A Practical Semi-External Memory Method for Approximate Pattern Matching

Daniel Saad Nogueira Nunes[1,2]
Departamento de Ciência da Computação
Instituto de Ciências Exatas
Universidade de Brasília[1]
Instituto Federal de Educação, Ciência e Tecnologia de Goiás[2]
Email: daniel.saad.nunes@gmail.com

Mauricio Ayala-Rincón[1]
Departamentos de Ciência da Computação e Matemática
Instituto de Ciências Exatas
Universidade de Brasília[1]
Email: ayala@unb.br

***Abstract*—The approximate pattern matching problem ($\mathcal{APM}$) consists in locating all occurrences of a given pattern P in a text T allowing a specific amount of errors. Due to the character of real applications and the fact that solutions do not have error-free nature in Computer Science, solving $\mathcal{APM}$ is crucial for developing meaningful applications. Recently, Ilie, Navarro and Tinta presented a fast algorithm to solve $\mathcal{APM}$ based on the well-known Landau-Vishkin algorithm. However, the amount of available memory limits the usage of their algorithm, since it needs that all the answer array be in memory. In this article, a practical semi-external memory method to solve $\mathcal{APM}$ is presented. The method is based on the direct-comparison variation from Ilie, Navarro and Tinta algorithm. Performance tests with real data of length up to 1.2 GiB showed that the presented method is about 5 times more space-efficient than Ilie *et al.*'s algorithm and yet, has a competitive trade-off regarding time.**

I. Introduction

Exact pattern matching is a important problem which gives birth to a variety of applications in Computer Science, such as: Document Retrieval, Text Edition, Compiler tools, and many others.

When a fixed number of errors is allowed, interesting applications, with different complexities, arise (cf. [1], [2], [3]), from which the following can be mentioned:

- Fragment Mapping and assembling in Computational Biology: A high throughput sequencer gives as output billions of short fragments. In order to do accurate genomic/transcriptomic analysis, it is often necessary to map or to assemble all the fragments in a reference genome. Since errors are inherent within the sequencer technology, an exact approach is limited.
- Sequence Alignment with respect to a score: given a set of sequences, one must identify the similarities (or the lack of them) among them in order to achieve relevant results. This can be a crucial step towards the construction of phylogenetic trees, which estimate the evolutionary distance between organisms.
- Signal processing: noise in channels is inherent to the nature of signal processing. Hence, more robust methods which can deal with errors are indispensable.
- Document Retrieval: when allowing errors, one can execute more complex queries for the retrieval of documents, which is difficult when depending only on exact pattern matching.

Several algorithms have been proposed in the literature to solve $\mathcal{APM}$ (cf. [4], [5], [6], [7]). Among them, the Landau-Vishkin algorithm relies on the extension of the diagonals in a dynamic programming table to find approximate occurrences of a pattern in a text increasing the number of admissible errors in each extension [8]. However, this algorithm has a high consumption of memory, since it is based on complex data structures such as Suffix Trees or Suffix Arrays to execute the diagonal extension in constant time [9]. A variation of this algorithm proposed by Ilie *et. al* [10] has been reported to be faster and more space-economical than the classical version, since it does not rely on complex data structures, for it is based on a brute-force, but reliable way to extend the diagonals.

Despite of being efficient, the amount of available memory limits the usage of their algorithm, since it needs that all the answer array must be in memory. For instance, in a ordinary machine with 4 GiB of RAM, it could only manipulate texts of length up to 750 MiB. Therefore, a more space-efficient strategy is necessary.

This work presents a practical semi-external memory method to solve $\mathcal{APM}$ by using the direct-comparison variation from [10]. Performance tests with real data of length up to 1.2 GiB. It was shown that this approach is $\approx 5\times$ more space-efficient than the pure direct-comparison variation and yet, has a competitive trade-off regarding time. Hence, by using both memory and disk in a clever way, it is feasible to manipulate larger files which were impossible to be treated before.

II. Background

Let Σ^* denotes the set of all strings over the finite alphabet $\Sigma = \{a_0, a_1, \ldots, a_{\sigma-1}\}$, such that $|\Sigma| = \sigma$. Specially, the empty string is denoted by ϵ, which has length $|\epsilon| = 0$.

The i^{th} symbol of a given string $X \in \Sigma^*$ is denoted by $X[i]$. Substrings of X are denoted by $X[i, j] = X[i]X[i + 1] \ldots X[j]$, $0 \leq i \leq j < |X| - 1$, otherwise $X[i, j] = \epsilon$. Suffixes $X[i, |X| - 1]$ are denoted by X_i.

Two particular strings, called the text and the pattern, are denoted respectively by T and P, with $|T| = n$ and $|P| = m$.

Definition 1 (Edit Distance):

The edit distance $\delta(X, Y)$ between any two strings X and Y, is the number of required operations which turns X onto Y.

A common distance function used is the Levenhestein distance, whose operations are based on insertion, deletion and substitution of symbols, each with cost 1. From now on, this will be the standard distance.

Definition 2 ($\mathcal{APM}$):

The Approximate Pattern Matching Problem can be formulated as, given T, P and a number of errors k, return all positions:

$$Occ = \{j | 0 \leq i \leq j < n \wedge \delta(P, T[i, j]) \leq k\} \quad (1)$$

Occ stands for the positions where P ends in T with at most k errors.

Figure 1 illustrates the occurrences of $P = ACA$ in $T = T = ACTAGACATAGCAA$ allowing at most one error (insertion, deletion or mismatch).

Definition 3 (LCE): The longest common string (LCE) of any given two strings X and Y corresponds to the length of the maximal prefix shared by X and Y:

$$LCE(X, Y) = \max\{k | X[0, k-1] = Y[0, k-1]\} \quad (2)$$

The Suffix Tree [11] is a fundamental text indexing data structure, as shown by [3]. However, its space consumption turns their usage unfeasible for large strings. The suffix array [12] is one of the most important space-efficient alternative which has been used to solve efficiently (in time and space) many string processing problems (c.f [13]).

Definition 4 (SA): A Suffix Array, SA for short, of a text T is a array A of integers containing the position of the Suffixes in lexicographical order induced by the alphabet symbols. Hence:

$$T_{A[0]} < T_{A[1]} < \ldots < T_{A[n-1]}$$

Definition 5 (ISA): The inverse Suffix Array of a text T is a array A^{-1} of integers containing in the i^{th} entry, the lexicographical position of the i^{th} suffix among the others.

While in the Suffix Array $A[i]$ is concerned about the i^{th} suffix in lexicographical order, $A^{-1}[i]$ corresponds to the lexicographical position of the suffix T_i among the other suffixes of T. Therefore $A[A^{-1}[i]] = i$.

Definition 6 (LCP): The Longest-Common-Prefix is defined as the length of the maximal prefix shared by two consecutive entries in the SA A. Formally:

$$LCP(i) = \begin{cases} 0, & i = 0 \\ LCE(T_{A[i-1]}, T_{A[i]}), & i > 0 \end{cases} \quad (3)$$

Table I shows a SA A augmented with A^{-1} and the LCP for the text $T' = ACTAGACATAGCAA\#ACA\$$.

Definition 7 (RMQ): Let the Range-Minimum-Queries RMQ_V over an array V be defined as:

$$RMQ_V(i, j) = \min\{\arg\min\{V[k] \mid i \leq j \leq j\}\} \quad (4)$$

Thus, $RMQ_V(i, j)$ holds the leftmost position in which occurs the minimum value on $V[i, j]$.

Suffix Arrays and their inverses can be build in $\Theta(n)$ time, as shown by Kärkkäinen and Sanders [14]. The LCP information also can be computed in $\Theta(n)$ time [15]. The RMQ_{LCP} support data structure can be computed in $\Theta(n)$ as well while allowing $\Theta(1)$ queries (c.f [16], [17]).

III. The Landau-Vishkin Algorithm

The Landau-Vishkin Algorithm, proposed originally in [8], solves the $\mathcal{APM}$. This algorithm is based on a dynamic programming technique which, at the k^{th} iteration, obtains the maximal extension of diagonals of the dynamic programming table allowing at most k errors. The diagonals refers to the Table of the classical Dynamic Programming technique which computes the minimal edit distance [18].

Considering k as the number of errors, and i as the i^{th} diagonal, the dynamic programming technique is based on the recurrence relation $L(i, k)$:

$$L(i,k) = \begin{cases} \text{For } (k = 0) \wedge (0 \leq i \leq n), \\ \quad L(i, 0) := LCE(P_0, T_i) \\ \\ \text{For } (-(m-1) \leq i < 0) \wedge (k = -i), \\ \quad L(i,k) := \text{let } j = (L(i+1, k-1) + 1) \text{ in} \\ \quad\quad \text{if } j \geq m + i \text{ then } m + i \\ \quad\quad \text{else } max(LCE(P_{-i}, T_0), j + LCE(P_{-i+j+1}, T_{j+1})) \\ \\ \text{For } (-(m-1) \leq i \leq 0) \wedge (k = -i + 1) \\ \quad L(i,k) := \text{let } j = max(L(i, k-1) + 1, L(i+1, k-1) + 1) \text{ in} \\ \quad\quad \text{if } j \geq m + i \text{ then } m + i \\ \quad\quad \text{else } j + LCE(P_{-i+j+1}, T_{j+1}) \\ \\ \text{For } (0 < k \leq m) \wedge (-(m-1) \leq i < n), \text{where} \\ \quad \text{for } (i \geq 0), (k + i \leq n) \wedge \text{ for } (i < 0), (k > -i + 1) \\ \quad L(i,k) := \text{let } j = \max(L(i-1, k-1), L(i, k-1) + 1, \\ \quad\quad\quad\quad\quad\quad L(i+1, k-1) + 1) \text{ in} \\ \quad\quad \text{if } j \geq m \vee i + j \geq n \text{ then } \min(m, n - i) \\ \quad\quad \text{else } j + LCE(P_{j+1}, T_{i+j+1}) \end{cases} \quad (5)$$

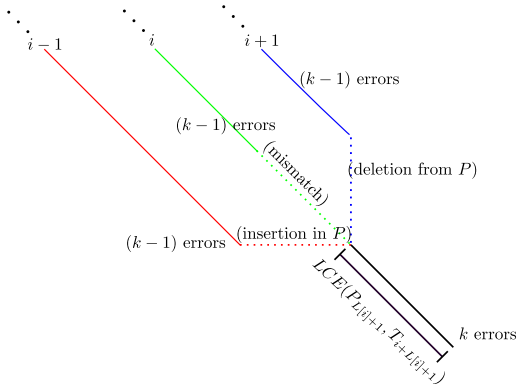


Figure 2: i^{th} diagonal extension in the Landau-Vishkin Algorithm via diagonal $i - 1$.

Thus, by this relation, $L(i, k)$ indicates that $P[0, L(i, k)-1]$ occurs at the position $i+L(i, k)-1$ of T with at most k errors.

This recurrence states that, for the diagonal i, allowing k errors, one has to look for the maximal extensions in diagonals $i-1$, i and $i+1$ allowing $k-1$ errors, introduce an error, and

$AC{-}TAGACATAGCAA$
ACA

$ACTAGACATAGCAA$
ACA

$ACTAGACATAGCAA$
$AC{-}A$

$ACTAGACATAGCAA$
$\quad ACA$

$ACTAGAC{-}ATAGCAA$
$\quad ACA$

$ACTAGACATAGCAA$
$\quad ACA$

$ACTAGACATAGCAA$
$\quad AC{-}A$

$ACTAGACATAGCAA$
$\quad ACA$

$ACTAGACATAGCAA$
$\quad A{-}CA$

$ACTAGACATAGCA{-}A$
$\quad ACA$

Figure 1: Ocurrences of $P = ACA$ in $T = ACTAGACATAGCAA$ allowing at most one error.

Table I: Suffix array for $T' = T\#P\$ = ACTAGACATAGCAA\#ACA\$$.

i	$T'_{A[i]}$	$A[i]$	$A^{-1}[i]$	$LCP[i]$
0	$\#ACA\$$	14	7	0
1	$\$$	18	14	0
2	$A\#ACA\$$	13	17	0
3	$A\$$	17	8	1
4	$AA\#ACA\$$	12	15	1
5	$ACA\$$	15	6	1
6	$ACATAGCAA\#ACA\$$	5	13	3
7	$ACTAGACATAGCAA\#ACA\$$	0	10	2
8	$AGACATAGCAA\#ACA\$$	3	18	1
9	$AGCAA\#ACA\$$	9	9	2
10	$ATAGCAA\#ACA\$$	7	16	1
11	$CA\$$	16	12	0
12	$CAA\#ACA\$$	11	4	2
13	$CATAGCAA\#ACA$	6	2	2
14	$CTAGACATAGCAA\#ACA\$$	1	0	1
15	$GACATAGCAA\#ACA\$$	4	5	0
16	$GCAA\#ACA\$$	10	11	1
17	$TAGACATAGCAA\#ACA\$$	2	3	0
18	$TAGCAA\#ACA\$$	8	1	3

then extend the diagonal maximally again. Graphically this is represented by Figure 2.

The implementation details of the dynamic programming technique are showed by Algorithm 1. One does not need to represent the entire Dynamic Programming Table, only the array L is necessary. Table II shows a snapshot of table L for $T = ACTAGACATAGCAA$ and $P = ACA$ allowing at most $k = 1$ errors.

The Landau-Vishkin Algorithm in the worst-case takes $\Theta(nk \cdot t_{LCE})$, where t_{LCE} stands for the time needed to compute the $LCE(P_{L[i]+1}, T_{i+L[i]+1})$ [8].

A. Classical Approach

The classical approach computes $LCE(P_{L[i]+1}, T_{i+L[i]+1})$ with the aid of Suffix Trees [8]. First, a generalized Suffix Tree is build for the text $T' = T\#P\$$ and then, Lowest Common Ancestor (LCA) queries are done between leaves in order to obtain the LCE value. Since LCA queries can be done in $\Theta(1)$ time with a $\Theta(n)$ time preprocessing [19], it is possible to extend a diagonal maximally in $\Theta(1)$ time.

However, Suffix trees have a high consumption of memory in practice. In fact, considering the worst case behavior, Suffix Trees have a consumption of $10\times$ to $15\times$ the text input size [20]. This huge factor turns the manipulation of larger texts unfeasible due the amount of memory available in ordinary machines.

B. Using Suffix Arrays

A variation proposed by Miranda and Ayala-Rincón [9] uses a more space-efficient data structure called Suffix Array (SA) [12]. This data structure is augmented with additional information: the inverse Suffix Array (ISA), the Longest-Common-Prefix (LCP) information and a support data structure for Range-Minimum-Queries (RMQ).

Table II: Snapshot of $L[i]$ for $T = ACTAGACATAGCAA$ and $P = ACA$

	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$L[0]$		1	-1	-1	0	-1	2	-1	0	-1	0	-1	-1	0	0
$L[1]$	2	2	1	2	2	2	2	2	2	1	1	2	2	1	1

Algorithm 1: Generic Landau-Vishkin Algorithm

Input: P, T, k
Output: $\{j | P$ ends in $T[j]$ with at most k errors$\}$

$L[i] = -2, -k \leq i \leq n$;
for $e \leftarrow 0; e \leq k; k++$ **do**
  $prev = -2$;
  $cur = -2 + e$;
  $next = L[-e + 1]$;
  for $i = -e; i < n; i++$ **do**
    $L[i] \leftarrow \max(prev, cur + 1, next + 1)$;
    $L[i] \leftarrow \max(L[i], m - 1)$;
    if $(i + L[i] + 1 < n)$ **then**
      $L[i] \leftarrow L[i] + \text{LCE}(P_{L[i]+1}, T_{i+L[i]+1})$;
    $prev = cur$;
    $cur = next$;
    $next = L[i + 2]$;
for $i = -k; i < n; i++$ **do**
  if $L[i] \geq |P| - 1$ **then**
    REPORTMATCH$(T_{i+|P|-1})$

In order to compute the value $LCE(P_{L[i]+1}, T_{i+L[i]+1})$ in $\Theta(1)$ time, one needs to build a complex data structure in the preprocessing phase. Building this data structure consist of: creating the Suffix Array A for $T' = T\#P\$$, computing its inverse A^{-1}, calculating its LCP information and building a support data structure RMQ_{LCP}.

The answer of a given $RMQ_{LCP}(k, l)$ can be obtained in constant time after the data structure is built. Thus, $LCE(P_{L[i]+1}, T_{i+L[i]+1}) = LCP[RMQ_{LCP}(k+1, l)]$, where $k = \min\{A^{-1}[L[i] + n + 2], A^{-1}[L[i] + i + 1]\}$ and $l = \max\{A^{-1}[L[i] + n + 2], A^{-1}[L[i] + i + 1]\}$. Hence, the LCE query is supported in constant time.

For example, in the Table I, $LCE(AGCAA\#ACA\$, ACA\$)$ corresponds to $LCP[RMQ_{LCP}(6, 9)] = LCP[8] = 1$. The resultant procedure is given by Algorithm 2.

Algorithm 2: Extension of diagonals by using LCP and RMQ queries

Input: $P, T, i, L[i], A^{-1}, LCP, RMQ_{LCP}$
Output: $LCE(P_{L[i]+1}, T_{i+L[i]+1})$

$i_1 = A^{-1}[|T| + L[i] + 2]$;
$i_2 = A^{-1}[i + L[i] + 1]$;
if $i_1 > i_2$ **then** SWAP(i_1, i_2);
return $LCP[RMQ_{LCP}(i_1 + 1, i_2)]$;

Since RMQ_{LCP} queries take $\Theta(1)$ time, $t_{LCE} \in \Theta(1)$. Therefore, the Landau-Vishkin algorithm takes $\Theta(kn)$ in the worst case when using this collection of data structures.

C. The Direct Comparison Variation

Despite of assuring a $\Theta(kn)$ worst case time, the classical solution is often considered unpractical, due to the large factors involved in the construction of Suffix Arrays, LCP information and the RMQ support data structure.

Moreover, according to Ilie *et. al*, LCE values tend to be quite small on average [10]. In fact, there is a proof that, the average $LCE(i, j)$ considering all strings of length n, is $\leq \frac{1}{\sigma - 1}$.

Due to this behavior, RMQ queries do not go well, since despite taking time in $\Theta(1)$, a meaningful overhead is present. A brute-force approach to compute LCE values is faster in practice, since they are usually small.

Ilie *et. al* showed that the Landau-Vishkin algorithm turns from a unpractical one to a practical one when direct-comparisons are made in order to compute $LCE(P_{L[i]+1}, T_{i+L[i]+1})$. The Algorithm 3 shows the direct-comparisons version of the diagonal extension.

Algorithm 3: Extension of diagonals by using direct-comparisons

Input: P, T
Output: $LCE(P_{L[i]+1}, T_{i+L[i]+1})$

$c = 0$;
while $P[L[i] + 1 + c] == T[i + L[i] + 1 + c]$ **do**
  $c++$;
return c;

The extension by direct-comparisons showed to be very effective [10]. Besides, it utilizes the computer resources in a more appropriate way, especially with respect to cache memory due to the great locality of reference, which was poor in the RMQ based queries of the classical variation.

Asymptotically, when using the direct-comparisons, $t_{LCE} = O(m)$ in the worst case. However, on average, $t_{LCE} = \frac{1}{\sigma - 1} \in \Theta(1)$. Therefore, a Landau-Vishkin approach, in the average case using direct-comparisons, would take $\Theta(kn)$ time.

IV. Proposed Semi-External Memory Method

A shared problem between the classical and the direct-comparison versions of the Landau-Vishkin algorithm is the amount of memory used. The straightforward classical variation needs $4n$ bytes for A, $4n$ bytes for A^{-1}, $4n$ bytes for

the LCP information, $4n$ bytes for the L array, n bytes for the text, m bytes for the patten and $\approx \frac{7}{8}n$ bytes for the RMQ_{LCP} support data structure based on [17]. The faster and more space-economical direct-comparison variation does not need any preprocessed data structure, only n bytes for the text, m bytes for pattern and $4n$ bytes for the L array.

Thus, an ordinary machine with 4 GiB of RAM could only manipulate texts of ≈ 200 MiB in the classical variation and ≈ 750 MiB in the pure direct-comparison variation. Therefore, larger texts cannot be treated by ordinary approaches. A more space-efficient strategy is necessary.

Semi-external algorithms are in between internal memory algorithms and external memory algorithms. An algorithm is named semi-external if the input is dependent to the amount of available memory, but it also uses external memory. In practice, a semi-external algorithm access the data os disk sequentially and maintain in memory the data which needs to be accessed randomly.

This work proposes a practical semi-external memory method based on the direct-comparisons variation from [10] to solve $\mathcal{APM}$. This method explores the fact that the array L is not entirely needed in memory. Thus, one can keep a block $B = L[j, j+|B|-1]$ in main memory to process the diagonals $j+1 \ldots j+|B|-2$. When the next block of diagonals need to be computed, one simply store B into disk and load the new chunk $B' = L[j+|B|-1, j+2|B|-2]]$ from the same disk into memory. Since these values would occupy a contiguous space in the disk, seeks are minimized, for the data tend to be in the same track, and hence, the technique does not struggle from the disk access bottleneck. Choosing $B \in \Theta(1)$, the overall space required is $n+m+\Theta(1)$, about $5\times$ less than the original direct-comparisons variation ($5n + m$ bytes).

The Table III summarizes the discussed until now.

V. Experimental Results

In order to evaluate the proposed semi-external variations, experiments were performed considering collections of large texts from Pizza-Chili[1] and Manzini[2] *corpora*. Table IV shows a description of the used texts. Each text has its own particularities and specific alphabets.

Comparisons were done with the respective Landau-Vishkin algorithm variations:

1) LV_RMQ: classical Variation with Suffix Arrays. Uses A, A^{-1} , LCP and RMQ_{LCP}. To build A, `libdivsufsort` was used [21]. The LCP was constructed by the authors using Kasai *et. al* method [15]. For the support RMQ_{LCP} data structure, code from [17] was used.
2) LV_DMIN: direct-comparison variation aided by LCP information. It uses direct-comparisons only if the distance in the Suffix Array from the suffixes $P_{L[i]+1}$ and $T_{i+L[i]+1}$ is large enough (suffixes which are close in the Suffix Array tend to share more symbols). Otherwise, a linear scan picking the minimum LCP value between these two suffixes is used. In this variation RMQ_{LCP} is not used.
3) LV_DC: our direct-comparison variation. Only needs P, T and L.
4) LV_DC_NAC: Navarro's direct-comparison variation from [10].
5) LV_DC_SE: our semi-external variation using a SATA 500 GB, 7200 RPM hard disk drive.
6) LV_DC_SE_SSD: our semi-external variation using a 64 GB, SATA SSD disk.

The code is based on C++11 standard and all tests ran in an `Ubuntu 12.04 64-bit`, `core i5-750`, $4GB$ 1333 Mhz RAM memory machine. All codes were compiled using `gcc 4.8` with the `-O2` flag for optimization purposes. The code is available on [22].

Each experiment was executed 3 times and the lowest wall clock time was chosen. At every experiment, a pattern with length 50 was chosen from the text randomly, as was done in [10], since the overall time showed to be largely independent of P and m. Errors from the set $\{0, 1, 2, 3, 6, 10, 20\}$ were considered in the experiments. Prefixes files from the *corpora* were also considered.

Figure 3 shows the experiments considering Manzini *corpus*. On every experiment, one can see that both the direct-comparison variations (LV_DC and LV_DC_NAV) are faster than any other variation. The slower one, is the classical approach using RMQ queries, which behave poorly for every file, confirming the high constants involved in the algorithm. The semi-external variations (LV_DC_SE and LV_DC_SE_SSD) have a good trade-off between space and time, for they are a order of magnitude slower if compared to the direct-comparison variations, but they require $\approx 5\times$ less memory. Finally, the hybrid LV_DMIN variation shows to be as competitive as the semi-external variations, since the data structures required take a considerable amount of time to be built, however, its space requirement is much higher than the semi-external variations.

We shall consider now experiments under Pizza-Chili *Corpus*.

For DNA files, Figure 4 shows the experiments performed. Considering the 200MiB prefix file from Figure 4a, it is clear that the classical variation is unpractical, taking two orders of magnitude more than the direct-comparison variations. Once again, the semi-external variations show to have a good compromise between time and space, being only one order of magnitude slower than the direct-comparison variation. Again, LV_DMIN shows a similar behavior with respect to the semi-external variations, but using much more memory. Figure 4b shows a situation where the classical and the LV_DMIN variations do not fit in memory. Once again, the semi-externals variations showed to be very competitive with respect to time/space trade-off. Also, it can be noticed that the LV_DC_SE_SSD variation has a significant difference in relation as the LV_DC_SE variation, which shows the potential of the SSD technology for larger files.

[1] http://pizzachili.dcc.uchile.cl/

[2] http://people.unipmn.it/~manzini/lightweight/corpus/

Table III: Memory, data structures and time required for Landau-Vishkin variations

	Variations		
	Classical	Direct-Comparisons	Semi-External
Data Structures	$P, T, A, A^{-1}, LCP, RMQ$ and L	P, T and L	P and T
Memory (bytes)	$18.875n + m$	$5n + m$	$n + m + \Theta(1)$
Worst-Case	$\Theta(kn)$	$\Theta(knm)$	$\Theta(knm)$
Average-Case	$\Theta(kn)$	$\Theta(kn)$	$\Theta(kn)$

Table IV: Texts used in the experiments.

Type	Size	Description
Manzini *Corpus*		
JDK	67MiB	HTML file documenting Java 2
HOWTO	38MiB	Describes Dfx graphics accelerator chip support for Linux
RCTAIL	110MiB	XML file containing news from Reuters
RFC	112MiB	A request for comments file
Pizza Chili *Corpus*		
DNA	386MiB	DNA texts collected from Gutenberg project
DBLP	283MiB	XML files containing bibliographical information from DBLP[3]
English	1024MiB	Natural Language texts collected from Gutenberg Project
Proteins	1.2GiB	Sequence of proteins obtained from Swissprot [4]
Sources	202MiB	Concatenated source code from gcc-4.0.0 and linux-2.6.11.6

Considering now the XML alphabet, Figures 5a and 5b show the experiments performed for a 200MiB and a 283MiB XML-DBLP files. As discussed before, the semi-external variations have an acceptable trade-off between time and space. Besides, Figure 5b shows that both the classical and the LV_DMIN variations are unfeasible for larger files, since they do not fit in main memory.

Figures 6, 7 and 8 show the results with respect to the English, Proteins and Sources texts, respectively. The same behavior from the previous experiments is observed. The interesting cases are shown in Figures 6b and 7b. In these two experiments, even the direct-comparison variations do not fit in memory, thus the semi-external variations show their truly utility. Once again, the LV_DC_SE_SSD shows to be more efficient than the LV_DC_SE variation for larger files.

VI. Conclusion

$\mathcal{APM}$ is a very important and recurrent problem in Computer Science with applications in many areas. The Landau-Vishkin method solves this problem by using the classical dynamic programming approach, but by extending maximally each diagonal of the table while the number of errors is increased by one. It has been shown that the classical variation performs poorly in practice. Due to an observation from [10], it was shown that the expected LCE between any two strings with length n is $1/(\sigma - 1)$. Hence, a brute-force approach to compute LCE values results faster than elaborated and complex solutions, such as the ones given in [8], [9].

This work proposed a semi-external variation of the Landau-Vishkin algorithm based on direct-comparisons. Despite of being only an order of magnitude slower in the majority of the experiments than the pure direct-comparison variations, the semi-external variation is a factor of $\approx 5\times$ more space-efficient, as stated by Table III. Besides, the semi-external variation is able to manipulate larger texts which cannot be possibly treated by the pure direct-comparison variation, since it does not fit in main memory. Consequently, the proposed semi-external memory approach is of practical value.

[3]http://dblp.uni-trier.de/db/

[4]ftp://ftp.ebi.ac.uk/pub/databases/swissprot/release_compressed/

VII. Future Work

Each symbol of T or P can be coded in $\log_2 \sigma$ bits, instead of a 8-bit ASCII representation. For example, analyzing the DNA alphabet $\Sigma = \{A, C, G, T\}$, one would need only 2 bits to represent each symbol. If the machine word length is 64-bits, one can pack 32 symbols in a integer and then, do a single comparison between integers to compare each pair of 32 symbols. Thus, the LCE extension would be a lot faster provided this preprocessing in T and P. The space requirements of T and P would decrease as well for small alphabets, since we would need only $n \log_2 \sigma$ bits for representing T and $m \log \sigma$ for P (25% of the original byte representation for the DNA alphabet).

Hence, a modification of the presented algorithm would benefit from texts and patterns with a small underlying alphabet with minimal preprocessing cost.

Acknowledgements

We would like to thank Gonzalo Navarro for making his direct-comparisons Landau-Vishkin code from [10] available for the experiments and Felipe Louza for meaningful suggestions.

REFERENCES

[1] G. Navarro, "A guided tour to approximate string matching," *ACM Computing Surveys*, vol. 33, no. 1, pp. 31–88, 2001.

[2] J. C. Setubal and J. Meidanis, *Introduction to Computational Molecular Biology*. PWS Publishing Company, 1997.

[3] D. Gusfield, *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*. Cambridge University Press, 1997.

[4] E. Ukkonen, "Algorithms for Approximate String Matching," *Information and Control*, vol. 64, no. 1-3, pp. 100–118, 1985.

[5] P. H. Sellers, "The Theory and Computation of Evolutionary Distances: Pattern Recognition," *J. Algorithms*, vol. 1, no. 4, pp. 359–373, 1980.

[6] W. J. Masek and M. Paterson, "A Faster Algorithm Computing String Edit Distances," *Journal of Computer and System Sciences*, vol. 20, no. 1, pp. 18–31, 1980.

[7] R. Cole and R. Hariharan, "Approximate String Matching: A Simpler Faster Algorithm," *SIAM Journal on Computing*, vol. 31, no. 6, pp. 1761–1782, 2002.

[8] G. M. Landau and U. Vishkin, "Fast Parallel and Serial Approximate String Matching," *Journal of Algorithms*, vol. 10, no. 2, pp. 157–169, 1989.

[9] R. de Castro Miranda and M. Ayala-Rincón, "A Modification of the Landau-Vishkin Algorithm Computing Longest Common Extensions via Suffix Arrays," in *Brazilian Symposium on Bioinformatics*, 2005, pp. 210–213.

[10] L. Ilie, G. Navarro, and L. Tinta, "The Longest Common Extension Problem Revisited and Applications to Approximate String Searching," *Journal of Discrete Algorithms*, vol. 8, no. 4, pp. 418–428, 2010.

[11] P. Weiner, "Linear pattern matching algorithms," in *14th Annual Symposium on Switching and Automata Theory, Iowa City, Iowa, USA, October 15-17, 1973*. IEEE Computer Society, 1973, pp. 1–11. [Online]. Available: http://dx.doi.org/10.1109/SWAT.1973.13

[12] U. Manber and E. W. Myers, "Suffix arrays: A new method for on-line string searches," *SIAM Journal on Computing*, vol. 22, no. 5, pp. 935–948, 1993.

[13] E. Ohlebusch, *Bioinformatics Algorithms: Sequence Analysis, Genome Rearrangements, and Phylogenetic Reconstruction*. Oldenbusch Verlag, 2013. [Online]. Available: http://www.oldenbusch-verlag.de/

[14] J. Kärkkäinen and P. Sanders, "Simple Linear Work Suffix Array Construction," in *International Colloquium on Automata, Languages, and Programming*, ser. Lecture Notes in Computer Science, J. C. M. Baeten, J. K. Lenstra, J. Parrow, and G. J. Woeginger, Eds., vol. 2719. Springer, 2003, pp. 943–955.

[15] T. Kasai, G. Lee, H. Arimura, S. Arikawa, and K. Park, "Linear-time longest-common-prefix computation in suffix arrays and its applications," in *Combinatorial Pattern Matching, 12th Annual Symposium, CPM 2001 Jerusalem, Israel, July 1-4, 2001 Proceedings*, ser. Lecture Notes in Computer Science, A. Amir and G. M. Landau, Eds., vol. 2089. Springer, 2001, pp. 181–192. [Online]. Available: http://link.springer.de/link/service/series/0558/bibs/2089/20890181.htm

[16] Johannes Fischer and Volker Heun, "Theoretical and Practical Improvements on the RMQ-Problem, with Applications to LCA and LCE," in *Combinatorial Pattern Matching*, ser. Lecture Notes in Computer Science, M. Lewenstein and G. Valiente, Eds., vol. 4009. Springer, 2006, pp. 36–48.

[17] J. Fischer and V. Heun, "A New Succinct Representation of RMQ-Information and Improvements in the Enhanced Suffix Array," in *Combinatorics, Algorithms, Probabilistic and Experimental Methodologies, First International Symposium, ESCAPE 2007, Hangzhou, China, April 7-9, 2007, Revised Selected Papers*, ser. Lecture Notes in Computer Science, B. Chen, M. Paterson, and G. Zhang, Eds., vol. 4614. Springer, 2007, pp. 459–470. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-74450-4_41

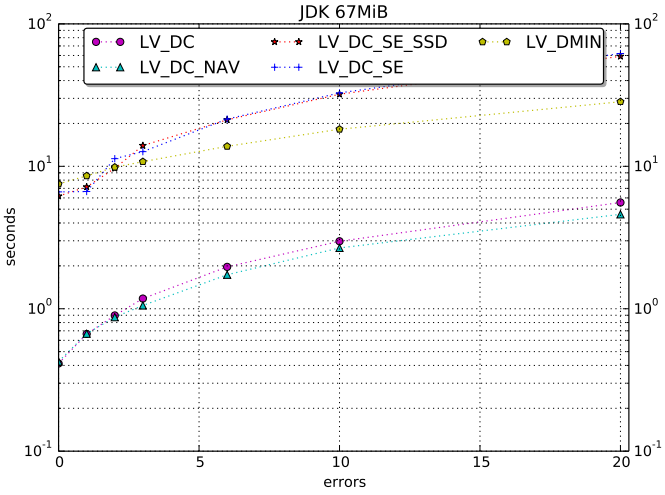
[18] D. S. Hirschberg, "A linear space algorithm for computing maximal common subsequences," *Commun. ACM*, vol. 18, no. 6, pp. 341–343, 1975. [Online]. Available: http://doi.acm.org/10.1145/360825.360861

[19] M. A. Bender and M. Farach-Colton, "The LCA Problem Revisited," in *LATIN*, ser. Lecture Notes in Computer Science, G. H. Gonnet, D. Panario, and A. Viola, Eds., vol. 1776. Springer, 2000, pp. 88–94.

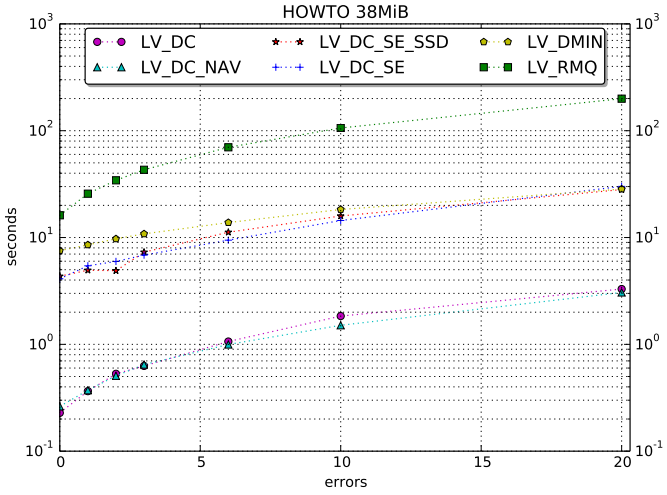
[20] S. Kurtz, "Reducing the Space Requirement of Suffix Trees," *Software: Practice and Experience*, vol. 29, no. 13, pp. 1149–1171, 1999.

[21] Y. Mori, "libdivsufsort - a lightweight suffix sorting library," 2007. [Online]. Available: https://code.google.com/p/libdivsufsort/

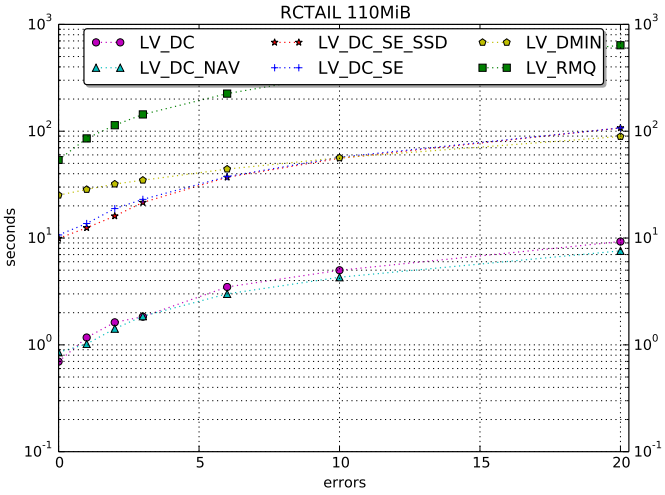
[22] D. S. N. Nunes, "A semi-external implementation of the Landau-Vishkin algorithm," 2015. [Online]. Available: https://github.com/danielsaad/Semi-External-Landau-Vishkin



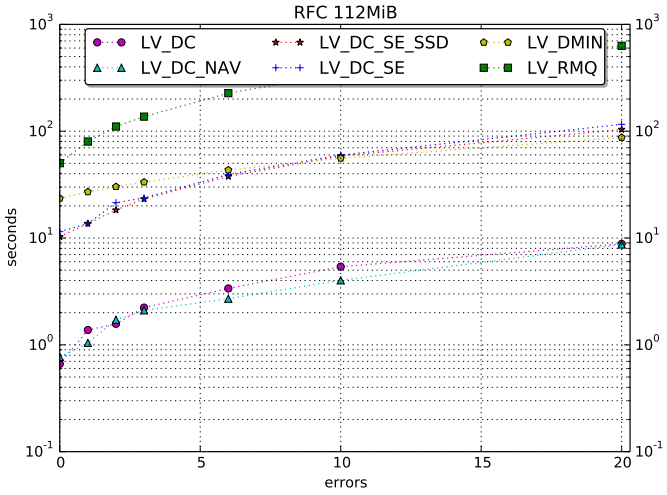
(a) 67MiB JDK File



(b) 38MiB HOWTO file.

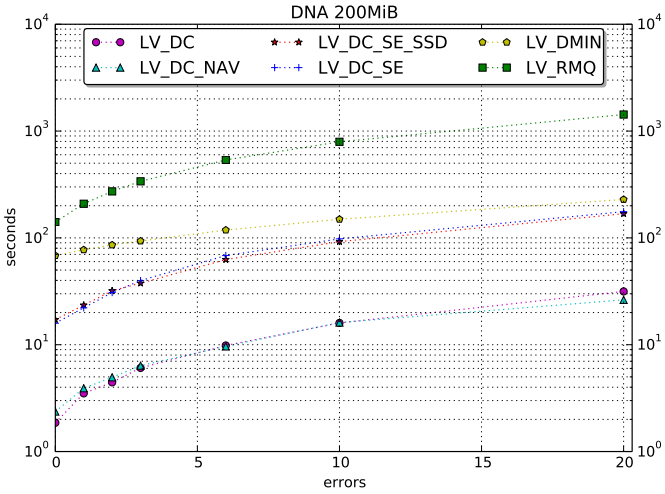


(c) 110MiB RCTAIL file.

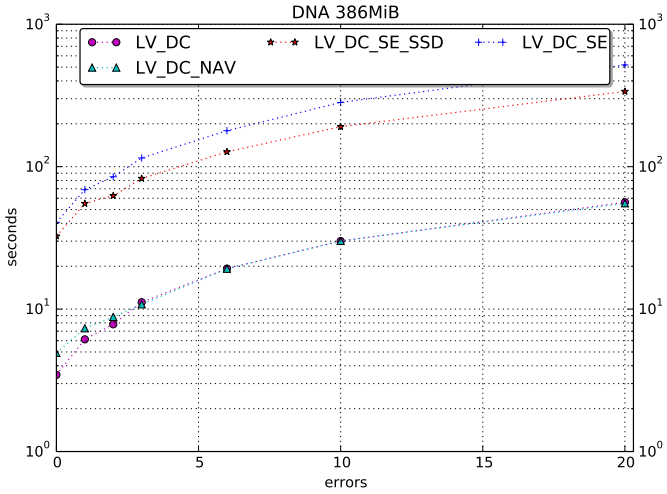


(d) 112MiB RFC file.

Figure 3: Manzini *corpus* files

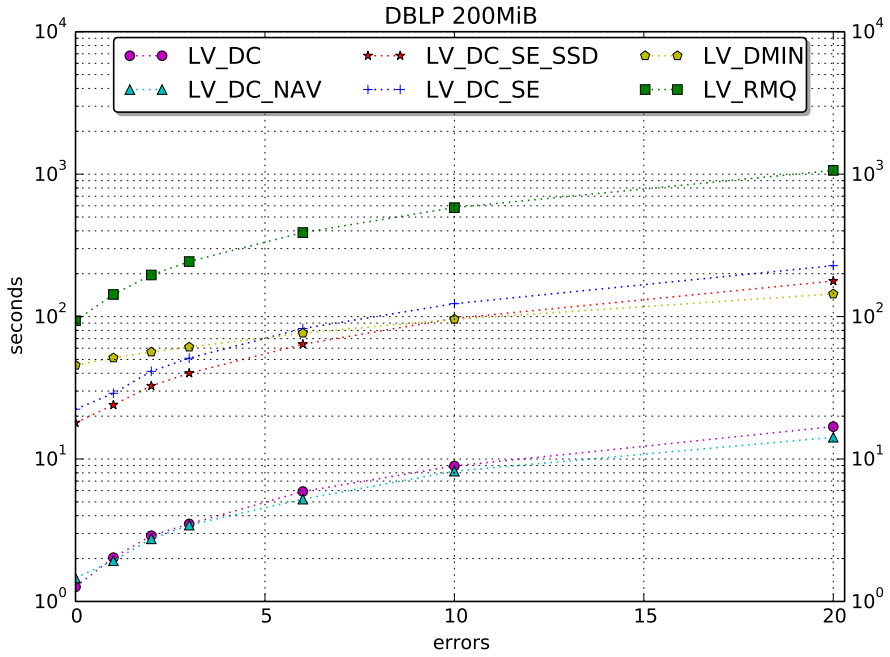


(a) 200MiB prefix DNA file.

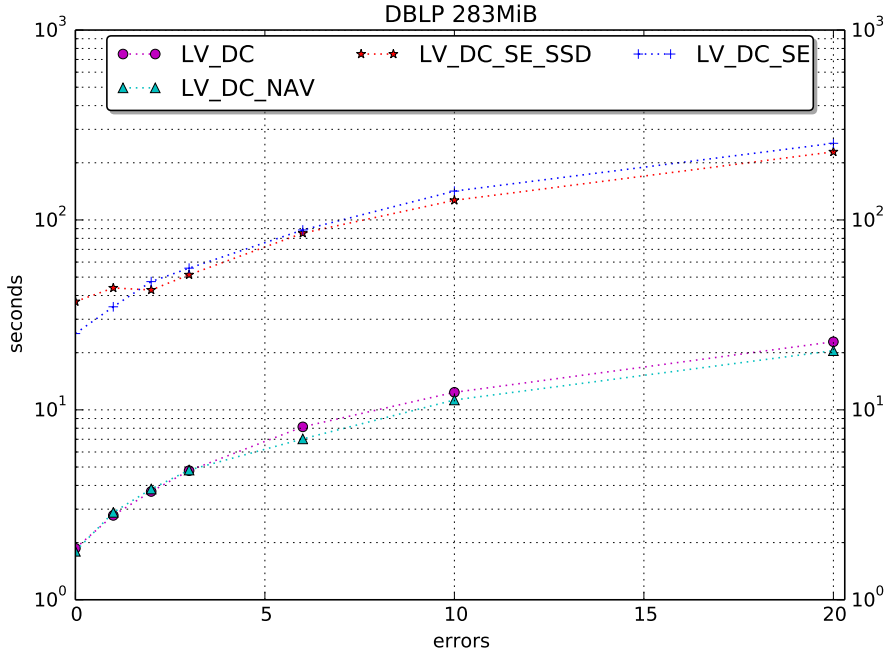


(b) 386MiB DNA file.

Figure 4: DNA files

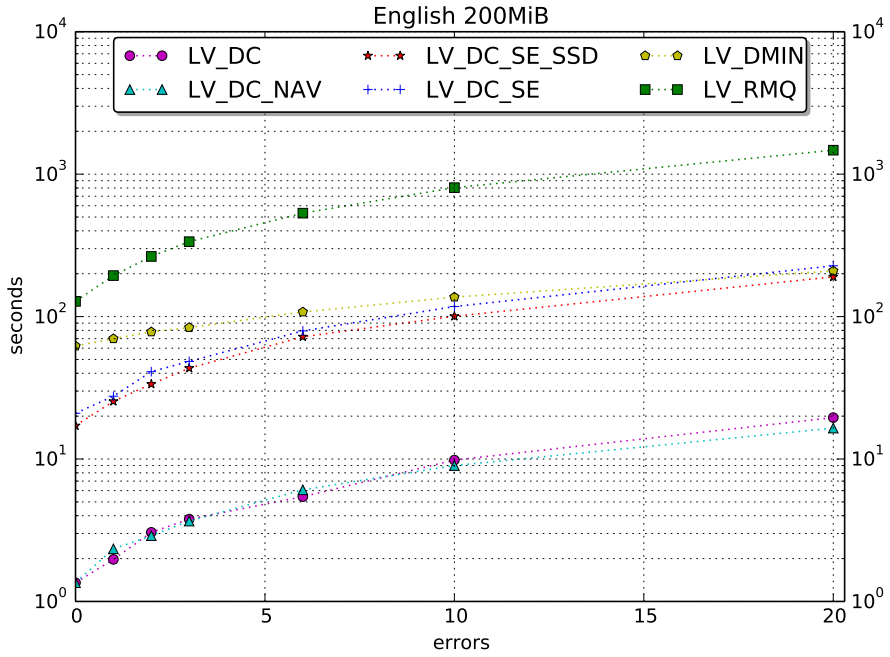


(a) 200MiB prefix DBLP-XML file.

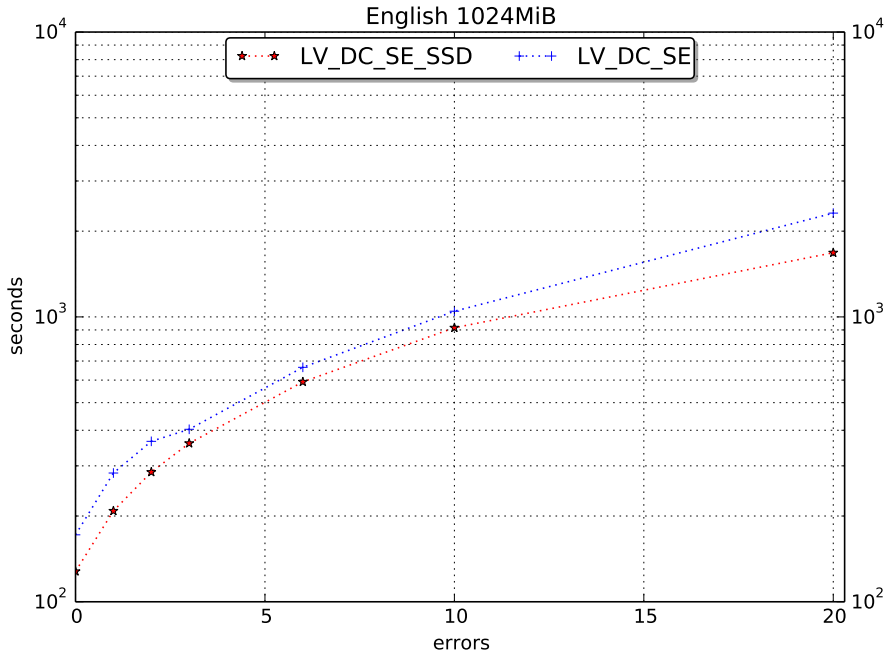


(b) 283MiB DBLP-XML file.

Figure 5: DBLP-XML files

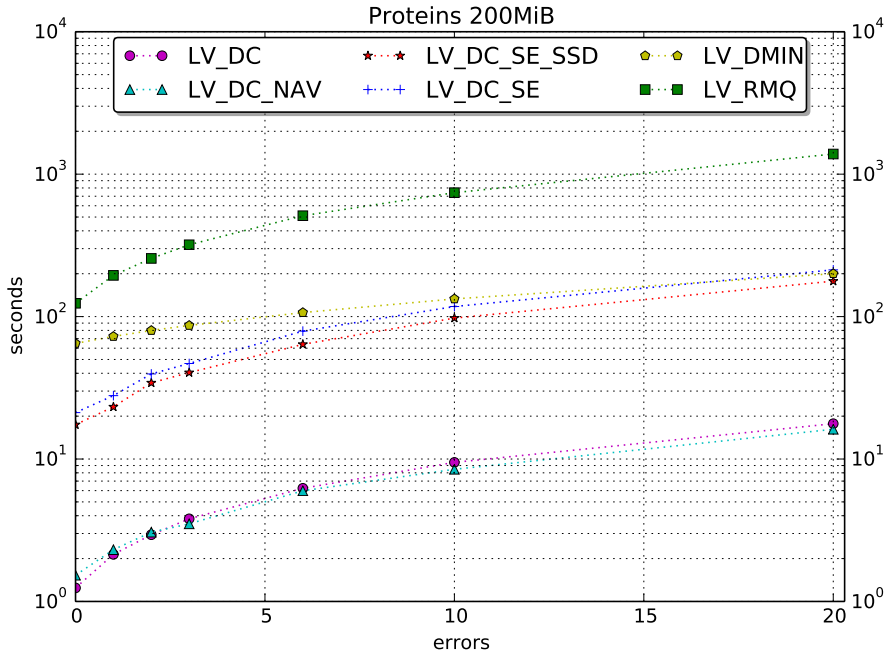


(a) 200MiB prefix English file.

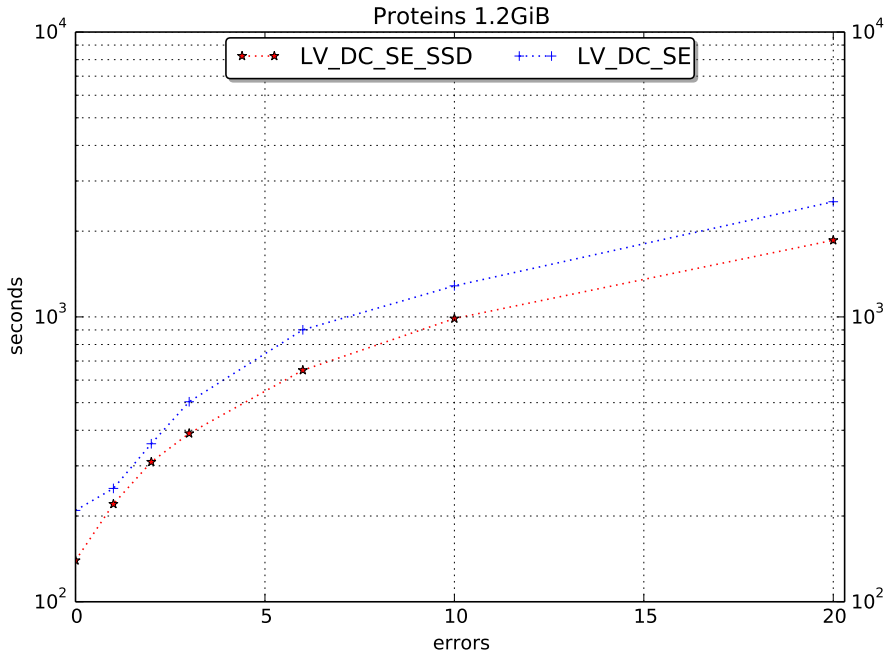


(b) 1024MiB English file.

Figure 6: English files



(a) 200MiB prefix Proteins file.



(b) 1.2GiB Proteins file.

Figure 7: Proteins files

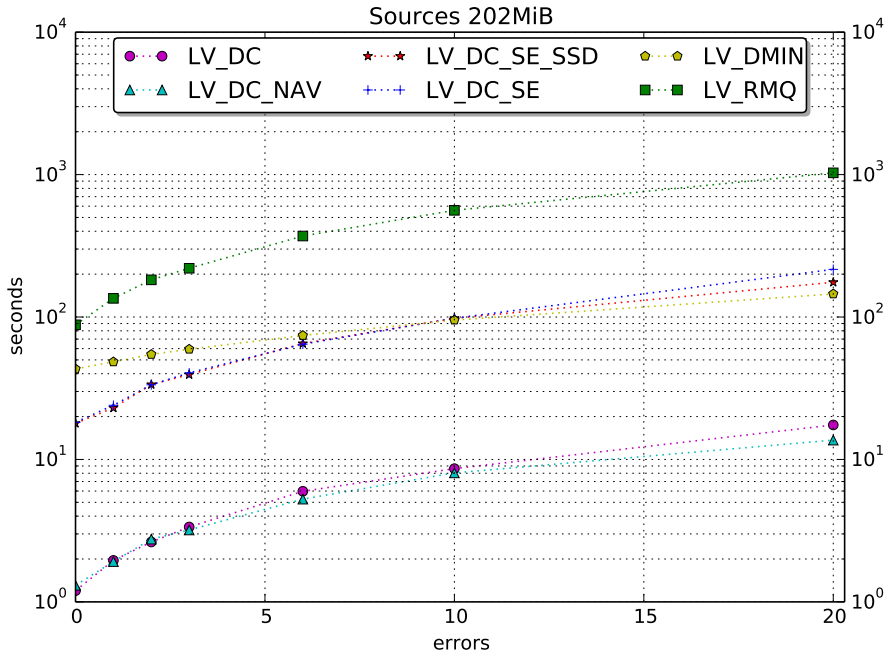


Figure 8: 202MiB Sources file