不等长 LCS 问题的探究

吴茁

摘要

本文主要探究长度比值固定的两个由 k 个字母组成序列 (不妨设长度分别为 an, bn) 的最长公共子序列长度的期望的不等式, 并证明了当 k, n 趋于无穷时, 该期望 E 具有性质 $\frac{E\sqrt{k}}{\sqrt{n}}$ 的极限是 $2\sqrt{ab}$.

1 导言

考虑两个字母序列 μ 和 ν, 其中每种序列由 k 种字母组成. 经典的公共子序列问题 (LCS, longest common subsequence 的简称) 考虑最大的可能值 L, 使得存在正整数 $1 \le i_1 \le i_2 \le ... \le i_L$, $1 \le j_1 \le j_2 \le ... \le j_L$, 满足 $\mu_{i_t} = \nu_{j_t}$ 对任意正整数 $1 \le k \le L$ 成立.

LCS 问题在几个截然不同的领域或多或少地出现, 例如, 计算机程序版本的比较, 或分子生物学. 这个问题的生物学动机是长分子, 如 DNA, 可以示意性地表示为一个有限字母表的序列. 从进化的角度来看, 通过找到它们最接近的共同祖先来比较两个 DNA 序列是很自然的. 如果假设这些分子只通过在表示字符串中插入新符号的过程来进化, 那么祖先就是表示分子的字符串的子字符串. 因此, 两个字符串的最长公共子序列的长度是衡量两个字符串之间的相似程度的合理尺度.

一般的传统 LCS 问题关心等长的字母序列的一些性质. 例如, 人们关心两个随机的 k 个字母的长度均为 n 的序列的最长公共子序列长度的期望. 在此方面的经典结果是, 1970 年, Chvátal 和 Sankoff 证明了当 k 固定时, 该期望与 n 的比值在 n 趋于无穷时趋近于一个常数. 记这个常数为 γ_k. 不过, 我们现在仍然无法具体计算出 γ_k 的值, 而是只有一个关于 γ_k 的一个简单的估计. 关于此方面最近的结果是, 在 2003 年, 文 [1] 中 Marcos Kiwi, Martin Loebl, Jiri Matousek 合作证明了下述的结论:

$$\lim_{k\to\infty} \gamma_k \sqrt{k} = 2.$$

在本文中, 我们将 LCS 问题一般化, 关心不等长的字母序列的 LCS 问题, 并利用文 [1] 中所证明的一个引理, 将上述结论推广到不等长的 LCS 问题. 我们的结论探究了长度成固定比例时 LCS 问题的推广, 因此, 我们的结论在一些需要比较不等长的结构的问题中仍然有着应用的价值.

2 结论的陈述

假设 A, B 是两个不交的有序集. 我们不妨将 A 中的按大小依次标号为 $1, 2, .., |A|$, B 中的元素按大小依次标号为 $1, 2, ..., |B|$. 假设二部图 G 分别以 A, B 作为两个部分, A, B 两部分分别称为 G 的上部和下部. 方便起见, 我们总是记连接图 G 里 A 中标号为 i 的元素和图 G 里 B 中标号为 j 的元素为边 ij, 即前面的数代表 A 中的标号, 后面的数代表 B 中的标号.

我们称 G 中的两条边 ab 和 $a'b'$ 是不交的, 如果 $(a' - a)(b' - b) > 0$; 二部图 G 中的一个匹配称为平面的, 如果该匹配中的任意两条边均不交. 记 $L(G)$ 为 G 最大的平面匹配的大小 (即该匹配

的边数). 这个定义容易被我们直观的看出.

我们定义随机图 $\Sigma(K_{r,s};k)$ 如下: 我们以等概率将 $K_{r,s}$ 的每个顶点涂上颜色 $1,2,...,k$ 之一, 然后连接所有颜色相同且在两个不同部分的两点. 那么, 如果我们假设这个二部图的两个部分的点的标号依次为 $1,2,..,r$ 和 $1,2,...,s$, 我们就可以自然的定义 $L(\Sigma(K_{r,s};k))$. 在下述行文中, 为了方便, 在不引起混淆的情况下, 我们会将随机图 $\Sigma(K_{r,s};k)$ 直接简记为随机图 $K_{r,s}$. 这个模型实际上就是长度分别为 r 和 s 的两个由 k 种字母组成的序列 μ 和 ν, 只不过我们采用了图论的描述以方便我们更清晰的刻画问题.

给定正整数 a,b. 在本文中, 我们主要研究随机图 $G = K_{an,bn}$(从下文起, 一切图 G 均表示 $K_{an,bn}$). 本文的主要目的是, 利用概率方法, 计算当 k 趋于无穷时 $L(G)$ 的期望的级别, 并给出关于 $L(G)$ 的分布的一个估计. 本文的主要结果是下述的定理:

定理 1. 对每个实数 $\epsilon > 0$, 存在正实数 k_0, C, c, 使得对一切正整数 $k > k_0$ 和正整数 $n > C\sqrt{k}$, 均有

$$\mathbf{P}[|L(\Sigma(K_{an,bn};k)) - \frac{2n\sqrt{ab}}{\sqrt{k}}| \geq \epsilon\frac{2n\sqrt{ab}}{\sqrt{k}}] \leq e^{-\frac{cn}{\sqrt{k}}}.$$

由定理 1, 我们容易得到下面的推论;

推论 2. 给定正整数 a,b. 记

$$\gamma_k^{a,b} = \lim_{n\to\infty} \mathbf{E}(L(\Sigma(K_{an,bn};k)/n)).$$

则对任意正整数 a,b, 我们有

$$\lim_{k\to\infty} \gamma_k^{a,b}\sqrt{k} = 2\sqrt{ab}.$$

实际上, 在文 [1] 中证明的主要结论即我们的证明中 $a = b = 1$ 的情形, 本文的主要目的即将文 [1] 中的主要结论推广到对于一般的 a 和 b 的情形.

3 工具

我们记

$$m_{\max} = (1+\epsilon)\frac{2\sqrt{ab}n}{\sqrt{k}}, m_{\min} = (1-\epsilon)\frac{2\sqrt{ab}n}{\sqrt{k}}.$$

那么定理 1 等价于分别证明以下两个不等式:

$$\mathbf{P}[L(\Sigma(K_{an,bn};k)) \geq m_{\max}] \leq e^{-\frac{cn}{\sqrt{k}}}, \ \mathbf{P}[L(\Sigma(K_{an,bn};k)) \leq m_{\min}] \leq e^{-\frac{cn}{\sqrt{k}}}.$$

我们将在第四节和第五节中分别证明这两个不等式. 本节我们先指出证明这两个不等式所使用的工具和方法. 和文 [1] 中的方法类似, 我们的主要技巧是通过将大的图利用一些技巧切割成小的子图, 导出大的图和小的图之间的一些概率关系, 再对小的图进行估计. 为了对小的图进行估计, 我们将沿用文 [1] 中关于小的图的估计的一个重要引理作为我们最主要的工具. 文 [1] 中利用 Talagrand's inequality 证明了下述关于比较小的图的引理:

引理 3. 对每个实数 $\delta > 0$, 存在一个常数 C 满足:

(a) 若 $rs \geq Ck$ 且 $r + s\sqrt{rs} \leq \delta \frac{k^{\frac{3}{2}}}{6}$, 则对任意实数 $t \geq 0$ 和 $m_u = 2(1+\delta)\frac{\sqrt{rs}}{k}$, 我们有

$$\mathbf{P}[L(\Sigma(K_{r,s};k)) \geq m_u + t] \leq 2e^{-\frac{t^2}{8m_u+t}}.$$

(b) 若 $rs \geq Ck$ 且 $r + s \leq \delta \frac{k}{6}$, 则对任意实数 $t \geq 0$ 和 $m_l = 2(1-\delta)\frac{\sqrt{rs}}{k}$, 我们有

$$\mathbf{P}[L(\Sigma(K_{r,s};k)) \geq m_l - t] \leq 2e^{-\frac{t^2}{8m_u}}.$$

引理 3 的证明参见文 [1]. 利用该引理, 我们可以得出关于一些比较小的图的一些估计, 然后结合一些合适的分块方式对原图进行估计. 注意, 本文中采取的分块方式和估计技巧并不与文 [1] 相同, 我们的方法改进了文 [1] 中的技术, 使得其更为简便, 同时能够适应关于不等长的 LCS 问题.

4 下界的证明

本节的目标是为了证明第三节中给出的第二个不等式. 记 $d = [\frac{\delta k}{6(a+b\sqrt{ab})}]$, 其中 $\delta = \delta(\epsilon) < 1$ 是一个待定的常数, 具体的条件由下面的计算确定. 则易知存在实数 C, 使得 $k > C$ 时, 我们有 $abd^2 \geq Ck$, 而此时显然有 $ad + bd \leq \delta \frac{k}{6}$ 和且 $ad + bd\sqrt{abd^2} \leq \delta \frac{k^{\frac{3}{2}}}{6}$, 因此, 我们可以对图 $K_{ad,bd}$ 运用引理 3 中两部分的条件.

我们现在从图 G 划分出若干个形如 $K_{ad,bd}$ 的子图; 具体来说, 设 $q = [\frac{n}{d}]$, 则我们记 G_i 为由 $K_{an,bn}$ 上部的点 $ad(i-1)+1, ad(i-1)+2, ..., a(i-1)+ad$ 和下部的点 $bd(i-1)+1, bd(i-1)+2, ..., b(i-1)+bd$ 诱导的子图. 那么, G_i 两两不交, 且我们显然有

$$L(G) \geq L(G_1) + \cdots + L(G_q).$$

注意到随机图 G_i 独立同分布且同构于 $K_{ad,bd}$, 因此, 若设 $\mathbf{E}(G_i) = \mu$, 则由引理 3,

$$\mu \geq (1-2\delta)\frac{2d\sqrt{ab}}{\sqrt{k}} \times \mathbf{P}[L(G_i) \geq (1-2\delta)\frac{2d\sqrt{ab}}{\sqrt{k}}] \geq (1-2\delta)\frac{2d\sqrt{ab}}{\sqrt{k}}(1 - 2\exp(-\frac{\delta^2}{4(1+\delta)} \times \frac{d}{\sqrt{k}})).$$

我们先取 δ 使得 $(1-2\delta)^2 > 1 - \frac{\epsilon}{2}$, 再取充分大的常数 K 使得

$$\exp(-K\frac{\delta^2}{4(1+\delta)}) < \delta,$$

则 $\frac{d}{\sqrt{k}} > K$ 时, 我们有

$$\mu > (1 - \frac{\epsilon}{2})\frac{2d\sqrt{ab}}{\sqrt{k}}.$$

注意 d 的定义, 这在 k 充分大时总成立. 使用完全类似的方法, 我们可以证明

$$\mu < (1 + \frac{\epsilon}{2})\frac{2d\sqrt{ab}}{\sqrt{k}}.$$

记 $v=(1-\frac{\epsilon}{2})\frac{2d\sqrt{ab}}{\sqrt{k}}$. 注意到

$$\mathbf{P}[L(G)\leq m_{\min}]\leq \mathbf{P}[\sum_{i=1}^{q}L(G_i)\leq m_{\min}],$$

记 $u=qv-m_{\min}$, 则

$$u=\frac{\epsilon}{2}\times\frac{2n\sqrt{ab}}{\sqrt{k}}-(1-\frac{\epsilon}{2})\times\frac{2(n-qd)\sqrt{ab}}{\sqrt{k}})\geq\frac{\epsilon}{2}\times\frac{2n\sqrt{ab}}{\sqrt{k}}-\frac{2d\sqrt{ab}}{\sqrt{k}}.$$

由于 $n>C\sqrt{k}$, 因此 k 充分大时我们有 $u\geq\frac{\epsilon}{4}\times\frac{2n\sqrt{ab}}{\sqrt{k}}$. 此时,

$$\mathbf{P}[\sum_{i=1}^{q}L(G_i)\leq m_{\min}]\leq\mathbf{P}[\sum_{i=1}^{q}L(G_i)\leq q\mu-u].$$

若 $L(G_i)>4\mu$, 则称 G_i 是坏的, 其余的称为好的. 注意每个坏的 G_i 为 $L(G_i)-3u$ 的和至少贡献 3μ, 而好的至多贡献 $-\mu$, 因此, 坏的 G_i 不多于 $t_0=[\frac{1}{3}q]$ 个.

假设 $G_1,G_2,...,G_t$ 是好的, 那么由 Chernoff 不等式, 我们有

$$\mathbf{P}[\sum_{i=1}^{t}L(G_i)\leq t\mu-u]\leq\exp(-\frac{u^2}{4q\mu}).$$

注意, 好的 $L(G_i)$ 的和小于等于 $t\mu-u$ 已经蕴含在所有的 $L(G_i)$ 之和大于等于 $q\mu-u$ 中. 因此, 原概率就小于等于对一切从 G_i 中选取 t_0 个构成的组, 这一组的 $L(G_i)$ 之和小于等于 $t_0\mu-u$ 这一族事件的概率之和.

而我们选出 t_0 个好的的方法数

$$\binom{q}{t_0}\leq(\frac{eq}{\frac{q}{3}})^{\frac{q}{3}}=e^{Tq},$$

其中 T 为某个常数. 故

$$\mathbf{P}[\sum_{i=1}^{q}L(G_i)\leq t\mu-u]\leq\exp(-\frac{u^2}{4q\mu}+Tq).$$

结合 $q=[\frac{n}{d}]>\frac{n}{2d}$, 我们有对于充分大的正整数 k, 总有

$$\mathbf{P}[\sum_{i=1}^{q}L(G_i)\leq m_{\min}]\leq\mathbf{P}[\sum_{i=1}^{q}L(G_i)\leq q\mu-u]\leq\exp(-\frac{\epsilon^2\sqrt{ab}}{100A(1+\epsilon)}\times\frac{2n}{\sqrt{k}}).$$

因此结论证毕. □

5 上界的证明

本节的目标是为了证明第三节中给出的第一个不等式. 在此之前, 先对本节中出现的记号进行一些简单的说明. 任取 $\epsilon>0$. 我们待定 $\delta=\delta(\epsilon)$ 为一个比 ϵ 小很多的常数, 具体的要求将在后面的

计算中给出. 同时, 在本节中我们要求 $k > k_0(\epsilon)$ 为充分大的正整数, k 的要求在后面也是显而易见的.

我们先简单说明本节的基本手法. 为了计算随机图 G 含一个个数至少是 $m_{\max}$ 的完美匹配的概率, 我们试图直接运用引理 3 中的结论. 和证明下界的方法不同, 我们不能直接对 G 分块使用引理 3 中的结论, 因为我们失去了求和的不等式. 因此, 对于每个完美匹配 M, 我们试图将 M 切成一些大小合适的块, 计算 G 分别含这些块的概率和切法的数量, 再运用引理 3 的结论.

为了选取合适的切块大小, 我们选取正整数 $l = [k^{\frac{2}{3}}]$. 注意, k 充分大时,

$$al \times bl \geq ck, \delta^2(al + bl)\sqrt{abl^2} \leq \frac{k^{\frac{2}{3}}\delta}{6}.$$

因此, 若 $a\delta l \leq i \leq al$, $b\delta l \leq j \leq bl$, 则 (i, j) 符合引理 3 中的条件. 此外, 我们可以假设 $n > k^{\frac{7}{10}}$, 否则直接运用引理 3 中的结论即可. 因此, 我们有 $n >> l$.

我们先把 M 切成大小小于 $K_{al,bl}$ 的块. 我们的基本手法是, 从 M 中最左边的第一条边开始, 对于 M 中的一条边 uv, 考虑以其作为第一条边生成的 $K_{al,bl}$(若 M 中有一条边 uv, 则原图中以上部的点 $u, u+1, ..., al+u-1$ 和下部的点 $v, v+1, ..., bl+v-1$ 诱导的子图称为以 uv 作为第一条边生成的 $K_{al,bl}$), 然后将完全涵盖在此图中的边作为一块, 再选取 M 中剩下的下一条边选一块, 以此类推. 具体来说, 我们给每个大小为 $m_{\max}$ 的完美匹配 M 以下述的方式赋予一个数组 $T = (a_1, b_1, e_1, a_2, b_2, e_2, ..., a_q, b_q, e_q, q)$:

- a_1b_1 为 M 中的第一条边.
- 选取 a_ib_i 后, $a_{i+1}b_{i+1}$ 为下一条满足 $a_{i+1} - a_i \geq al$ 或 $b_{i+1} - b_i \geq bl$ 的边; 如果这样的边不存在, 则令 $i = q$.
- e_i 为 M 中从边 a_ib_i 到边 $a_{i+1}b_{i+1}$ 之间 (不包含 $a_{i+1}b_{i+1}$) 经过的边的数目. 特别地, e_q 为 M 中中从边 a_qb_q 起剩余的边的数目.

我们称 T 为 M 的类型. 记 $\mathcal{T}$ 为一切可以作为某个 M 的类型的这样的数组的集合, P_T 为随机图 G 包含一个类型为 T 的匹配的概率. 那么, 我们有

$$\mathbf{P}[L(\Sigma(K_{an,bn}; k)) \geq m_{\max}] \leq \sum_{T \in \mathcal{T}} P_T \leq |\mathcal{T}| \max_T P_T.$$

我们首先来对 $|\mathcal{T}|$ 进行放缩.

引理 4. 我们有

$$|\mathcal{T}| \leq \exp(C\frac{n}{l}\log l)$$

对某个常数 C 成立.

证明: 注意到对任意正整数 i, $a_{i+1} - a_i \geq al$ 或 $b_{i+1} - b_i \geq bl$ 至少有一个成立, 因此, 必有一种不等式至少对 $\frac{q}{2}$ 个 i 成立, 故 $\frac{q}{2}al \leq an$ 或 $\frac{q}{2}bl \leq bn$, 即 $q \leq \frac{2n}{l}$.

显然, 对固定的 q, $(a_1, a_2, ..., a_q)$ 和 $(b_1, b_2, ..., b_q)$ 均至多有 $\binom{(a+b)n}{q}$ 种取值, 而我们又显然有

$$c_1 + c_2 + ... + c_q = m_{\max} \leq (a+b)n,$$

故 $(e_1, e_2, ..., e_q)$ 也至多有 $\binom{(a+b)n}{q}$ 种取值. 因此,

$$|T| \leq n \max_q \binom{(a+b)n}{q}^3 \leq n \binom{n}{\frac{2(a+b)n}{l}}^3.$$

将不等式

$$\binom{n}{q} \leq (\frac{en}{q})^q$$

代入上式即得结论. 引理证毕. □

我们接下来对每个固定的 T, 对 P_T 进行放缩. 我们将证明下述的引理:

引理 5. 对任意的类型 $T \in \mathcal{T}$, 我们有

$$P_T \leq \exp(-c\epsilon^2 \frac{n}{\sqrt{k}})$$

对某个常数 $c > 0$ 成立.

证明: 补充定义 $a_{q+1} = an + 1, b_{n+1} = bq + 1$. 定义

$$a'_i = \min\{a_{i+1} - 1, a_i + al - 1\}, b'_i = \min\{b_{i+1} - 1, b_i + bl - 1\}.$$

并设 G_i 是 G 中由点 $a_i, ..., a'_i$ 和 $b_i, ..., b'_i$ 诱导的子图. 那么:

- G_i 显然两两不交.
- 由边 $a_{i+1}b_{i+1}$ 的定义, G_i 中包含了从 a_ib_i 到 $a_{i+1}b_{i+1}$ 之间 (不含后者) 的所有边, 即 e_i 条边.
- $r_i = a'_i - a_i + 1 \leq al$, $s_i = b'_i - b_i + 1 \leq bl$.

注意到 G_i 同构于 K_{r_i,s_i}, 故

$$P_T \leq \prod_{i=1}^{q} \mathbf{P}[L(\Sigma(K_{r_i,s_i}; k)) \geq e_i].$$

这里 G_i 就是我们切出的块. 该不等式成立是由于 G_i 两两无交, 故 $L(G_i) \geq e_i$ 这一族事件两两无关. 此时我们已经保证了 $r_i \leq al, s_i \leq bl$. 我们还要让 r_i, s_i 足够大才能使用引理 3. 为此, 我们把较小的 r_i 适当调大, 来保证引理 3 中的条件成立. 我们记

$$r'_i = \max\{r_i, a\delta l\}, s'_i = \max\{s_i, b\delta l\},$$

则

$$P_T \leq \prod_{i=1}^{q} \mathbf{P}[L(\Sigma(K_{r'_i,s'_i}; k)) \geq e_i].$$

这个不等式成立是由于显然的单调性. 此时我们可以对 K_{r_i,s_i} 使用引理 3 了. 为了使用我们的引理, 我们记 $t_i = 2(1+\delta)\sqrt{\frac{r_i's_i'}{k}}$. 则 $e_i \geq t_i$ 时, 我们可以对所选取的部分使用引理 3. 由对称性, 我们不妨设 $e_1, e_2, ..., e_t$ 是大于等于 A 的那些 e_i, 则由引理 3, 我们有

$$P_T \leq \prod_{i=1}^{t} \mathbf{P}[L(\Sigma(K_{r_i',s_i'}; k)) \geq e_i] \leq \prod_{i=1}^{t} 2e^{\frac{(e_i-t_i)^2}{8e_i}}.$$

即

$$-\ln P_T \geq \frac{1}{8}\sum_{i=1}^{t} \frac{(e_i-t_i)^2}{e_i} - q\ln 2.$$

而

$$\sum_{i=1}^{t} \frac{(e_i-t_i)^2}{e_i} = \sum_{i=1}^{t} e_i - 2\sum_{i=1}^{t} t_i + \sum_{i=1}^{t} \frac{t_i^2}{e_i} \geq \sum_{i=1}^{t} e_i - 2\sum_{i=1}^{t} t_i + \frac{(\sum_{i=1}^{t} t_i)^2}{\sum_{i=1}^{t} e_i},$$

因此, 若我们记 $E = \sum_{i=1}^{t} e_i, T = \sum_{i=1}^{t} t_i, L = E - T$,(注意, $e_i \geq t_i$, 故 $L \geq 0$), 则上式即

$$E - 2T + \frac{T^2}{E} = \frac{L^2}{T+L}.$$

由均值不等式,

$$\sum_{i=1}^{q} t_i \leq (1+\delta)\frac{1}{\sqrt{abk}}\sum_{i=1}^{q}(ar_i' + bs_i'),$$

且

$$\sum_{i=1}^{q} r_i' \leq qa\delta l + \sum_{i=1}^{q} r_i = an(1+2\delta), \sum_{i=1}^{q} s_i' \leq qb\delta l + \sum_{i=1}^{q} s_i = bn(1+2\delta),$$

故

$$T \leq \sum_{i=1}^{q} t_i \leq 2(1+2\delta)\frac{n}{\sqrt{k}},$$

且

$$L = E - T \geq \sum_{i=1}^{q}(e_i - t_i) = m_{\max} - \sum_{i=1}^{q} t_i \geq 2(\epsilon - 2\delta)\frac{n}{\sqrt{k}}.$$

故

$$\frac{L^2}{T+L} = L \times \frac{1}{\frac{T}{L}+1} \geq \frac{2(\epsilon-2\delta)^2}{\epsilon+1} \times \frac{n}{\sqrt{k}} \geq \frac{n\epsilon^2}{\sqrt{k}}.$$

而 $q \leq \frac{2n}{l}$, 故

$$-\ln P_T \geq c\frac{n\epsilon^2}{\sqrt{k}}$$

对某个实数 $c > 0$ 成立. 结论得证. □

因此, 结合引理 4 和引理 5, 我们知

$$|\mathcal{T}| \max_T P_T \leq \exp(\Omega(\epsilon^2\frac{n}{\sqrt{k}}))$$

对充分大的正整数 k 成立. 这就完全证明了上界部分. □

参考文献 [1]. Marcos Kiwi and Martin Loebl and Jirı Matousek, Expected length of the longest common subsequence for large alphabets, Advances in Mathematics, 197:480-498, November 2004.