L K G T F A T L S E L H C D

算法评测

❖ 如何客观地测量与评估串匹配算法的性能？具体采用什么标准与策略？

❖ 随机T + 随机P？不妥！

❖ 以 $\Sigma = \{0, 1\}^*$ 为例

$|\{ \text{长度为 } m \text{ 的P} \}| = 2^m$

$|\{ \text{长度为 } m \text{ 且在T中出现的P} \}| = n - m + 1 < n$

匹配成功的概率 $= n/2^m << 100{,}000 / 2^{100} < 10^{-25}$

如此，将无法对算法做充分测试

❖ 随机T，对成功、失败的匹配分别测试

成功：在T中，随机取出长度为m的子串作为P；分析平均复杂度

失败：采用随机的P；统计平均复杂度