arXiv:2303.03086v1 [cs.FL] 6 Mar 2023

基于交换与不匹配距离的等距词 *

M. Anselmo[1] G. Castiglione[2] M. Flores[1]D. Giammarresi[3] M. Madonia[4] S. Mantaci[2]

[1] 意大利萨莱诺大学计算机系。{manselmo, mflores }@unisa.it[2] 意大利巴勒莫大学数学与计算机系{ giuseppa.castiglione, sabrina.mantaci }@unipa.it[3] 意大利罗马第二大学“Tor Vergata”数学系。giammarr@mat.uniroma2.it[4]意大利卡塔尼亚大学数学与计算机系。madonia@dmi.unict.it

摘要

编辑距离是一种用于度量词之间差异的度量，它通过计算将一个词变换为另一个词所需的编辑操作次数来量化这种差异。若对于任意一对 f-自由词 u 和 v，都存在一条从u 到 v 的最短变换路径，且该路径通过相关编辑操作实现，并且所有中间词也都是 f-自由的，则称词 f 相对于某种编辑距离是等距的。“等距”这一形容词源于这样一个事实：如果考虑汉明距离（即仅考虑不匹配），那么等距词便与超立方体等距子图的定义相关联。

我们考虑含交换与不匹配的编辑距离情形。我们将其与仅含不匹配的情形进行比较，并证明等距词的一些性质，这些性质与其重叠的特定特征有关。

关键词：交换与不匹配距离、等距词、含错重叠。

1 引言

编辑距离是计算机科学许多领域中的核心概念。它在定义字符串族的组合性质，以及设计许多经典字符串算法方面发挥着至关重要的作用；这些算法应用于自然语言处理、生物信息学以及一般的信息检索问题。编辑距离是一种字符串度量，用于量化两个字符串之间的差异，其基础是计算将一个字符串转换为另一个字符串所需的最少编辑操作次数。

*本研究部分由 INdAM-GNCS 2022和2023项目、萨莱诺大学 FARB ORSA229894项目、卡塔尼亚大学 TEAMS项目，以及授予罗马第二大学数学系的 MIUR卓越系项目 MatMod@TOV 资助。

编辑距离的不同定义采用不同的编辑操作集合。在字符串中插入、删除和替换字符的操作，定义了莱文斯坦距离，这可能是最广为人知的编辑距离（参见 [17]）。另一方面，最基本的编辑距离是汉明距离，它仅适用于长度相同的一对字符串，并统计其中存在失配的位置；这相当于限制只能使用替换操作。因此，汉明距离在检测和纠正字符串中的错误方面有着直接应用，也是带失配的字符串匹配算法中的重要组成部分（见 [13]）。

等距词（或字符串）的概念将编辑距离与这样一种性质结合起来：一个词不会作为因子出现在其他词中。需要注意的是，这一性质在组合学研究以及 DNA序列相似性或距离的研究中都十分重要，其中被规避的因子被称为缺失词 [8, 9, 10]。基于汉明距离的等距词最早在 [15] 中被引入，被定义为一类特殊的二进制字符串，它们不会作为因子出现在某些字符串变换中。如果一个字符串不包含 f 作为因子，则称其为f-自由 。若对于任意一对 f-自由词 u 和 v，都存在一系列符号替换操作，将u 变换为 v，且所有中间词也都是 f-自由的，则词 f 是等距的。

等距词与超立方体的等距子图定义相关，这类子图称为广义斐波那契立方体。超立方体图 Q_n 是一个图，其顶点是长度为 n的（二进制）词；当对应的词恰好在一个符号上不同时，两个顶点相邻。因此，两个顶点之间的距离就是对应顶点词的汉明距离。令 $Q_n(f)$表示 Q_n 的子图，该子图仅包含 f-自由的顶点。那么，如果 f 是等距的， $Q_n(f)$中顶点之间的距离与在整个Q_n中计算得到的距离相同。斐波那契立方体由 Hsu 于 [14]引入，对应于 $f = 11$的情形。在 [15, 16, 19, 20, 22]中，字母表大小为 2 且距离为汉明距离时，非等距词的结构得到了完整刻画，并与其重叠上的特定性质联系起来。字母表大小大于 2 且距离为李距离的更一般情形在 [3, 4, 5]中进行了研究。利用这些刻画， [7]给出了一些线性时间算法，用于检验二进制词是否为汉明等距词，以及四元词是否为李等距词。这些算法经过扩展，可以进一步提供有关非等距词的信息，同时仍保持[4]中的线性复杂度。二进制汉明等距二维词也已在 [6]中得到研究。

字符串纠错中的许多难题都源自计算生物学。在 DNA序列的染色体操作中，对于基因突变和复制，考虑交换操作似乎是顺理成章的；该操作由交换两个相邻符号构成。Damerau-Levenshtein距离也将交换纳入所有编辑操作之中。在 [18], 中，Wagner证明，包含插入和交换的编辑距离是NP难的，而每种单独的情形都可以在多项式时间内求解。此外，包含插入、删除、替换和交换的一般编辑距离可以在多项式时间内求解。 [1, 12],中研究了交换匹配问题，并在 [2, 11]中给出了计算相应编辑距离的算法。

本文研究基于交换和失配的编辑距离下等距词的概念。我们用波浪号 符号表示这种距离，这个符号在某种程度上让人联想到交换操作。等长词 u 和 v 之间的波浪号距离 $\mathrm{dist}_{\sim}(u, v)$，是将 u 变换为 v所需的替换和交换操作的最少次数。由此很自然地得出波浪号等距 词的定义。如果对于任意一对等长且不含 f的词 u 和 v，都存在一种将 u变换为 v的变换，且该变换恰好使用 $\mathrm{dist}_{\sim}(u, v)$ 次替换

和交换操作，并且所有中间词仍然避开 f，那么结论便成立。事实证明，在定义中加入交换操作会使情况变得更加复杂，但也让它在应用中更有趣。这并非对汉明字符串等距性的简单推广，因为会出现一些特殊情况。事实上，一次交换操作等价于两次替换，但在计算波浪号距离时只计作一次。此外，u 变换为 v 可能有不同的方式，因为对于某些连续符号三元组，可以从左到右先交换再替换，也可以先替换再交换。我们给出了一些波浪号等距词的例子，它们并非汉明等距词，反之亦然。根据定义，要证明给定的字符串 f 不是波浪号等距词，就应举出一对不含 f 的词（$\tilde{\alpha}, \beta$），使得从 $\tilde{\alpha}$ 到 β 的任何长度为 $\mathsf{dist}_{\sim}(\tilde{\alpha}, \beta)$ 的变换都必须经过包含 f 的词。这对词称为 f 的波浪号见证。我们基于错误重叠的概念，证明了判定 f 为非等距词的一些必要条件，并在许多情况下给出了波浪号见证的显式构造。

2 预备知识

设 Σ 是一个有限字母表。长度为 $|w| = n$ 的词（或字符串）w 是 $w = a_1 a_2 \cdots a_n$，其中 a_1，a_2，…，a_n 是 Σ 中的符号。所有由 Σ 构成的词的集合记作 Σ^*，所有由 Σ 构成且长度为 n 的词的集合记作 Σ^n。最后，ϵ 表示空词，而 $\Sigma^+ = \Sigma^* - \{\epsilon\}$.。对于任意词 $w = a_1 a_2 \cdots a_n$，其反转是词 $w^{rev} = a_n a_{n-1} \cdots a_1$。若 $x \in \{0, 1\}$，我们用 $\bar{x}$ 表示 x 的相反词，即当 $x = 0$ 时，$\bar{x} = 1$，反之亦然。接着，我们定义词 $\bar{w} = \bar{a_1}\bar{a_2} \cdots \bar{a_n}$ 的取补为 $\bar{w}$。

令 $w[i]$ 表示 w 在位置 i 处的符号，即 $w[i] = a_i$。那么，对于 $1 \le i \le j \le n$，$w[i..j] = a_i \cdots a_j$ 是 w 的一个因子。长度为 l 且满足 $1 \le l \le n-1$ 的 w 的前缀（分别为后缀）是 $\mathsf{pre}_l(w) = w[1..l]$（分别为 $\mathsf{suf}_l(w) = w[n-l+1..n]$）。当 $\mathsf{pre}_l(w) = \mathsf{suf}_l(w) = u$ 时，u 在此称为 w 的一个长度为 l 的重叠；它也称为边界或双前后缀。若词 w 满足 f-自由，则表示 w 不含作为因子的 f。

一个编辑操作是一个函数 $O: \Sigma^* \to \Sigma^*$，它将一个词转换为另一个词。最常见的编辑操作包括插入、删除或替换一个字符，以及交换两个相邻字符。设 OP 为一个编辑操作集合。两个词 $u, v \in \Sigma^*$ 的编辑距离是 OP 中将 u 转换为 v 所需的最少编辑操作数。. 本文中，我们考虑仅使用替换和交换的编辑距离。注意，这两种操作不会改变词的长度。下面给出形式化定义。

定义 1 设 Σ 为一个有限字母表，$w = a_1 a_2 \ldots a_n$ 为 Σ 上的一个词。在位置 i 处以 $x \in \Sigma$ 替换 w 中的字符 $x \neq a_i$，称为替换操作（简称替换），定义如下：

$$R_{i,x}(a_1 a_2 \ldots a_{i-1} a_i a_{i+1} \ldots a_n) = a_1 a_2 \ldots a_{i-1} x a_{i+1} \ldots a_n.$$

交换操作（简称交换）在 w 的位置 i 处交换位置 i 和 $i+1$ 上的字符，前提是这两个字符不同，$a_i \neq a_{i+1}$，

$$S_i(a_1 a_2 \ldots a_i a_{i+1} \ldots a_n) = a_1 a_2 \ldots a_{i+1} a_i \ldots a_n.$$

当字母表 $\Sigma = \{0, 1\}$ 在给定位置 i 只有一种可能的替换时，因此我们写作 $R_i(w)$，而不写作 $R_{i,\,x}(w)$。

给定两个等长的词 $u = a_1 \cdots a_n$和 $v = b_1 \cdots b_n$，如果 $a_i \neq b_i$，则它们在位置 i 有一个失配错误（或失配）；如果 $a_i a_{i+1} = b_{i+1} b_i$，则它们在位置 i 有一个交换错误（或交换），其中 $a_i \neq a_{i+1}$。如果 u 和 v 在位置 i 有失配或交换错误，我们就说它们在该位置有错误。

请注意，一次交换对应两个相邻的失配。

如果对于任意一对 f-自由词 u 和 v，都存在一个长度最短的替换操作序列，可将 u 变换为 v，且所有中间词也都是 f-自由的，那么词 f 就是等距的。本文将这种等距定义称为汉明等距。在 [21], 中，如果存在 l，使得 $\mathrm{pre}_l(w)$和 $\mathrm{suf}_l(w)$有两个不匹配错误，那么词 w 就有二错误重叠。接着，他们证明了以下刻画。

命题 2 词 f 是汉明等距的，当且仅当 f 有二错误重叠。

3 波浪号距离与波浪等距词

本节考虑基于交换和替换操作的编辑距离，我们称之为波浪号距离，记作 $dist_{\sim}$。首先，我们给出一些定义和记号，并举出一些例子，同时证明一些初步性质。

定义 3 设 u、$v \in \Sigma^*$ 是长度相等的词。u 与 v 之间的波浪号距离 $\mathrm{dist}_{\sim}(u, v)$是将 u 变换为 v所需的最少替换和交换次数。

定义 4 设 $u, v \in \Sigma^*$ 是长度相等的词。从 u 到 v 的长度为 h 的波浪变换 τ是一个词序列$(w_0, w_1, \ldots, w_h)$，满足 $w_0 = u, w_h = v$，且对任意 $k = 0, 1, \ldots, h-1$，$\mathrm{dist}_{\sim}(w_k, w_{k+1}) = 1$。此外，若对任意$i = 0, 1, \ldots, h$，词 w_i 是 f-***free***，则 τ 是 f-free。

从 u 到 v 的波浪变换 $(w_0, w_1, \ldots, w_h)$ 与一系列 h 个操作 $(O_{i_1}, O_{i_2}, \ldots O_{i_h})$ 相关联，并且满足：对任意 $k = 1, \ldots, h$，$O_{i_k} \in \{R_{i_k,x}, S_{i_k}\}$ 和 $w_k = O_{i_k}(w_{k-1})$；它可以表示如下：

$$u = w_0 \xrightarrow{O_{i_1}} w_1 \xrightarrow{O_{i_2}} \cdots \xrightarrow{O_{i_h}} w_h = v.$$

稍微滥用一下记号，在下文中，我们将把波浪变换既称为词序列，也称为操作序列。下面给出一些例子。

例5 设 $u = 1011, v = 0110$。下面展示从 u 到 v的两种不同波浪变换。注意，τ_1 的长度对应于 $\mathrm{dist}_{\sim}(u, v) = 2$。

$$\tau_1 : 1011 \xrightarrow{S_1} 0111 \xrightarrow{R_4} 0110 \qquad \tau_2 : 1011 \xrightarrow{R_1} 0011 \xrightarrow{R_2} 0111 \xrightarrow{R_4} 0110$$

此外，考虑将 $u' = 100$变换为 $v' = 001$的以下波浪变换：

$$\tau_1' : 100 \xrightarrow{S_1} 010 \xrightarrow{S_2} 001 \qquad \tau_2' : 100 \xrightarrow{R_1} 000 \xrightarrow{R_2} 001$$

注意，τ_1'和 τ_2'的长度相同，等于 $\mathrm{dist}_{\sim}(u', v') = 2$。有趣的是，在 τ_1' 中，位置 *2* 上的符号被改变了两次。

下一个引理表明，对于双字母字母表的情形，我们可以将范围限制为每个字符至多改变一次的波浪号变换。

引理6 设 $u, v \in \{0,1\}^m$ 满足 $m \geq 1$。那么，存在一个将 u 变换为 v 的波浪变换，其长度为 $\mathrm{dist}_{\sim}(u,v)$，使得对于任意 $i = 1$、2、.....、m，位置 i 上的字符至多改变一次。

证明： 设 $u = a_1 \cdots a_m$ 和 $v = b_1 \cdots b_m$，并设 τ 是一个将 u 变换为 v、长度为 $d = \mathrm{dist}_{\sim}(u,v)$ 的波浪变换。假设对于某个 i，位置 i 上的字符被 τ 改变了不止一次，并令 O_t 和 O_s 分别为首次和第二次修改位置 i 上字符的操作。注意，位置 i 上的字符可以由操作 R_i、S_{i-1} 或 S_i 改变。

假设 $O_t = S_i$ 和 $O_s = R_i$。那么，符号 a_i 被改变了两次，操作 S_i 和 R_i 可以替换为一个 R_{i+1}。这样就会得到一个将 u 变换为 v、长度严格小于 d 的波浪变换；这与波浪号距离的定义矛盾。对于 $O_t = R_i$ 和 $O_s = S_i$、$O_t = S_{i-1}$ 和 $O_s = R_i$、$O_t = R_i$ 和 $O_s = S_{i-1}$ 的情形，同理可证。

最后，如果 $O_t = S_{i-1}$ 且 $O_s = S_i$，那么位置 $i-1$、i 和 $i+1$ 上的三个字符都会改变，但位置 i 上的字符会改变两次。因此，交换操作 S_{i-1} 和 S_i 可以替换为 R_{i-1} 和 R_{i+1}，从而得到一个将 u 变为 v 的波浪变换，其长度相同，为 d，但只涉及位置 $i-1$ 和 i 一次（见例5中的 τ_2'）。 □

注释7 引理6 仅适用于二元字母表。事实上，如果 $\Sigma = \{0,1,2\}$，并取 $u = 012$ 和 $v = 120$，那么 距离$_{\sim}(012,120) = 2$，因为存在波浪变换 $012 \xrightarrow{S_1} 102 \xrightarrow{S_2} 120$。相反，若要使每个字符至多改变一次，则需要三次替换操作。

定义8 设 Σ 是一个有限字母表，且 u、$v \in \Sigma^+$。从 u 到 v 的波浪变换若其长度等于 距离$_{\sim}(u,v)$，且每个位置上的字符至多修改一次，则称其为 最小的 。

引理6保证，在二元情形下，最小波浪变换总是存在的。在下文中，这将是研究最多的情形。现在，我们根据交换与不匹配距离来定义等距词。

定义 9 若对于任意一对长度为 $m > n$ 的 f-自由词 u 和 v，都存在一个从 u 到 v 的最小波浪变换，且该变换是 f-自由的，则称带有 $n \geq 1$、f 的 $f \in \Sigma^n$ 是波浪等距的。若它不是波浪等距的，则称其为波浪号非等距的。

要证明一个词是波浪号非等距的，只需举出一对与定义 9 矛盾的词（u，v）。这样的一对词将称为 f 的波浪号见证。下面给出一些例子。

定义 10 Σ^m 中的一对词（u、v）是 f 的一对波浪号见证，当且仅当：

1. u 和 v 是 f-自由的 *2.* $\mathrm{dist}_{\sim}(u,v) \geq 2$ *3.* 不存在从 u 到 v 且为 f-自由的最小波浪号变换。

例 11 词 $f = 1010$ 是波浪号非等距的，因为 $u = 11000$ 和 v =10110 是 f的波浪号见证。事实上，从 u到 v 唯一可能的最小波浪号变换是 $11000 \xrightarrow{S_2} 10100 \xrightarrow{R_4} 10110$ 和 $11000 \xrightarrow{R_4} 11010 \xrightarrow{S_2} 10110$，且在这两种情况下，1010 都会在第一步之后作为因子出现。

注12 当一个变换包含彼此相邻的交换和替换时，可能存在许多不同的最小波浪号变换，它们涉及不同的操作集合。例如， $u = 010$ 和 $v = 101$的对(u, v)具有如下最小波浪号变换：

$$010 \xrightarrow{S_1} 100 \xrightarrow{R_3} 101 \qquad 010 \xrightarrow{S_2} 001 \xrightarrow{R_1} 101$$

当只允许替换时，这种情况不会发生。因此，研究波浪等距词比汉明情形更复杂。

例 11 展示了一个波浪号非等距词。要证明给定的词是波浪等距词则困难得多，因为这要求给出不存在波浪号见证的证据。现在我们将用特设的技术证明词**111000**是等距的。

例 13 词 $f = 111000$ 是波浪等距词。反设 f是波浪号非等距词，并令(u, v) 为 f 的一对具有最小波浪号距离的波浪号见证。若 u 和 v 仅有不匹配错误，这就是汉明距离的情形，该理论的结果 *[4, 19]* 表明 $u = 1110\mathbf{01}1000$ 和 $v = 1110\mathbf{10}1000$；它们不是波浪号见证，因为$\mathrm{dist}_{\sim}(u, v) = 1$.
因此， u 和 v 在某个位置 i发生了交换错误；不妨设 $u[i..i + 1] = 01$。$\mathrm{dist}_{\sim}(u, v)$ 的最小性意味着， $S_i(u)$ 并非 f-自由的。于是，因子111000出现在 $S_i(u)$ 中，位置从 $i - 2$开始，并且 $u[i - 2..i + 3] = 110100$。由于 v 是 f-自由的，因此 u 中还存在另一个错误，涉及 $[i - 2..i + 3]$中的某些位置。它既不可能是交换（因为不存在尚未改变的相邻不同符号），也不可能是位置 $i - 1$、 $i + 2$ 上的失配（因为相应的替换无法使 f 出现）。因此，它是位置 $i - 2$ 或 $i + 3$上的失配。考虑位置 $i+3$ 上的失配情形（另一种情形类似）。于是， $u[i+3..i+8] = 011000$，并且 $[i+4..i+8]$中还存在另一个错误，事实上位于位置 $i+6$ 或 $i+8$。继续进行类似推理，最终会回到前一种情形。这是矛盾的，因为 u 的长度是有限的。

从现在起，我们仅考虑二元字母表 $\Sigma = \{0, 1\}$ ，并研究以 **1** 开头的等距二元词，因为下面的引理可由结合定义轻易推得。

引理14 设 $f \in \{0, 1\}^n$。以下陈述等价：

1. f 是波浪等距的 2. f^{rev} 是波浪等距的
3. $\underline{f}$ 是波浪等距的。

我们以比较波浪等距词与汉明等距词来结束本章节。尽管波浪号距离比汉明距离更一般，但二者不可比较，正如下述命题所述。

命题 15 存在一个波浪等距但汉明非等距的词，也存在一个波浪号非等距但汉明等距的词。

证明： 词 $f = 111000$ 是波浪等距的（见例 13），但根据命题 2，f 是汉明非等距的。

反之，$f' = 1010$ 是波浪号非等距的（见例 11），但根据命题 2，它是汉明等距的。

□

4 波浪等距词与波浪错误重叠

在本章节中，我们将重点研究具有波浪号非等距这一词性质的词，并将其与这些词的重叠中的错误数量联系起来。这一思路使人想起命题 2 中回顾的汉明等距词的刻画，但正如命题 15 所指出的，交换操作改变了所有视角。

定义 16 设 $f \in \{0,1\}^n$。那么，f 具有一个长度为 l 的 q-波浪错误重叠，其中 $1 \le l \le n-1$ 和 $0 \le q \le l$，若 $\mathrm{dist}_{\sim}(\mathrm{pre}_l(f)、\mathrm{suf}_l(f)) = q$。

换言之，如果 f 具有长度为 l 的 q-波浪误差重叠，那么就存在一个从 $\mathrm{pre}_l(f)$ 到 $\mathrm{suf}_l(f)$、长度为 q 的最小波浪号变换 τ。下文中，当 $q = 2$ 时，为了说明错误的类型，若 τ 由两次替换构成，则称 2-波浪误差重叠为 RR 型；若由两次交换构成，则称为 SS 型；若由一次替换和一次交换构成，则称为 RS 型；若由一次交换和一次替换构成，则称为 SR 型。如果两个错误分别位于位置 i 和 j，且 $i < j$，我们称 f 在 i 和 j 处具有 2-波浪误差重叠；等价地，称 i 和 j 为该 2-波浪误差重叠的 误差位置。

设 $f \in \{0,1\}^n$ 在位置 i 和 j 处具有 2-波浪误差重叠，其中 $i < j$，位移为 r，长度为 $l = n - r$。对于某些 w_1、w_2、w_3、$w_4 \in \{0,1\}^*$ 和 $|w_1| = r$，可能出现以下情况（见图 1）。

RR：$f = w_2\boldsymbol{f}[i]w\boldsymbol{f}[j]w_3w_4 = w_1\overline{w_2\boldsymbol{f}[i]}\overline{w\boldsymbol{f}[j]}w_3$SR：
$f = w_2\boldsymbol{f}[i]\boldsymbol{f}[i+1]w\boldsymbol{f}[j]w_3w_4 = w_1w_2\boldsymbol{f}[i+1]\boldsymbol{f}[i]w\boldsymbol{f}[j]\cancel{w_3}$RS：
$f = w_2\boldsymbol{f}[i]w\boldsymbol{f}[j]\boldsymbol{f}[j+1]w_3w_4 = w_1w_2\boldsymbol{f}[i]\underline{w\boldsymbol{f}}[j+1]\boldsymbol{f}[j]w_3$SS：
$f = w_2\boldsymbol{f}[i]\boldsymbol{f}[i+1]w\boldsymbol{f}[j]\boldsymbol{f}[j+1]w_3w_4 = w_1w_2\boldsymbol{f}[i+1]\boldsymbol{f}[i]w\boldsymbol{f}[j+1]\boldsymbol{f}[j]w_3$

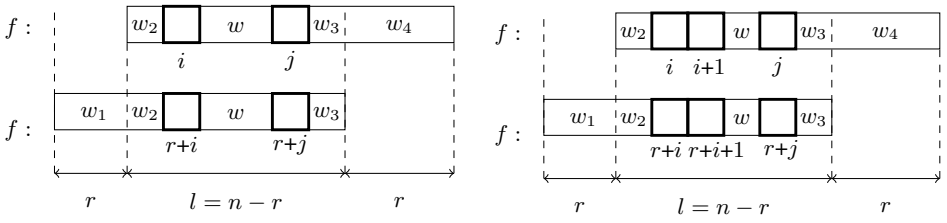


图 1：词 f 及其 RR 型（a）和 SR 型 2-波浪误差重叠

如果 $w=\epsilon$ 我们称这两个错误为相邻错误。具体而言，对于位置 i 和 j上的 RR 型和 RS 型 2-tilde-错误重叠，当$j=i+1$时，这两个错误是相邻的。注意，对于具有 $f[i]\neq f[i+1]$ 的相邻 RR 型错误，我们有一个属于交换的 1-tilde-错误重叠，并将其称为 S 型。对于位置 i和 j上的 SR 型和 SS 型 2-tilde 错误重叠，当 $j=i+2$时，这两个错误是相邻的。

注 17 设 $f\in\{0,1\}^n$ 是一个波浪号非等距词，且(u,v)（其中 $u,v\in\Sigma^m$）是 f 的一对波浪号见证，在所有长度为 m 的波浪号见证对中，u 与 v 的 $d=$距离$_\sim(u,v)$ 最小。

设 $\{O_{i_1},O_{i_2},\ldots,O_{i_d}\}$ 是从 u 到 v, $1\leq i_1<i_2<\cdots<i_d\leq m$ 的最小波浪变换的操作集合。那么，对于任意 $j=1,2,\ldots,d-1$,$O_{i_j}(u)$ 都在区间 $[k_j..(k_j+n-1)]$，中出现，而该区间至少包含一个由 O_{i_j} 修改的位置。事实上，如果 O_{i_j} 是交换操作，那么它会同时改变两个位置，即位置 i_j 和 i_j+1，而区间 $[k_j..(k_j+n-1)]$ 可能包含两个位置，也可能只包含其中一个。请注意，当区间中只包含一个位置时，该位置位于区间边界。这意味着，尽管边界位置上的错误看起来可能是由替换造成的，但实际上也可能是由涉及边界两侧位置的隐藏交换造成的。

命题 18 如果 $f\in\{0,1\}^n$ 是波浪号非等距的，那么

1. 要么 f 有一个 *S* 型 *1-tilde*-错误重叠；
2. 要么 f 有一个2-*tilde*-错误重叠。

证明：设 f是一个波浪号非等距词，(u,v) 是 f的一对波浪号见证，而 $\{O_{i_1},O_{i_2},\ldots,O_{i_d}\}$ 如注 17 所述。那么，对于任意 $j=1$，2，...，$d-1$，$O_{i_j}(u)$ 在区间 $[k_j..k_j+n-1]$中有 f 的一次出现，其中至少包含一个被 O_{i_j}修改的位置。注意，这次 f 的出现必须在从 u 到 v的波浪号变换中消失，因为 v 是 f-自由的。因此，区间 $[k_j..k_j+n-1]$中包含一个由 $\{O_{i_1},O_{i_2},\ldots,Oi_d\}$中的另一个操作修改的位置。根据鸽笼原理，存在 s、$t\in\{i_1,i_2,\ldots i_d\}$，使得 $O_s(u)$ 在 $[k_s..k_s+n-1]$中有 f的一次出现，且其中至少包含一个被 O_t修改的位置；而 $O_t(u)$ 在 $[k_t..k_t+n-1]$ 中有 f的一次出现，且其中至少包含一个被 O_s修改的位置。不失一般性，假设 $k_s<k_t$。$[k_s..k_s+n-1]$与 $[k_t..k_t+n-1]$ 的交集在 $O_t(u)$中截取了 f 的一个前缀，并在 $O_s(u)$中截取了 f 的一个后缀，其长度为 l。这样的交集可能包含 O_s 和 O_t所修改的位置中的两个、三个或四个，其中至少一个由 O_s 修改，且至少一个由 O_t修改。

考虑这样一种情况：$[k_s..k_s+n-1]$ 与 $[k_t..k_t+n-1]$的交集中包含 O_s 和 O_t 修改的位置中的两个，记为 i 和 j，其中 $1\leq i<j\leq l$。如果这些位置不相邻，则 f 存在 2-波浪误差重叠（RR 型）。否则，若 $f[i]\neq f[i+1]$，则 f 存在 1-tilde-错误重叠（S 型）。若 $f[i]=f[i+1]$，则 f 存在 2-波浪误差重叠（RR 型）。

假设 $[k_s..k_s+n-1]$ 与 $[k_t..k_t+n-1]$ 的交集包含 O_s 和 O_t 所修改位置中的三个。在这种情况下，这两个操作中至少有一个必须是交换；假设 O_s 是一次交换。那么，O_t 可以是对第三个位置进行替换，也可以是交换（如果第三个位置位于 $[k_t..k_t+n-1]$ 的边界）。无论哪种情况，f 都有一个 2-波浪误差重叠（SR 型或 SS 型）。

现在假设 $[k_s..k_s+n-1]$ 与 $[k_t..k_t+n-1]$ 的交集包含 O_s 和 O_t 所修改位置中的四个。在这种情况下， O_s 和 O_t 各涉及两个位置，而 f 有一个 SS 型 2-波浪误差重叠。

□

5 波浪见证的构造

如第 4 节所述，要证明一个词是波浪号非等距的，只需给出一对波浪号见证。命题 18 指出，如果一个词是波浪号非等距的，那么它要么有一个交换型 1-tilde-错误重叠，要么有一个 2-tilde-错误重叠。本节将说明如何从一个词的错误重叠出发构造波浪号见证。我们先从 1-tilde-错误重叠的情况开始。

命题 19 如果 f 有一个1-*tilde*-错误重叠，且类型为 *S*，那么它是波浪号非等距的。

证明： 设 f 在位置 i 有一个位移为 r、类型为 S 的 1-tilde-错误重叠。数对 $(u,\ v)$:

$$u = \mathrm{pre}_r(f)R_i(f) \qquad v = \mathrm{pre}_r(f)R_{i+1}(f)$$

是 f的一对波浪号见证。事实上，可以证明它们满足定义10中的条件。

□

例 20 词 $f = 101$ 在位置 1具有一个 *S* 型1-波浪误差重叠，因此它是波浪号非等距的。事实上，按照前一个命题的证明，数对 (u, v) （其中 $u = 1001$ 和 $v = 1111$ ）是一对波浪号见证。

现在我们来引入一些常用作波浪号见证的特殊词。设 $f \in \{0,1\}^n$ 在位置 i 和 j具有位移为 r 的 2-波浪误差重叠，则

$$\tilde{\alpha}_r = \mathrm{pre}_r(f)O_i(f) \text{ and } \tilde{\beta}_r = \mathrm{pre}_r(f)O_j(f) \tag{1}$$

例如，沿用前文对 SR 型错误的记号，我们有

$$\tilde{\alpha}_r(f) = w_1w_2f[i{+}1]f[i]wf[j]w_3w_4 \text{ and } \tilde{\beta}_r(f) = w_1w_2f[i]f[i{+}1]w\overline{f[j]}w_3w_4 \tag{2}$$

引理 21 设 $f{\in}\{0,1\}^n$ 具有位移为 r的2-波浪误差重叠，则 $\tilde{\alpha}_r(f)$ 是 f-自由的。

证明：假设 $f \in \{0,1\}^n$ 具有一个 2-波浪误差重叠。如果它是 RR 型，那么根据 [19], 中引理 2.2 的断言 1，即使在相邻错误的情况下，$\tilde{\alpha}_r(f)$ 也是 f-自由的。如果它是 SR 型，位移为 r，位于位置 i 和 j，且 $i < j$，那么根据公式 (1)，对于任意 $1 \le k \le n$，满足 $k \ne i$ 和$k \ne i+1$，我们有 $\tilde{\alpha}_r(f) = w_1S_i(f)$，于是 $\tilde{\alpha}_r[r+k] = f[k]$。如果 f在 $\tilde{\alpha}_r$ 中出现在位置 r_1+1，则有 $1 < r_1 < r$（如果 $r_1 = 1$，那么$f[i] = f[i+1]$，且位置 i 处不存在交换错误），并且 $\tilde{\alpha}_r[r_1+1\ldots r_1+n]$ =$f[1\ldots n]$。最后，根据公式 (2)，对于 $k \ne r+j$，有 $\tilde{\alpha}_r[k] = f[k]$。总之，我们有 $f[i] = \tilde{\alpha}_r[r_1+i] = f[r_1+i]$ （显然，$r_1+i \ne r+j$）。此外，$f[r_1{+}i] = \tilde{\alpha}_r[r{+}r_1{+}i]$（$r_1{+}i \ne i$和 $r_1{+}i \ne i{+}1$，因为 $r_1 > 1$）。但是 $\tilde{\alpha}_r[r+r_1+i] = f[r+i]$，于是我们得到矛盾 $f[i] = f[r+i]$。如果该 2-波浪误差是 RS 型或 SS 型，证明类似。为清楚起见，请注意，即使在相邻错误的情况下，假设 f 出现在 $\tilde{\alpha}_r$ 中，也会在 $f[i]$中导致矛盾，而该矛盾不受 j影响。

□

~ ~注意，虽然 $\tilde{\alpha}_r$ 总是 f-自由的，但 β_r 并非如此。事实上，性质 β_r 不是 f-自由的，这与关于 f的重叠的一个条件有关。我们给出如下定义。

定义 22 设 $f \in \{0,1\}^n$，并考虑 f 的一个2-波浪误差重叠，其位移为 r，误差位置为 i、j，且 $1 \le i < j \le n-r$。该2-波浪误差重叠若为 ***RR*** 型或 ***SS*** 型，则满足$Condition^{\sim}$：

$$\begin{cases} r \ is \ even \\ j - i = r/2 \\ f[i..(i+r/2-1)] = f[j..(j+r/2-1)] \end{cases} \qquad (\textbf{\textit{Condition}}^{\sim})$$

~**引理 23** 设 $f \in \{0,1\}^n$ 有一个位移为 r 的2-波浪误差重叠，则 $\beta_r(f)$ 不是f-自由的，当且仅当该2-波浪误差重叠满足 $Condition^{\sim}$。

证明： 假设 $f \in \{0,1\}^n$ 有一个满足 $Condition^{\sim}$ 的 2-波浪误差重叠。注意，RS 型或 SR 型的 2-波浪误差重叠不可能满足 $Condition^{\sim}$。现在，~若该 2-波浪误差重叠为 RR 型，则可如 [19] 中引理 2.2 的断言 2 的证明所示，证明 $\beta_r(f)$ 不是 f-自由的。若该 2-波浪误差重叠为 SS 型，则必须对该证明作适当修改。更确切地说，设 i, j，其中$1 \le i < j \le n-r$，是位移为 r 且满足 $Condition^{\sim}$ 的 2-波浪误差重叠的误差位置。

令 $f[i] = f[j] = x$、$f[i+r] = f[j+r] = x$、¯$f[i+1] = f[j+1] = x$ 和¯$f[i+1+r] = f[j+1+r] = x$。

可以证明，对于某个 k_1，$k_2 \ge 0$，我们可以写成

$$f = \rho(uw)^{k_1} uwuw\overline{u}w\overline{u}(w\overline{u})^{k_2}\sigma$$

其中 $u = x\overline{x}$、$w = f[i+2..j-1]$（若 $j = i+2$，则 w 为空），并且 ρ 和 σ 分别是 w 的后缀和前缀。现在，我们有

$$\tilde{\beta}_r(f) = \rho(uw)^{k_1+1} uwuw\overline{u}w\overline{u}(w\overline{u})^{k_2+1}\sigma$$

~，因此，β_r（f）不是 f-自由的。

现在假设~不是 f-自由的，并设 f 的一个副本出现在 β_r（f）中的位置 r_1+1 处。类似于引理 21 证明中所采用的推理可知，若 i 和 j 是误差位置，则 $j-i = r_1$ 和 $j-i = r-r_1$。因此，$r = 2r_1$ 为偶数且 $j-i = r/2$。所以，对于 $0 \le t \le r/2$，有 $f[i+t] = f^b[i+t] = f[i+t+r/2] = f[j+t]$，即 $f[i..$（$i+r/2-1$）$] = f[j..$（$j+r/2-1$），并且 2-波浪误差重叠满足 $Condition^{\sim}$。

□

在本章节的其余部分，我们讨论 2-波浪误差重叠情形下波浪号见证的构造。我们区分非相邻错误和相邻错误两种情况。非相邻错误可以用标准技术处理，而相邻错误的情况可能会出现新的问题。例如，当 f 存在一个类型为 SR、错误块为**101**（与 010 对齐）的 2-错误重叠时，也可以将其视为类型 RS。

此外，请注意，所有相邻的错误均可列举如下（取补和反转的情形除外）。S 型的二错误重叠可能具有（错误）块 **1010** 或**1001**；SR 型或 RS 型可能具有块 **100**、**101** 或 **110**；RR 型的块为 **11**（与 01 对齐的块**10** 对应于一次交换）。请注意，对于某些错误类型，我们还需要

区分与这些错误块相邻的不同字符所对应的子情形。我们在下面的命题中汇总了所有情形。由于篇幅所限，仅对情形 2 详细证明。对于其余情形，则通过展示一对可证明为一对波浪号见证的词来概述证明。

定理 24 设 $f \in \{0,1\}^n$。以下任一条件（至多相差取补和反转）足以使 f 波浪号非等距。

1. f 与误差位置不相邻的对象存在 2-波浪号误差重叠 *2.* f 与误差位置相邻的对象存在 *S* 型 *S* 型 2-波浪号误差重叠 *3.* f 与块**101** 存在 2-波浪号误差重叠（属于 *SR* 型或 *RS* 型） *4.* 在特定情形 $f = x\mathbf{100}1z = yx\mathbf{011}$ 下，对于某些 x、y、$z \in \{0,1\}^*$，f 与块 **100** 存在 2-波浪号误差重叠（属于 *SR* 型或 *RS* 型）*5.* 在特定情形 f 以**110**开头并以**100**结尾时，f 存在 2-波浪号误差重叠 *RR*。

证明： 对于列表中的每种情形，我们都给出 f 的一对波浪号见证。

情形 *1.* 若 2-波浪误差重叠不满足 $Condition^{\sim}$，根据定义 10，可以证明，公式 (1) 中的配对 $(\tilde{\alpha}_r, \beta_r)$ 是 f 的一对波浪号见证。否则，可以证明，$(\tilde{\eta}_r, \tilde{\gamma}_r)$，其中 $\tilde{\eta}_r = \mathrm{pre}_r(f)O_i(f)\mathrm{suf}^{r/2}(f)$ 和 $\tilde{\gamma}_r = \mathrm{pre}_r(f)O_j(O_t(f))\mathrm{suf}^{r/2}(f)$，是 f 的一对波浪号见证。

情形 *2.* 由引理 25 证明。

情形 *3.* 对于某些 $w_1, w_2, w_3, w_4 \in \{0,1\}^*$，有 $f = w_2\mathbf{101}w_3w_4 = w_1w_2\mathbf{010}w_3$。根据定义 10，配对 $(\tilde{\alpha}_r, \beta_r)$，其中包含 $\tilde{\alpha}_r = w_1w_2\mathbf{011}w_3w_4$ 和 $\beta_r = w_1w_2\mathbf{100}w_3w_4$，是一对波浪号见证。

情形 *4.* 对于某个 $w_1, w_2, w_3 \in \{0,1\}^*$，我们有 $f = w_2\mathbf{100}1w_3 = w_1w_2\mathbf{011}$。在这种情形下，我们需要采用不同的方法来构造这对波浪号见证 $(\tilde{\alpha}_r, \delta_r)$。我们令 $\tilde{\alpha}_r = w_1w_2\mathbf{0101}w_3$，并令 $\delta_r = w_1w_2\mathbf{1010}w_3$。这里我们证明 δ_r 是 f-自由的。事实上，假设 f 的一个副本出现在 δ_r 中，起始位置为 r_1。根据 δ_r 的定义和 f 的结构进行一些考察后可知，要么 $r_1 = 2$，要么 $r_1 = 3$，而且可以证明这会导致矛盾。

情形 *5.* 对于某个 $w_1, w_2 \in \{0,1\}^*$，我们有 $f = \mathbf{110}w_1 = w_2\mathbf{100}$。按照定义 10，可以证明，满足 $\tilde{\alpha}_r = w_2\mathbf{1010}w_1$ 且 $\delta_r = w_2\mathbf{0101}w_1$ 的有序对 $(\tilde{\alpha}_r, \delta_r)$ 是一对波浪号见证。注意，在这种情况下，方程1中的有序对 $(\tilde{\alpha}_r, \beta_r)$ 不是一对波浪号见证，因为 $\mathrm{dist}_{\sim}(\tilde{\alpha}_r, \beta_r) = 1$。

□

下面详细证明前一个定理的情形 2 成立。

引理 25 如果 f 具有一个 *SS* 型的2-波浪误差重叠，且错误彼此相邻，那么 f 是波浪号非等距的。

证明： 设 $f \in \{0,1\}^n$ 具有一个位移为 r 的 SS 型 2-波浪误差重叠，且错误彼此相邻。那么，除取补外，可能出现以下两种情形：

情形 *1*：$f = w_2\mathbf{1010}w_3w_4 = w_1w_2\mathbf{0101}w_3$，$|w_1| = r$

如果该 2-波浪误差重叠不满足条件~，那么β($\tilde{\alpha}_r$，$\tilde{\alpha}_r$ =)，其中$\tilde{w}_1w_2\mathbf{0110}w_3w_4$ 和 $\beta_r = w_1w_2\mathbf{1001}w_3w_4$是根据定义 10 得出的一对波浪号见证。事实上：

~1. 根据引理 21，$\tilde{\alpha}_r$ 不含 f；根据引理 23，β_r 不含 f。

~2. $\text{dist}_{\sim}(\tilde{\alpha}_r, \beta_r) = 2$，很直接。

~3. 从 $\tilde{\alpha}_r$ 到 β_r 的最小波浪号变换由两次交换 S_i 和 S_j组成，分别与 $i = |w_1 w_2| + 1$ 和 $j = i + 2$进行。如果将 S_i 应用于 $\tilde{\alpha}_r$ 作为第一次操作，那么 f 会作为后缀出现；而如果先将 S_j 应用于 $\tilde{\alpha}_r$，那么 f 会作为前缀出现。

如果 2-波浪误差重叠满足条件~，那么 $w_3 = \mathbf{10}w_3'$，并且根据定义 10，$(\tilde{\eta}_r, \tilde{\gamma}_r)$ 是一对波浪号见证，其中 $\tilde{\eta}_r = w_1 w_2 \mathbf{011010} w_3' w_4 w_5$ 和$\tilde{\gamma}_r = w_1 w_2 \mathbf{100101} w_3' w_4 w_5$，且 $w_5 = suf_{r/2}(f)$。

1. 可以证明，$\tilde{\eta}_r$ 和 $\tilde{\gamma}_r$ 是 f-自由的。

2. $\text{dist}_{\sim}(\tilde{\eta}, \tilde{\gamma}) = 3$。

3. 从 $\tilde{\eta}_r$ 到 $\tilde{\gamma}_r$ 的最小波浪号变换由三个交换操作S_i、S_j 和 S_t 组成，其中 $i = |w_1 w_2| + 1$、$j = i + 2$、$t = j + 2$。如果将 S_i 作为第一个操作应用于 $\tilde{\eta}_r$，那么 f 出现在位置 $|w_1| + 1$；如果先应用 S_j，那么 f 会作为前缀出现；而如果先应用 S_t，那么 f会作为后缀出现。

情形 *2*：$f = w_2\mathbf{1001}w_3 w_4 = w_1 w_2 \mathbf{0110} w_3$、$|w_1| = r$ 这一对 $(\tilde{\alpha}_r, \tilde{\beta}_r)$ 其中 $\tilde{\beta}_r = w_1 w_2 \mathbf{1010} w_3 w_4$ 是根据定义 10 得到的和一对~波浪号见证。在这种情况下，根据引理 23，β_r 是 f自由的。事实上，条件~ 从不成立，因为 $f[i]$ 与 $f[j]$不同。

□

下面的例子使用引理 25。

例 26 词 $f = \mathbf{1001}0110 = 1001\mathbf{0110}$ 在位置 $1, 3$ *25,* $\tilde{\alpha}_4$ β_4 $\tilde{\alpha}_4 =$上有一个移位为 $r = 4$、类型为 *SS* 的2波浪号错误重叠~。根据引理，(、) 与$1001\mathbf{0101}0110$ 以及 $\beta_4 = 1001\mathbf{1001}0110$ 构成一对波浪号见证。因此，f 是波浪号非等距的。注意，f 是汉明等距的。

定理 24 列出了词 f 为波浪号非等距的条件，而~证明给出了所有相应的见证对。迄今使用的 $(\tilde{\alpha}_r, \beta_r)$ 和 $(\tilde{\eta}_r, \tilde{\gamma}_r)$ 的构造，受到汉明距离下类似构造的启发（参见 [19]），并在此改编用于波浪号距离。相反，在情况 4 和 5 中，需要新的构造，因为通常的见证对不再满足定义 10。δ_r 的构造是波浪号距离特有的。它解决了注 17 中所述的情形：失配错误看似由替换造成，实际上却是由涉及相邻位置的隐藏交换造成。

总之，我们在本文中采用的交换与不匹配距离开辟了新的应用场景，并带来了值得进一步研究的有趣新情况。

参考文献

[1] Amihood Amir、Richard Cole、Ramesh Hariharan、Moshe Lewenstein 和 Ely Porat。重叠匹配。*Information and Computation*，181(1)：57–74，2003。

[2] Amihood Amir、Estrella Eisenberg 和 Ely Porat。交换与失配编辑距 离。Algorithmica，45(1)：109–120，2006。[3] Marcella Anselmo、Manuela Flores 和 Maria Madonia。四元 n-立方体与等距词。载于《组合学与词》，计算机科学讲义第 12842 卷，第 27–39 页，2021。[4] Marcella Anselmo、Manuela Flores 和 Maria Madonia。避免因子的趣味老虎机与词变换。载于第十一届算法趣味国际会议，LIPIcs 第 226 卷，第 4:1–4:15 页，2022。[5] Marcella Anselmo、Manuela Flores 和 Maria Madonia。关于 k 元 n 立方体和等距词。理论计算机科学，938：50–64，2022。[6] Marcella Anselmo、Dora Giammarresi、Maria Madonia 和 Carla Selmi。坏图：与重叠相关的一些结构性质。载于 DCFS 2020，计算机科学讲义第 12442 卷，第 13–25 页，2020。[7]Marie-Pierre B´eal 和 Maxime Crochemore。在线性时间内检验一个词是否为汉明等距。理论计算机科学，933：55–59，2022。[8] Marie-Pierre B´eal、Filippo Mignosi 和 Antonio Restivo。极小禁用词与符号动力学。载于 STACS 96，第 13 届计算机科学理论方面年度研讨会，计算机科学讲义第 1046 卷，第 555–566 页，1996。[9] Giuseppa Castiglione、Sabrina Mantaci 和 Antonio Restivo。基于缺失词的相似度度量的一些研究。Fundamenta Informaticae，171(1-4)：97–112，2020。[10] Panagiotis Charalampopoulos、Maxime Crochemore、Gabriele Fici、Robert Mer- cas 和 Solon P. Pissis。使用缺失词进行无比对序列比较。Information and Computation，262：57–68，2018。[11] Yair Dombb、Ohad Lipsky、Benny Porat、Ely Porat 和 Asaf Tsur。近似交换与失配编辑距离。理论计算机科学，411(43)：3814–3822，2010。[12] Simone Faro 和 Arianna Pavone。一种用于交换匹配的高效跳跃搜索方法。The Computer Journal，61(9)：1351–1360，2018。[13] Zvi Galil 和 Kunsoo Park。一种改进的近似字符串匹配算法。SIAM Journal on Computing，19：989–999，01 1990。[14] W.-J. Hsu。斐波那契立方体——一种新的互连拓扑。IEEE 并行与分布式系统汇刊，4(1)：3–12，1993。[15] Aleksandar Ili´c、Sandi Klavˇzar 和 Yoomi Rho。二进制词的指标。理论计算机科学，452：100–106，2012。[16] Sandi Klavˇzar 和 Sergey V. Shpectorov。等距广义斐波那契立方体的渐近数量。European Journal of Combinatorics，33(2)：220–226，2012。[17] V. I. Levenshtein。能够纠正删除、插入和反转的二进制码。控制论与控制理论，10：707–710，1966。

[18] Robert A. Wagner. 关于扩展字符串到字符串校正问题复杂性的研究。载于 William C. Rounds 等人编，《第七届 *ACM* 年度计算理论研讨会论文集，*1975*年》，第 218–223 页，1975年。[19] 魏建新。坏词的结构。*European Journal of Combinatorics*，59：204–214，2017年。[20] 魏建新、杨玉军和朱雪娜。非等距二进制词的刻画。*European Journal of Combinatorics*，78：121–133，2019年。[21] 魏建新、杨玉军和朱雪娜。非等距二进制词的刻画。*European Journal of Combinatorics*，78：121–133，2019年。[22] 魏建新和张和平。关于广义斐波那契立方体两个猜想的证明。*European Journal of Combinatorics*，51：419–432，2016年。