

L-MinHash：一种蛋白质同源性 搜索的高效算法

答 辩 人：项溢馨

指导老师：王和兴 老师

2022 年 6 月 9 日

内容提要

01 选题背景与研究意义

02 蛋白质同源性搜索问题

03 蛋白质同源性搜索的高效算法L-MinHash

理论基础 : Jaccard相似度是编辑相似度的上界

如何解决 l^2 : Jaccard相似度估计的近似算法MinHash

如何解决 n^2 : 基于Jaccard相似度的LSH聚类算法

04 实验结果

05 总结与展望

内容提要

01 选题背景与研究意义

02 蛋白质同源性搜索问题

03 蛋白质同源性搜索的高效算法L-MinHash

理论基础 : Jaccard相似度是编辑相似度的上界

如何解决 l^2 : Jaccard相似度估计的近似算法MinHash

如何解决 n^2 : 基于Jaccard相似度的LSH聚类算法

04 实验结果

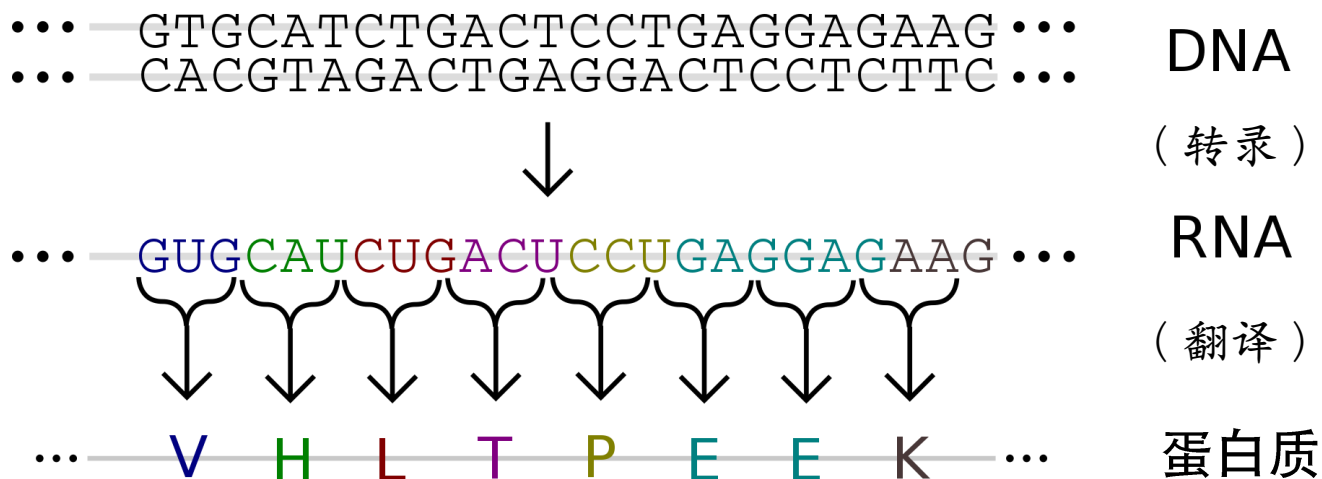
05 总结与展望

选题背景与研究意义

分子生物学

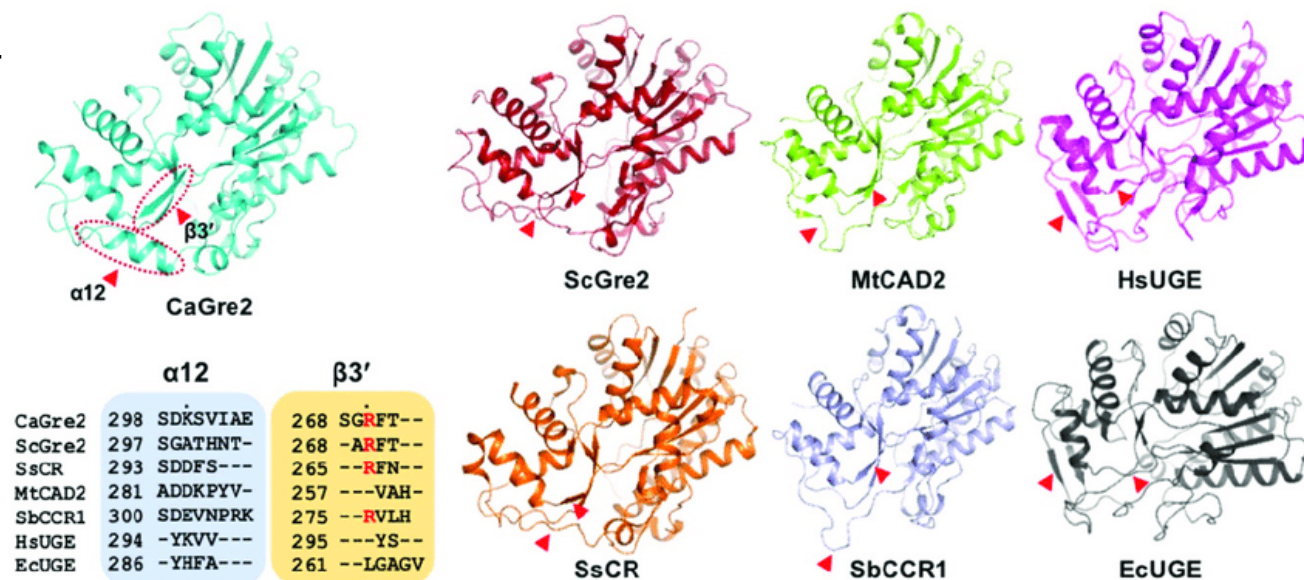
- DNA: 由A、C、G、T四种碱基
- RNA: 由A、C、G、U四种碱基
- 蛋白质: 由 20 种氨基酸

组成的分子序列
(字符串)



选题背景与研究意义

同源蛋白质



- 由一个**共同祖先**演化而来的多个结构间所具有的**共性**特征
- **高度相似性**意味两个序列极有可能拥有共同祖先
- 蛋白质的氨基酸序列决定其三维结构，从而决定其**功能特性**
→ 进一步分析，应用（制药等）

内容提要

01 选题背景与研究意义

02 蛋白质同源性搜索问题

03 蛋白质同源性搜索的高效算法L-MinHash

理论基础 : Jaccard相似度是编辑相似度的上界

如何解决 l^2 : Jaccard相似度估计的近似算法MinHash

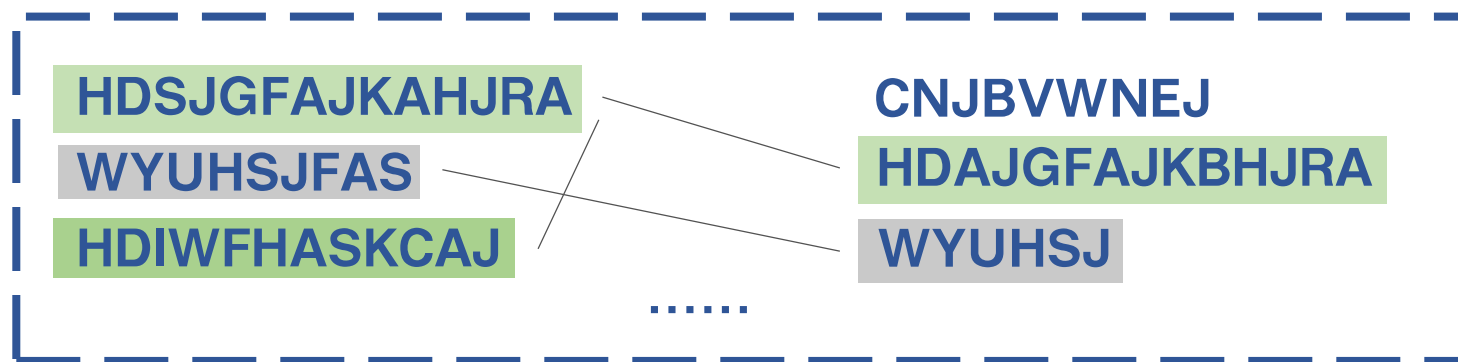
如何解决 n^2 : 基于Jaccard相似度的LSH聚类算法

04 实验结果

05 总结与展望

蛋白质同源性搜索问题

蛋白质同源性搜索问题



编辑相似度 $E_s(a, b)$

- 编辑相似度的定义基于编辑距离
- 编辑距离指两个字符串之间,通过增删改的编辑操作,由一个转换成另一个所需的最少操作次数

增	删	改
AB	ABB	AC
ABB	AB	AB

编辑相异度

$$E_d(a, b) = E_d(b, a) = \frac{\text{lev}(a, b)}{\max(|a|, |b|)}$$

编辑距离

$$E_s(a, b) = 1 - E_d(a, b)$$

蛋白质同源性搜索问题

动态规划求解编辑距离 $O(l_a l_b)$

		A	B	B	C
	0	1	2	3	4
A	1	0	1	2	3
C	2	1	1	2	3
B	3	2	2	1	2

$$\text{lev}(a, b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \text{lev}(\text{tail}(a), \text{tail}(b)) & \text{if } a[0] = b[0], \\ 1 + \min \begin{cases} \text{lev}(\text{tail}(a), b) \\ \text{lev}(a, \text{tail}(b)) \\ \text{lev}(\text{tail}(a), \text{tail}(b)) \end{cases} & \text{otherwise} \end{cases}$$

蛋白质同源性搜索问题的时间复杂度

$O(n^2 l^2)$

n : 蛋白质序列数量 8×10^6

$3 \times 10^{18} \approx 3 \times 10^{10}$ 秒

l : 蛋白质序列平均长度 200

≈ 1000 年

蛋白质同源性搜索问题

研究现状及挑战

- 序列比对算法： Smith-Waterman算法和Needleman-Wunsch算法及其优化
- 启发式策略加速： FASTA和BLAST软件
- 哈希算法避开序列直接比较： Mash、Mashmap和Mhap
- **挑战**： 超长的蛋白质序列和蛋白质序列规模庞大
- **目标**： 在可接受时间范围内快速搜索出高相似度的序列对

内容提要

01 选题背景与研究意义

02 蛋白质同源性搜索问题

03 蛋白质同源性搜索的高效算法L-MinHash

理论基础 : Jaccard相似度是编辑相似度的上界

如何解决 l^2 : Jaccard相似度估计的近似算法MinHash

如何解决 n^2 : 基于Jaccard相似度的LSH聚类算法

04 实验结果

05 总结与展望

蛋白质同源性搜索的高效算法L-MinHash

设计思路

- 对蛋白质序列进行聚类，在更小的类中计算编辑相似度

		长度	相似度	概率
[HDSJGFAJKAHJRA	1	100%	1/20
	HDAJGFAJKBHJRA			
	HDIWFHASKCAJ			
.....				
]	CNJBVWNEJ	2	50%	1/160
	WYUHSJ WYUHSJFAS			

如果将 8×10^6 条序列分成10000簇

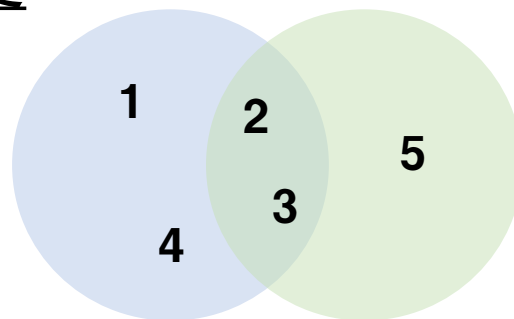
在每个簇中分别求解 时间复杂度从 3×10^{18} 变成 3×10^{14}

- 何种距离度量 编辑距离 $O(l^2)$
- 何种聚类算法 ~~k-means~~ 聚类 $O(nkt)$

蛋白质同源性搜索的高效算法L-MinHash

距离度量：Jaccard相似度

$$J_s(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

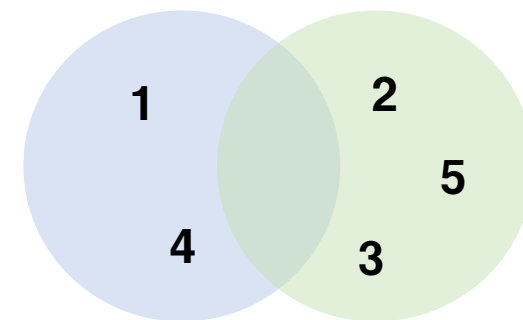
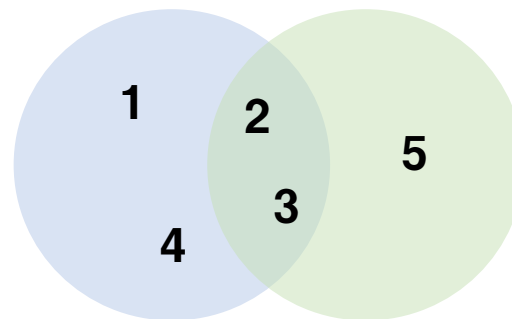
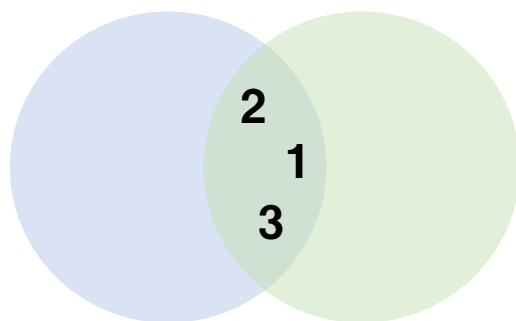


- Jaccard相似度是编辑相似度的上界（已证明）
- Jaccard相似度不超过阈值，则编辑相似度一定不高于阈值，
两个序列肯定不相似
- 计算的时间复杂度 $O(n + m)$

蛋白质同源性搜索的高效算法L-MinHash

Jaccard相似度的近似估计算法MinHash

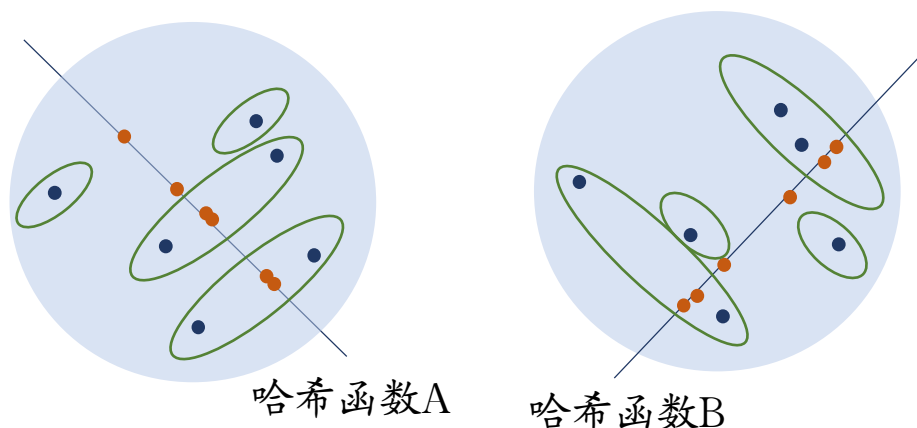
$$J_s(A, B) = \frac{|A \cap B|}{|A \cup B|}$$



- 交集中的数命中最小值则MinHash相等
- $\Pr[h_{min}(A) = h_{min}(B)] = J_s(A, B)$
- 重复多次hash即可估计交集大小

蛋白质同源性搜索的高效算法L-MinHash

聚类算法：基于Jaccard相似度的局部敏感哈希（LSH）

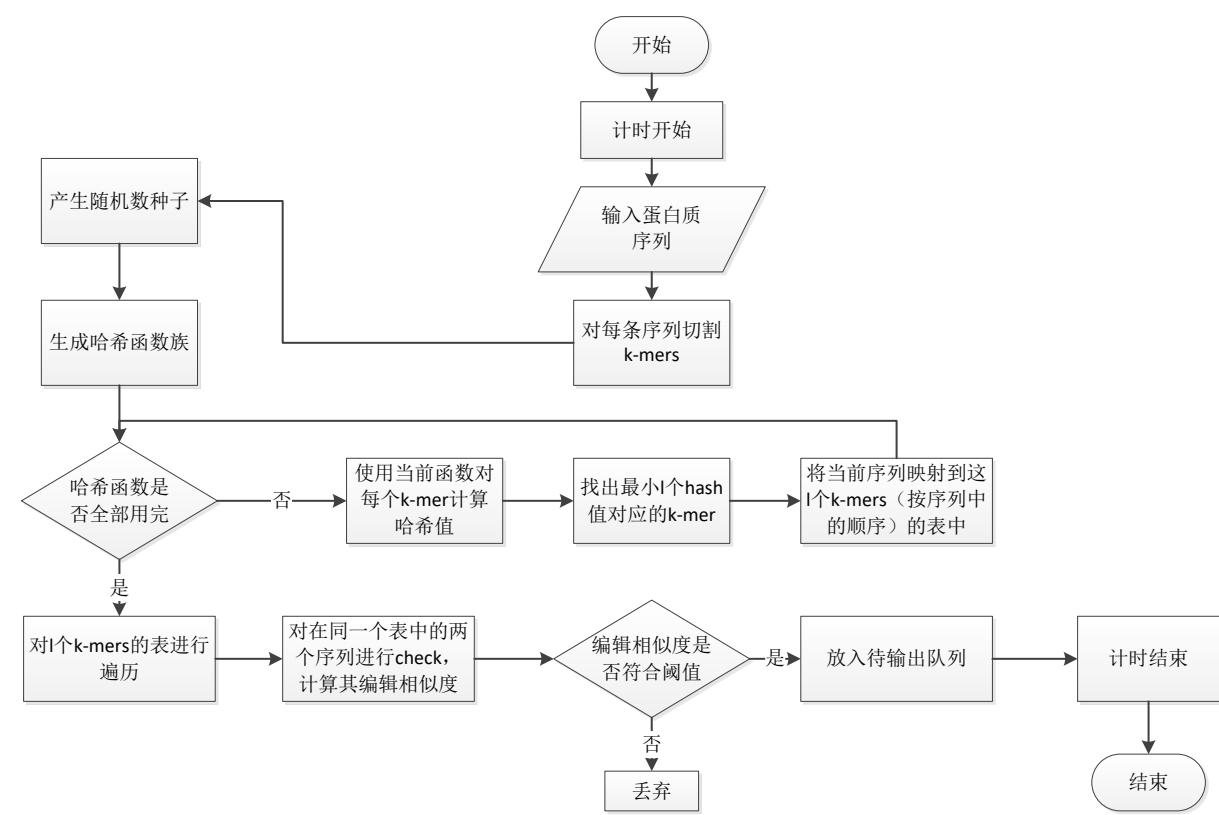
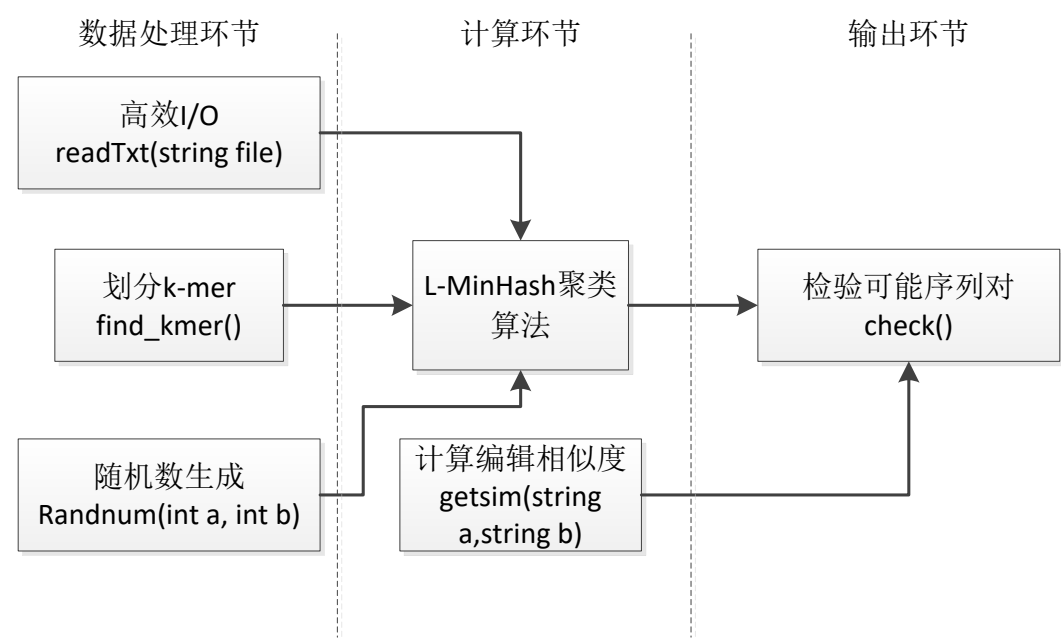


- 效率高，步骤简洁，误差大
- 线性的时间复杂度
- 增加哈希函数个数可以迅速降低误差率

蛋白质同源性搜索的高效算法L-MinHash

算法框架

$L\text{-MinHash} = \text{MinHash} + \text{LSH}$



蛋白质同源性搜索的高效算法L-MinHash

算法伪代码（部分）

算法 4.1 k-mer 分割算法

输入：长度为 L 的蛋白质序列 S ，k-mer 的长度 k
输出：输入序列 S 的全部 k -mer 加权集合 V

```
1:  $V \leftarrow []$  //初始化集合
2: for  $i = 1$  to  $L-k+1$  //每条序列拥有的 k-mer 数量为  $L-k+1$ 
3:    $s \leftarrow S[i..i+k-1]$ 
4:    $V \leftarrow V \cup \{s, H[s]\}$ 
5:    $H[s] = H[s]+1$ 
6: end for
```

算法 4.2 Order Min Hash 算法

输入：n 组蛋白质序列 $S[]$ ，k-mer 的长度 k ，散列函数模数 p ，全域散列模数 pp
输出：聚类后存在同源性可能的序列对集合 V

```
1:  $V \leftarrow []$ 
2: let  $A[1..L]$  be a new array //A 存放散列函数的基数
3: for  $i=1$  to  $L$ 
4:    $A[i] = \text{RANDOM}(1, p-1)$  //生成 L 组散列函数
5: end for
6: for  $ii = 1$  to  $L$ 
7:   for  $i = 1$  to  $n$ 
8:      $m \leftarrow |k\_mer[i]|$  //m 是第 i 个序列 k-mer 的数量
9:     for  $j = 1$  to  $m$ 
10:       $x \leftarrow k\_mer[i][j]$ 
11:       $val = 1$  //val 用来计算 k-mer 的散列值
12:      for  $jj = 1$  to  $k$ 
13:         $val = (val * A[ii] + x[jj]) \% p$ 
14:      end for
15:       $hash[j] = val \% pp$ 
```

算法 4.3 动态规划求解编辑相似度

输入：长度分别为 l_1, l_2 的蛋白质序列 S_1 和 S_2
输出：输入序列 S_1 和 S_2 的编辑相似度 s

```
1: let  $dp[0..l_1][0..l_2]$  be a new array
2: for  $i = 0$  to  $l_1-1$ 
3:    $dp[i][0] = i$ 
4: end for
```

```
5: for  $i = 0$  to  $l_2-1$ 
6:    $dp[0][i] = i$ 
7: end for
8: for  $i = 1$  to  $l_1$ 
9:   for  $j = 1$  to  $l_2$ 
10:    if  $S_1[i] = S_2[j]$ 
11:       $dp[i][j] = dp[i-1][j-1]$ 
12:    else  $dp[i][j] = 1 + \min\{ dp[i-1][j], dp[i][j-1], dp[i-1][j-1] \}$ 
13:    end for
14: end for
15:  $s = 1 - dp[l_1][l_2] / \max(l_1, l_2)$ 
15: return  $s$ 
```

```
16: end for
17: select  $hash[1..m]$  corresponding to the smallest  $l$  values
//将最小的 l 个值放置在 hash 序列最左边，l 个值内部保持原有序列排序
18: for  $j = 1$  to  $l$ 
19:    $v \leftarrow v \cup hash[j]$ 
20: end for
21:  $set[v] \leftarrow set[v] \cup i$  //对包含 l 个值和相对位置信息的列表做一个散列
22: end for
23: for  $i = 1$  to  $|set[ ]|$ 
24:   for  $j = 1$  to  $|set[i]|$ 
25:     for  $jj = j + 1$  to  $|set[i]|$ 
26:        $V \leftarrow V \cup \{ set[i][j], set[i][jj] \}$  //散列相同的两个序列可能同源
27:     end for
28:   end for
29: end for
30: end for
31: return  $V$ 
```


内容提要

01 选题背景与研究意义

02 蛋白质同源性搜索问题

03 蛋白质同源性搜索的高效算法L-MinHash

理论基础 : Jaccard相似度是编辑相似度的上界

如何解决 l^2 : Jaccard相似度估计的近似算法MinHash

如何解决 n^2 : 基于Jaccard相似度的LSH聚类算法

04 实验结果

05 总结与展望

实验结果

模型的效率与准确性的对比

- 初步模型相较于现有模型提速约30%
- 在一些适当参数下，初步模型能达到较高的准确率

初步模型在真实数据集上的效果

- 在将 8×10^6 条序列中共找到992821对相似度高于0.7的序列对

```
1 ILSSTXCGJGAKSMJFXZJGCJPOGTLVXXPGNBMNJDJTAXCTJBGLGWJJBKHSTUPKKMGJLGXPPGKVXUPZKPLJHHLXKKTLKKTJBKUMAJPJPJXGLWXWFGIVWXWJLXPKMXXGGJLGSUJUANWKPFK
2 ILSSTXCGJGAKSMJFXZJGCJPOGTLVXXPGNBMNJDJTAXCTJBGLGWJJBKHSTUPKKMGJLGXPPGKVXUPZKPLJHHLXKKTLKKTJBKUMAJPJPJXGLWXWFGIVWXWJLXPKMXXGGJLGSUJUANWKPFK
3 0.999493
4 IPPOGGPAALASTBJABHAXTGPUWXOOWPXTJGJAAJPXL TZOXSDSTJAJAUTSGXPLJTGABWLSWAMLGKPKAGXSBSAVVGHMJHBKRUOLJHGBOBTPHAPGVWGTABUGWOOWPWNGWZTGBTIWWIFHWGHI
5 IPPOGGPAALASTBJABHAXTGPUWXOOWPXTJGJAAJPXL TZOXSDSTJAJAUTSGXPLJTGABWLSWAMLGKPKAGXSBSAVVGHMJHBKRUOLJHGBOBTPHAPGVWGTABUGWOOWPWNGWZTKBTIWWIFHWGHI
6 0.999486
7 ITPXTA0OLKZJLPMKKTAMAWKSOTWSTKPLWUSWKAMLVPLKTUTSUJXHMTCP0BXVPSWUSAPGVASVUXLTTHAAGGGRLVPHLKMCKGWPPHLXPHXRTCOMLKKOHGOWGAUAJHSDZKSABXPJEIKPWFS
8 ITPXTA0OLKZJLPMKKTAMAWKSOTWSTKPLWUSWKAMLVPLKTUTSUJXHMTCP0BXVPSWUSAPGVASVUXLTTHAAGGGRLVPHLKMCKGWPPHLXPHXRTCOMLKKOHGOWGAUAJHSDZKSABXPJEIKPWFS
9 0.99935
10 IUPGPPVGPAPXATASSTPPKJOLPOUTGDAWPKLILFOBALLWTFAXPGCSBGFXOOTTWTMKAAMOCMABXLLAAXKPPPGPOHHOCLTBGASWSPBSPXBBXWTATXP5JISAMMOGGGKPCXPXZPLWAAYGOO
11 IUPGPPVGPAPXATASSTPPKJOLPOUTGDAWPKLILFOBALLWTFAXPGCSBGFXOOTTWTMKAAMOCMABXLLAAXKPPPGPOHHOCLTBGASWSPBSPXBBXWTATXP5JISAMMOGGGKPCXPXZPLWAAYGOO
12 0.999332
13 IPFLLOPOSPLALLLLPLPSTAPIDATGG000LGMJXGTGA00LOLFOTLUOJTGKKYATAHGTEGTAGTKAHLTTAGSTA0UBTPPPPWUBABPOOPSLOOCTGQGATTLVYUHATTHKTTTTTOXOGZMGABMAHI
14 IPFLLOPOSPLALLLLPLPSTAPIDATGG000LGMJXGTGA00LOLFOTLUOJTGKKYATAHGTEGTAGTKAHLTTAGSTA0UBTPPPPWUBABPOOPSLOOCTGQGATTLVYUHATTHKTTTTTOXOGZMGABMAHI
```

内容提要

01 选题背景与研究意义

02 蛋白质同源性搜索问题

03 蛋白质同源性搜索的高效算法L-MinHash

理论基础 : Jaccard相似度是编辑相似度的上界

如何解决 l^2 : Jaccard相似度估计的近似算法MinHash

如何解决 n^2 : 基于Jaccard相似度的LSH聚类算法

04 实验结果

05 总结与展望

总结与期望

- 本文提出了一种无需对所有序列进行两两比对而找出高相似度序列对的新算法。
- 首先**证明**了序列间的Jaccard相似度是编辑相似度的上界。
- 利用**最小哈希和局部敏感哈希**设计了L-MinHash算法
- 最小哈希实现了Jaccard相似度的**快速估计**
- 局部敏感哈希基于相似度估计值对全部序列进行了**聚类**

- 分别在**模拟数据**和**真实数据**上对不同算法的性能进行了测试，实验结果验证了新算法具有出色的性能优势，值得进行深入研究。
- **不足**：系统框架较粗糙，算法性能仍有待改进
- **目标**：下一步继续完善算法，进行实验，对论文进行细致修改后投稿。并对系统进行封装，使之可以真正应用于生物学研究。

谢 谢

请各位老师批评指正