SPINE: Putting Backbone into String Indexing

Naresh Neelapala Romil Mittal Jayant R. Haritsa*
Dept. of Computer Science and Automation
Indian Institute of Science, Bangalore 560012, INDIA

Abstract

The indexing technique commonly used for long strings, such as genomes, is the suffix tree, which is based on a vertical (intra-path) compaction of the underlying trie structure. In this paper, we investigate an alternative approach to index building, based on horizontal (inter-path) compaction of the trie. In particular, we present SPINE, a carefully engineered horizontally-compacted trie index. SPINE consists of a backbone formed by a linear chain of nodes representing the underlying string, with the nodes connected by a rich set of edges for facilitating fast forward and backward traversals over the backbone during index construction and query search. A special feature of SPINE is that it collapses the trie into a linear structure, representing the logical extreme of horizontal compaction.

We describe algorithms for SPINE construction and for searching this index to find the occurrences of query patterns. Our experimental results on a variety of real genomic and proteomic strings show that SPINE requires significantly less space than standard implementations of suffix trees. Further, SPINE takes lesser time for both construction and search as compared to suffix trees, especially when the index is disk-resident. Finally, the linearity of its structure makes it more amenable for integration with database engines.

1. Introduction

A wide variety of applications require searching for exact or approximate matches over long text strings [6]. For example, performing global alignment between a pair of genomes that each run to millions or billions of nucleotides is a common task undertaken by biologists, the core operation of which is searching for maximal unique matches across the genomic strings [4]. Since brute-force searching techniques do not scale to such long strings, there has been extensive research on the design of high-performance *index structures* for strings.

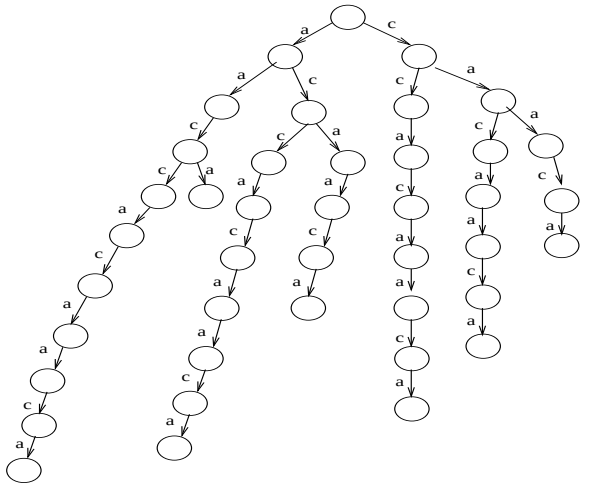

Figure 1. TRIE (for *aaccacaaca*)

In this context, a trie [13] which holds all suffixes of the data string, has become a popular starting point for developing index structures [6]. An example trie for the string aaccacaaca is shown in Figure 1. A space-efficient version of the trie structure, called the *suffix tree* [6], can be created by collapsing every unary node of the trie into its parent. When special edges called *suffix links*[1] are incorporated into this structure, it is possible to devise construction algorithms that have linear (in the size of the data) time and space complexity, and search algorithms that have linear (in the size of the query) time complexity. Given these remarkable performance properties, it is not surprising that suffix trees have become the defacto standard string index, featuring in popular tools like MUMmer [4], a global alignment software for genomes. The suffix tree corresponding to the example string aaccacaaca is shown in Figure 2.

From an abstract view-point, the suffix tree can be viewed as an *intra-path*, or "vertical", compaction of the

* Contact Author: haritsa@dsl.serc.iisc.ernet.in

1 A suffi x link connects a node representing aW to the node representing W, where a is a single character and W is an arbitrary string.

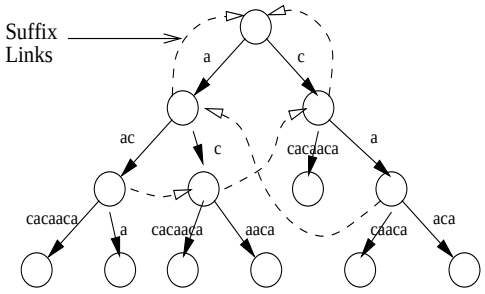

Figure 2. Suffix Tree (for *aaccacaaca*)

original trie since, as mentioned above, unary child nodes are merged into their parents. In this paper, we present a new index structure that is based on a novel *inter-path*, or "horizontal", compaction of the trie. Our motivation stems from the simple observation that there is considerable duplication of patterns *across* the various paths in the trie – for example, in Figure 1, the pattern *cacaaca* appears *thrice* in the trie structure. Eliminating this repetition holds out the promise of significantly reducing the number of nodes in the index and thereby reducing its resource consumption. However, achieving horizontal compaction is a significantly more complex task as compared to vertical compaction. This is because, unlike vertical compaction which is a simple structural merging that is independent of the content of the compacted nodes, horizontal compaction is based on merging character *patterns* across paths in the trie, thereby immediately running into the risk of generating *false positives* in the compaction process.

1.1. The SPINE Index

In this paper, we present a carefully designed horizontally-compacted trie index structure called SPINE (String Processing INdexing Engine). The SPINE index for the example string aaccacaaca is shown in Figure 3. As seen here, SPINE consists of a backbone formed by a linear chain of nodes representing the underlying data string, with the nodes connected by a rich set of edges for facilitating fast forward and backward traversals over the backbone during index construction and query search. All edges of the index are assigned labels during the construction process, and these labels are used to avoid false positives while traversing the index during the search process.

At a structural level, SPINE provides *all* the standard functionalities provided by suffix trees. Additionally, it has a variety of other attractive features:

- The entire trie is collapsed into a single *linear* structure, representing the logical extreme of horizontal compaction. Further, the number of nodes is *always* equal to the string length. This is in marked contrast to suffix trees where the number of nodes may go upto *double* the length of the string.

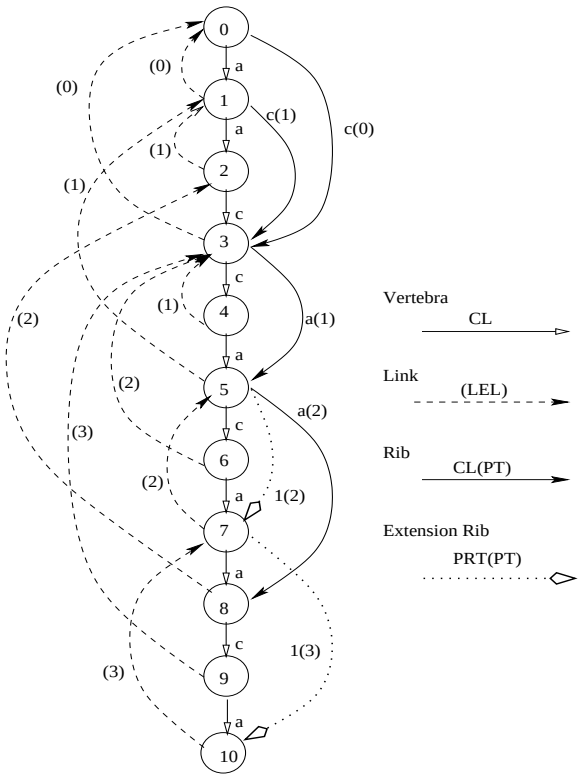

Figure 3. SPINE Index (for *aaccacaaca*)

- Since there is one edge (vertebra) on the backbone corresponding to each character in the string, the data string is *not required* any more once the index is constructed. This property does not hold for most of the other string indexes, including suffix trees.
- A SPINE index can be constructed in an *online* manner, not requiring prior knowledge of the entire data string. Further, a single SPINE index can be used to index multiple different strings, using techniques similar to those employed in Generalized Suffix Trees [6].
- Since index-growth always occurs at the tail of the structure, the node creation order and the node logical order are *identical* in SPINE. The utility of this feature is that it makes SPINE *prefix-partitionable* – that is, given a SPINE index for a string, the index for a prefix of this string is simply the corresponding initial fragment of the index.

  Note that the prefix-partitioning property is *not* supported by suffix trees since a node that is logically high up in the tree may be created much after nodes from lower levels in the tree.

- In comparing Figures 2 and 3, it might appear at first glance that a SPINE index may require more resources than a suffix-tree since it has 11 nodes and 26 edges while the suffix tree has 13 nodes and 16 edges. That is, the node reduction is offset by an increased number of edges. However, as discussed later in this paper, a variety of optimizations can be implemented to minimize the size of the SPINE index such that it is about a third smaller as compared to the equivalent suffix-tree.
- For finding all the matching substrings between two strings, the number of suffixes processed by SPINE is considerably smaller than those processed by suffix-trees because they process suffixes on an individual basis, whereas SPINE processes them on a *set* basis.
- Finally, due to the simple linearity of SPINE's structure, it is easy to develop efficient buffering policies, a mandatory requirement for good disk performance.

From a performance perspective, we demonstrate through a variety of experiments on real genetic strings, whose lengths are of the order of several millions of characters, that the horizontal compaction approach of SPINE results in significant improvements over vertical compaction. Overall, SPINE takes less space and time to construct, and has better search performance – that is, it wins on both construction and usage metrics. An implication of the lower space requirement is that, for a given memory budget, SPINE is able to process much longer strings than those supported by suffix trees. Even more attractive is that the performance differentials *increase* in moving from fully memory-resident indexes to disk-based implementations.

1.2. Contributions

To summarize, the main contributions of this work are the following:

1. We investigate horizontal compaction of tries and demonstrate that indexes that achieve complete horizontal compaction are feasible.
2. We describe the SPINE index structure and present online algorithms for its construction as well as for searching the index for query strings. We prove that, by virtue of the edge labeling strategy, the searches are guaranteed to not return false positives.
3. We present a variety of optimizations that drastically reduce the memory requirements of the SPINE index.
4. We profile the performance of SPINE against suffix trees over a variety of extremely long genetic strings for both memory-resident and disk-resident scenarios and show that SPINE offers significant benefits with regard to both space and time metrics.

1.3. Organization

The remainder of this paper is organized as follows: The SPINE structure is presented in Section 2, while the construction and search algorithms are described in Sections 3 and 4, respectively. The specifics of our prototype implementation are outlined in Section 5. Experimental results on the performance of this prototype are highlighted in Section 6. Related work is overviewed in Section 7. Finally, in Section 8, we summarize the conclusions of our study and outline future avenues to explore.

2. The SPINE Index Structure

In this section, we first overview the SPINE index structure and then describe its components in detail.

The central component of SPINE is the "backbone" of nodes connected by forward (or downstream) directed edges called "vertebras", as shown in Figure 3. Each vertebra corresponds to a character in the input data string, and this character is used to provide a *character label* (CL) for the vertebra. The vertebras appear in the same order as the associated characters in the input string.

While the backbone forms one source of forward connectivity between the nodes, there are additional downstream edges that connect nodes across the backbone. These edges are called "ribs" (full lines in Figure 3) and "extribs" (dotted lines in Figure 3). Similar to vertebras, each rib is labeled with a character label (CL), corresponding to the character that it represents in the associated suffix. The set of forward edges collectively represent all possible suffixes of the data string, and are used during the search process.

The backward (or upstream) edges, called "links" (dashed lines in Figure 3) are created and used during the SPINE construction process. They provide the ability to process suffixes on a set basis.

2.1. Avoiding False Positives

As mentioned in the Introduction, the SPINE index represents the complete horizontal compaction of all the suffixes in the corresponding trie. An implication of merging of all the matching paths into a single path is that all paths that were there in the original trie continue to be represented in SPINE, and therefore there is no possibility of *false negatives*. However, *false positives*, that is, invalid substrings, may arise. For instance, in Figure 3, a path for accaa appears to exist in the SPINE index even though it is not a substring of the data string.

To avoid such false positives, we take recourse to a numeric labeling strategy for the edges during the construction process. Specifically, each rib and extrib is assigned an integer label, called *Pathlength Threshold (PT)*. The extribs

have an additional integer label called *Parent Rib Threshold (PRT)*. In order to be able to assign the correct PT values to the ribs/extribs, each link is assigned an integer label called *Longest Early-Terminating suffix Length (LEL)*. For example, in Figure 3, the rib from Node 3 has a PT of 1, the extrib from Node 5 to Node 7 has a PRT of 1 and PT of 2, while the link from Node 8 to Node 2 has an LEL of 2. These labels, which are assigned during the index construction process, determine when forward edges can be traversed during the subsequent search process, as described in Section 4.

2.2. Notation and Terminology

In the remainder of this section, we describe the components of SPINE in detail. Our discussion assumes that the data string which is being indexed is composed of M characters. For ease of presentation, we use the notation shown in Table 1. While the table entries are mostly self-explanatory, the *termination* concept requires elucidation: A suffix s_{ij} is said to terminate at node p ($p \leq i$) if there is a valid traversal path from the root node to node p whose string of character labels match the suffix. A suffix s_{ij} whose termination node is strictly less than i is said to be an *early-terminating* suffix, otherwise it is called *end-terminating*.

To make the above notation clear, consider Node 5 in Figure 3, for which S_5 = aacca, s_{52} = ca, $AllSuf_5$ = {aacca, acca, cca, ca, a}, $EndSuf_5$ = {aacca, acca, cca, ca}, and $EarlySuf_5$ = {a}.

Notation	Meaning
N_i	*Node i*
S_i	*String on backbone from root to* N_i
s_{ij}	*Suffix of* S_i *of length* j
$AllSuf_i$	*Set of all suffixes of* S_i
$AllSuf_i(k)$	*Set of suffixes of* S_i *of length* $\leq k$
$EndSuf_i$	*Set of suffixes in* $AllSuf_i$ *terminating at* N_i
$EndSuf_i(k)$	*Set of suffixes of* $EndSuf_i$ *of length* $\leq k$
$EarlySuf_i$	*Set of suffixes in* $AllSuf_i$ *not terminating at* N_i
$Link(N_i).LEL$	*LEL of the link of* N_i
$Link(N_i).Dest$	*Destination node of link of* N_i
$Rib(N_i).Dest(c)$	*Destination node of rib at* N_i *for character* c
$Rib(N_i).PT(c)$	*PT of rib at* N_i *for character* c

Table 1. Notation

2.3. Vertebra Backbone

During construction, the backbone is initially created with a single node, called the *root node*, and for each character in the data string, a new node is added sequentially using a vertebra edge labeled with the corresponding character. The node that is currently at the bottom of the backbone is referred to as the *tail node*, N_{tail}. Each node has an integer identifier which is set equal to the length of the backbone string above that node. With this naming convention, the root node has identifier 0, the first node has identifier 1, and so on until the tail node of the entire index which will have identifier M.

2.4. Links

Links are meant to record, at each node, the information about the node's early-terminating suffixes, namely, $EarlySuf_i$. Specifically, only the *longest* early-terminating suffix (hereafter referred to as LET-suffix) is explicitly kept track of since, by definition, all shorter suffixes are also early-terminating suffixes and they would themselves have been linked up earlier. For example, in Figure 3, there is a link from N_5 to N_1 to represent a, the LET-suffix. If a node has no early-terminating suffixes (i.e. $EndSuf_i = AllSuf_i$), then its link points to the root node, N_0, which can be interpreted as representing the null suffix. N_3 in Figure 3 is an example of this scenario. Finally, as a special case, the root node has no link edge since it is the starting node.

Link Labels The LEL label of a link is the length of the LET suffix which it represents. Intuitively, if we have a link from N_i to N_j with a LEL 'k', then it means $s_{ik} = s_{jk}$. More formally, $AllSuf_i$ can be defined as follows:

$AllSuf_i = EndSuf_i \cup EarlySuf_i$ and
$EarlySuf_i = AllSuf_j(k)$

where $k = Link(N_i).LEL$ and $j = Link(N_i).Dest$.

2.5. Ribs

When the SPINE index that has been built for S_i is extended by one more character from the data string, we need to extend all the suffixes of S_i by this additional character, c_{tail}. For the end-terminating suffixes, the newly added node on the backbone, N_{tail}, *automatically* records this extension through its vertebra edge. For the early-terminating suffixes, however, the extension must be explicitly recorded and this is achieved through the addition of rib edges. Specifically, the link chain from N_i is traversed and if a rib/vertebra does not already exist for c_{tail} at any node, say N_j, in the link chain, a new rib is created from node N_j to N_{tail}.

The traversal of the link chain terminates if either the root node is reached, or a node having an outgoing edge labeled with c_{tail} is reached. The first stopping condition is obvious since no further traversal is possible, while the other condition reflects the fact that the suffix in question has *already* been previously extended. And there is no need to explicitly handle the remaining smaller suffixes as they would also have been extended automatically.

Rib Labels When a new rib is created at N_j, its CL is set to c_{tail}, and its Pathlength Threshold (PT) is set to the length of the longest suffix of S_i terminating at that node, which is given by the LEL of the last traversed suffix link. Intuitively, the rib PT represents the length of the longest prefix that can be traversed from the root before the rib is traversed. This is because the rib was created to extend the suffix of that length.

2.6. ExtRibs

As mentioned above, we stop the link-chain traversal for rib addition if we find that the current node already has a matching rib (i.e. with CL = c_{tail}). However, the following situation may now arise: The PT of the pre-existing rib may be *less* than the LEL of the link used to reach this node, which means that this rib is *not valid* to represent the extension of the associated early-terminating suffix. To address this issue, the solution that immediately comes to mind is to update the rib's PT to be equal to the LEL value. However, this is not correct since it may permit illegal paths resulting in false positives. We therefore take an alternative approach of extending the rib itself through edges called *extribs* (extension ribs). For example, in Figure 3, the extrib (dotted line) from N_5 to N_7 is an extension of the "parent" rib connecting N_3 to N_5.

At a given node, there may be multiple extribs, each corresponding to a different parent rib that terminates at this node. From an implementation perspective, this is problematic since it makes the node size to be variable. Therefore, we take the alternative approach of maintaining the extribs in a *chained* fashion. That is, the first extrib in the chain is located at the destination node of the rib which failed the validity test, and the second extrib is located at the destination node of the first extrib, and so on. This ensures that at any node there is at most *only one extrib*. So, whenever we need to create an extrib, instead of creating it from the destination of the parent rib, we traverse to the node at the end of the extrib chain, and then create a new extrib from this node to the tail node. All the extribs created for a rib are its children.

ExtRib Labels Each extrib has an associated Pathlength Threshold (PT), which is the length of the longest suffix that it is extending, as well as a PRT, which is the PT value of the parent rib. The reason for including the PRT value is to be able to uniquely identify the extrib. Note that a character label is not required for an extrib as it is implicitly represented by the CL of the incoming rib or extrib at its source node. And hence, a complete extrib chain represents a single character. In Figure 3, an example chain is the extrib from N_5 to N_7, and then from N_7 to N_{10}.

2.7. Prefix-Partitioning

It is easy to see from the above discussion that SPINE is *prefix-partitionable*, i.e. given a SPINE index for a string, the index for a prefix of the string is simply the corresponding initial fragment of the index.

3. SPINE Construction Algorithm

In the previous section, we presented an overview of the SPINE index structure. We now move on to presenting an *online* algorithm for constructing this structure. The pseudo code for the main algorithm is given in Figure 4 (subroutines are described in [12]).

We start off with the SPINE index initially consisting of just the root node and then, for each new character in the string, a node is appended to the tail of the index. The vertebra connector to the newly-added node is labeled with the new character, and the associated links and ribs are created as required.

As mentioned earlier, every node, excepting the root, has a link associated with it. When the first character is appended, a link is created from the new node to the root node. For all subsequent nodes, the following process is followed: The link edge of the immediate predecessor of N_{tail} is traversed upstream. Let the destination node of this link be N_{curr}. At N_{curr}, it is checked whether a vertebra/rib already exists for c_{tail}. If it is not present, a new rib is constructed from N_{curr} to N_{tail}. Then, the link at N_{curr} is traversed upwards and the same process is repeated with the new N_{curr}.

The above process stops with the creation of a new link, which happens when one of the following cases occur during the upward traversal of the link chain:

A vertebra is found with CL = c_{tail}: In this case, a link is created from N_{tail} to the destination node of the vertebra. The LEL of the link is set one greater than the LEL of the last link traversed.

A rib is found with CL = c_{tail}: In this case, if the threshold test does not fail, then a link is created from N_{tail} to the node referenced by the rib, and the LEL of the link is set one greater than the LEL of the last link traversed.

Otherwise, the extrib chain is traversed to find a child extrib with PT $\geq$ LEL of the link. If found, then a

link is created from N_{tail} to the destination of that extrib. The PT of the link is set one greater than the PT of the last link traversed. Otherwise, a new extrib is created from the end of the extrib chain to N_{tail} and a link is also created from N_{tail} to the destination node of the last traversed extrib with PRT equal to the PT of the rib which failed the validity test. The link is assigned a LEL which is one more than the PT of the last rib or sibling extrib that has been traversed.

A rib is created from the root node: Here, a link with LEL set to 0 is created from N_{tail} to the root node.

APPEND $(n+1)^{th}$ character
01. c_{tail} = $(n+1)^{th}$ character
02. Append N_{n+1} to the SPINE using a vertebra
03. N_{tail} = N_{n+1}
04. N_{curr} = $Link(n).Dest$
05. *edgeFound* = FALSE
06. WHILE (NOT *edgeFound*)
07. IF ($N_{curr} \neq NULL$)
08. l = Most recently traversed link
09. Check for a c_{tail} vertebra/rib at N_{curr}
10. IF (a matching edge e is found)
11. IF (e is vertebra)
12. AddLink($N_{tail}, e.Dest, l.LEL + 1$)
13. ELSE IF (e is a rib)
14. IF ($l.LEL > e.PT$)
15. HandleExtribs($l.LEL, e.PT$)
16. ELSE
17. AddLink($N_{tail}, e.Dest, l.PT + 1$)
18. edgeFound = TRUE
19. ELSE
20. AddRib($N_{curr}, N_{tail}, c_{tail}, l.LEL$)
21. N_{curr} = $Link(N_{curr}).Dest$
22. END-IF
23. ELSE // link chain ends
24. AddLink (N_{tail},N_{root},0)
25. edgeFound = TRUE
26. END-IF
27. END-WHILE

Figure 4. SPINE Construction Algorithm

3.1. Construction Example

To help clarify the above discussion, we now describe how the SPINE index is created for the same input string used in Figure 3, i.e. aaccacaaca.

In the beginning, a root node is created with identifier 0. Subsequently, whenever a new node is added to the backbone, we start traversing the link chain beginning from the parent node of the newly added node. An example scenario for each of the conditions in the construction algorithm given in Figure 4 is discussed below.

CASE 1: Vertebra Exists (Line 11)
This case occurs when a vertebra for c_{tail} exists at N_{curr}. For example, consider appending N_2. Here, we traverse the link of N_1 to reach N_0 and find a vertebra for a. Hence, we create a link from N_2 to N_1 and assign it a LEL of 1 (= LEL of last traversed link + 1).

CASE 2: Rib With Required PT Exists (Line 17)
This case occurs when there already exists a rib/vertebra for c_{tail} with sufficient PT. Consider appending N_4. In this case, we find that a rib for c with sufficient PT exists at N_0. Hence, a link is created from N_4 to N_3 (the destination of the rib) with a LEL of 1 (= LEL of last link traversed + 1).

CASE 3: Rib Creation (Line 20)
This case occurs when there exists no rib/vertebra for c_{tail}. Consider appending N_3. Traverse the link of N_2 to reach N_1. Since there exists no rib/vertebra for c, create a rib from N_1 to N_3 and assign it a LEL of 1. Now traverse the link of N_1 to reach N_0. Again, since no rib or vertebra exists for c, a rib is created for character c from N_0 to N_3 with PT equal to 0. Since the root node has no link, we end the process by creating a link from N_3 to N_0 with LEL = 0.

CASE 4: ExtRib Creation (Line 15)
This case occurs when there exists a rib whose PT is less than the desired value. Consider appending N_7. Traverse the link of N_6 to reach N_3. At N_3, there exists a rib for character a but with PT of 1 which is less than the LEL of the last traversed link(= 2). And we see that there is no extrib from N_5 (the destination node of the rib). So, an extrib is created from N_5 to N_7 and its PT and PRT are set to 2 (LEL of the last traversed link) and 1 (PT of the parent rib), respectively. Then, a link is created from N_7 to N_5 (the last traversed rib/extrib with the same PRT as the newly created extrib) with a LEL of 2 (= PT of last traversed rib/extrib with same PRT + 1).

4. Searching with SPINE

In this section, we discuss how a SPINE index can be used for efficient searching. We begin by defining valid search paths, then present an example search process, and conclude with a comparison of searches in suffix-trees.

Valid Paths A search path in a SPINE index is a *valid path* if and only if (a) the path originates at the root node, and (b) all the ribs/extribs in the traversed path satisfy the Path-length Threshold (PT) constraints. A formal proof that the

valid paths in the SPINE index correspond exactly to the set of substrings that occur in the data string is given in [12].

To make the above clear, a rib/extrib can be traversed only if the length of the path traversed so far (i.e. from the root node till that point) is less than or equal to the PT of the rib/extrib. For example, the accaa path will not be permitted in Figure 3 because when we traverse the path from the root for acca, after we reach Node 5, the rib for a violates the constraint since its PT of 2 is less than the current pathlength of 4. Thus, accaa is not a valid substring of the given data string.

Search Example For illustrative purposes, we will assume a complex matching operation wherein the goal is to find, given a data string S1 on which a SPINE index has been built, and a query string S2, all maximal matching substrings, including repetitions, between S1 and S2, whose lengths are above a threshold value. A practical application of this matching operation is in establishing local alignments across genetic strings.

For example, given the following strings S1 and S2, and a threshold value of 6, the output should contain the substrings shown in boldface.

S1 *acacc***gacgat***acga***gattacgaga***cgagaatacaacag**
S2 *catagagagac***gattacgaga***aaacgggaaa***gacgat***cc**

For the above operation, the SPINE matching would proceed as follows: To start off, the entire query string is searched for in the SPINE index of the data string. As soon as the first mismatch is found, the length matched so far is reported. Now, we check if the mismatched character follows any of the shorter suffixes in the matched part of the query string, and the process is repeated again. The shorter suffixes are reached by traversing the link chain upwards.

The procedure for finding a match is as follows: We start at the root and traverse the forward edges (vertebras, ribs and extribs) according to the characters in the query string. A vertebra edge can be traversed at any time. Before traversing a rib, however, a check is made as to whether the length traversed thus far is $\leq$ PT of the rib. If this test fails, then this rib's extrib chain is followed until either the extrib chain ends, or we find the child extrib whose PT is greater than or equal to the current pathlength and having a PRT which is equal to PT of the rib which failed the test.

The intuition behind our searching scheme is simple: Each valid path starting from the root to a node corresponds to some suffix of the string on the backbone till that node. And while more than one suffix might terminate at a node, each such suffix would be of a different length. So, at a given node if it is valid to traverse a rib after a pattern p (suffix till that node) of length k, then it has to be valid after a pattern q whose length is less than k and which ends at the same node, because q would be a suffix of p.

The above matching process finds the first occurrence of a match in the data string. But our goal is to find *all* occurrences of the match. This is achieved using a simple technique that exploits the property that a link with *LEL v* from node N_a to node N_b indicates that a string of length *v* above N_a is the same as the string of length *v* above N_b. Specifically, after we find the first occurrence of the match, the node indexing the first occurrence is stored in a *target node buffer*. Then, all the nodes downstream are scanned successively to check if their links point to the node in the target node buffer, i.e. the node indexing the first occurrence and have an LEL greater than length of pattern being searched. If so, then that node is also stored in the target node buffer. Again, downstream scanning is started from this node and the process is repeated until the end of the backbone is reached. Searching in the target node buffer is performed in binary fashion to improve the performance.

To clarify the above, consider Figure 3 with a query string ac. Here, after locating the first occurrence, the target node buffer will contain N_3. Moving downstream, at N_6 we find a link with LEL = 2 (length of string ac) pointing to N_3. And so, N_6 is also added to the target node buffer. On moving further downstream, at N_9, a link with appropriate LEL is found pointing to a node in the target node buffer (N_3), and therefore it is also added to the buffer. In this manner, the target node buffer finally gives the *end* nodes of all occurrences of the pattern in the string. As a last step, their starting positions can be trivially determined by merely subtracting the query pattern length from each of the node identifiers in the target node buffer.

While we could, in principle, search for all occurrences of a matching pattern immediately after it is found, this would be wasteful since it would require a traversal of the backbone for each matching pattern. Instead, we defer this step until the first occurrences of all matches are found, and then, in one single final sequential scan of the backbone, the repeated occurrences of all matching patterns are concurrently found.

The detailed pseudocode to find the first occurrence of the query string in the data string is available in [12].

4.1. Comparison with Suffix-Trees

Similar to SPINE's use of links, suffix-trees use *suffix links* to assist in finding the suffixes of the matched substrings. But, the number of suffixes checked by suffix-tree search algorithms is *far more* than those checked by SPINE. The reason for this is as follows: In suffix trees, a suffix link points from a node indexing string aW to the node indexing W, where 'a' is a character and W is a string [6]. In the case of a mismatch, after checking for aW, we retrieve the node indexing the suffix W and check if the mismatched character follows W. This process iterates till a complete

match is found or there are no more suffixes remaining to be checked.

In SPINE, however, each node N_i in a link chain represents a *set* of suffixes, namely $EndSuf_i$. Therefore, only *one check* is sufficient for all the suffixes in that set, reducing the computational effort.

To make the above analysis clear, consider the index structures shown in Figures 2 and 3. Here, assuming that while matching accaa a mismatch is found in the suffix-tree after matching acca, then the next suffix to be checked will be cca (length 3) i.e. the one indexed by the destination node of the suffix link. On the other hand, in SPINE, the link from Node 5 *directly* points to Node 1 which represents the suffixes of length 1 or less. This means that the (unnecessary) checks for suffixes of length 3 and length 2 are not made. Therefore, for long strings, only a small number of suffixes are actually checked in the SPINE index.

5. Implementation Details

We have developed a prototype version of SPINE, and in this section, we discuss its implementation details. While SPINE is general in its applicability, for ease of presentation, we assume in the following discussion that it is DNA genomic strings, which are over an alphabet of size four, that are being indexed; proteomes, which are over an alphabet of size twenty, are discussed at the end of the section.

Our implementation strategies are based on our experience with a variety of DNA genomes, each of which is several million characters in length. In particular, we will present results for the following representative genomes:

ECO : E.coli earthworm genome of length 3.5 million characters;

CEL : C.Elegans bacterial genome of length 15.5 million characters;

HC21 : Human chromosome 21 genome of length 28.5 million characters;

HC19 : Human chromosome 19 genome of length 57.5 million characters.

Field Name	Space (Bytes)	Count	Total (Bytes)
CharacterLabel	0.25	1	0.25
VertebraDest	4	1	4
Link Dest	4	1	4
Link LEL	4	1	4
Rib Dest	4	3	12
Rib PT	4	3	12
ExtRib Dest	4	1	4
ExtRib PT	4	1	4
ExtRib PRT	4	1	4

Table 2. Index Node Content

The information associated with each node of the SPINE index and the associated space requirements are shown in Table 2, corresponding to storing one vertebra, one link, a maximum of three ribs (for DNA alphabet), and one extrib at the node. As can be seen from the table, with a straightforward implementation, the worst-case space required by each node is huge (48.25 bytes). However, the SPINE index exhibits a variety of both structural and empirically-observed features using which the actual space required can be drastically reduced – in fact, as we will show next, it can be brought down to *less than 12 bytes* when all these features are taken into account.

5.1. Node Size Optimizations

In this section, we present the optimizations which reduce the space requirements of SPINE index.

Implicit Vertebra Edge Since, as mentioned earlier, SPINE grows sequentially at the tail of the backbone, the physical order and the logical order of the nodes are identical. We can take advantage of this feature to not explicitly represent the vertebra edge destination since the neighboring nodes are physically contiguous, i.e. the *Vertebra Dest* field can be eliminated.

Small Numeric Label Values Table 3 gives the maximum value observed for the various numeric labels (PT, LEL, PRT) when the SPINE index was constructed on the representative biological genomes mentioned above. As can be observed here, the label values *never exceed* 25000 even for very long genomes like the human chromosomes. Therefore, only two bytes, rather than four, need to be allocated for the length fields.

Genome	Max Value
ECO	1785
CEL	8187
HC21	21844
HC19	12371

Table 3. Maximum Label Values

However, to ensure that the index works robustly, we have a mechanism in place to handle even those rare cases where the numeric values may exceed 65536 (the maximum value that can be represented in two bytes). We allocate separate entries for these cases in an *overflow table*. The node space normally used for storing the label value is now used to index into the overflow table, and a one-bit flag is used in the main node structure to indicate whether the space is storing a value or a pointer.

Sparse Rib Distribution While all nodes have upstream edges (links), the same is not true with respect to downstream edges (ribs/extribs). In fact, we found that only around *30 to 35 percent* of the nodes have downstream edges emanating from them – Table 4 shows the distribution of their number for the various genomes. Specifically, the columns labeled 1 through 4 represent the percentage of nodes having that many forward edges emanating from them, with the maximum corresponding to having the full complement of downstream edges (3 ribs and 1 extrib).

Genome	Number of Ribs				Total
	1	2	3	4	
ECO	15%	9%	6%	4%	33%
CEL	15%	8%	6%	4%	33%
HC21	14%	8%	6%	4%	32%
HC19	13%	7%	5%	3%	28%

Table 4. Rib Distribution across Nodes

The reason for this distributional behavior is that after some length of the data string has been processed, the remaining part mostly contains *repetitions* of previously occurred patterns, and therefore fresh downstream edges are rarely created. Based on this observation, we do not allocate space for downstream edges at every node, since considerable space would be wasted. Instead, we store information about the links and the downstream edges separately in a *Link Table* (LT) and a *Rib/Extrib Table* (RT), respectively. One entry for each character in the string is allocated space statically in the *LT*, while space for downstream edges is allocated dynamically in the *RT* for only those nodes from which a rib/extrib emanates. Therefore, the total number of entries in the RT is less than 35 percent of that in the LT.

Further, from Table 4 it is clear that the number of nodes with a given rib fanout decreases with the fanout value. For example, only about 4% of the nodes have the full complement of downstream edges. Therefore, to avoid the space wasted for the edges which are not present, we use *multiple* RTs. Specifically, there is one RT for each possible fanout, resulting in four RTs in total: *RT1, RT2, RT3, and RT4*.

While this optimization results in considerable space savings, it might appear at first glance that the construction time of SPINE would degrade due to the movement of nodes across the RTs, which would occur whenever a node acquires an additional downstream edge. However, we have experimentally observed that this impact is negligible.

Final Node Layout Based on the above discussion, the optimized implementation of the SPINE index consists of a **Link Table** (LT) and four **RibTables** (RTs), whose entries are shown in Figure 5. The LT contains one entry for each node (character) in the string. It stores its LEL as one of its columns while the other column represents either the destination node of that link (the LD field) or a pointer to an entry in one of the RTs (the PTR field). In particular, the LT stores the link destinations only for the nodes that don't have any ribs/extrib. For the remaining nodes, they are stored in the RT entries only.

Each node features in at most one RT table. A RT entry for a node stores the destination node of the link from that node and also the destination nodes (the *RD* fields) and the threshold values (the *PT* fields) of all the ribs/extrib emanating from the node. And, lastly, the *PRT* field denotes the PRT value of the extrib.

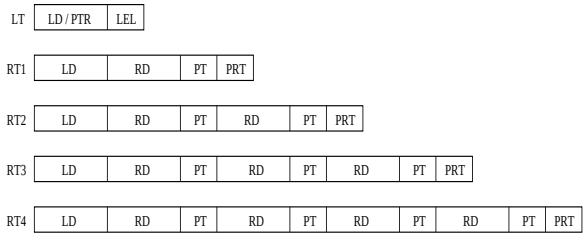

Figure 5. Optimized SPINE Implementation

By implementing all the above optimizations, the net effect is that the average node size in SPINE is *less than 12 bytes*, that is, the index takes upto 12 bytes per indexed character. The advantage of smaller node sizes is reflected not only in space occupancy but also in improved construction and searching times, as is quantitatively demonstrated in the following section.

5.2. SPINE Implementation for Proteins

The above implementation focused on DNA strings which have an alphabet size of 4. When we consider proteins strings, where the alphabet size increases to 20, we observed that the numeric label values are *even smaller* than those found with DNA strings. Our experiments were conducted with the E.Coli Residue (1.5 M), Yeast Residue (3.1 M) and Drosophila Residue (7.5 M) proteomes. Moving on to the rib distribution, we observed that here too there is a steep decay in the percentages of nodes having multiple ribs. And again, the total number of nodes with any rib/extrib is less than 30%. Therefore, the overall behavioral characteristics of proteomes are similar to that of genomes with the only practical difference being that each character label requires 5 bits to code as opposed to the 2 bits used for DNA.

6. Experimental Analysis

We conducted a detailed evaluation of the performance of the SPINE index prototype, and these results are pre-

sented in this section. Our experiments were conducted with the same set of genomes mentioned earlier in this paper (i.e. E.Coli, C.Elegans, HumanChromosome 21, and HumanChromosome 19). For comparison purposes, we also evaluated the performance of the suffix tree, hereafter referred to as ST – the code base was taken from the MUMmer software [4] to reflect an industrial-strength implementation.

Our experiments were conducted on a Pentium IV 2.4 GHz machine with 1 GB RAM, 40GB IDE disk and running Linux 7.3 operating system. The performance metrics in our experiments were the following:

Index Construction Time: This is the overall time taken to build the complete index for a string.

Index Search Times: This refers to the time taken to perform the complex matching operation discussed in section 4, wherein we need to output all maximal matching substrings, including repetitions, between the source strings.

6.1. In-Memory Environment

The performance of ST and SPINE with regard to in-memory index construction times are shown in Figure 6. Firstly, note that the indexes take less than two seconds construction time per Mbp, which means that with sufficient resources, a complete in-memory index construction of the human genome (approximately 3Gbp in length) can be done in under two hours. Second, SPINE takes only marginally lesser time to construct than ST, especially for longer strings. This is not surprising since all operations are done in memory and therefore the structural differences do not really play a role in determining the construction time. But, these features do show up with regard to the *maximum string length* that can be successfully handled for a given budget. This is evident in Figure 6, where no results are shown for ST with regard to the HC19 string as it ran out of memory due to its larger space requirements. In contrast, SPINE was able to complete the index build successfully – in general, SPINE can handle approximately *30 percent more string length* than the maximum that can be supported by ST.

Moving on to the search times, Table 5 gives the times required to find all the exactly matching substrings (including multiple occurrences) for SPINE and ST for various genome pairs. We observe here that SPINE takes around 30 percent lesser time than ST. This is entirely due to its efficiency in handling a much smaller number of suffixes, as described earlier in section 4.1, and quantitatively shown in Table 6.

We hasten to add here that while the results we show here are for complete genomes in order to demonstrate scalability, the same performance differences held even when the query strings were much smaller (e.g.,, of length 1K).

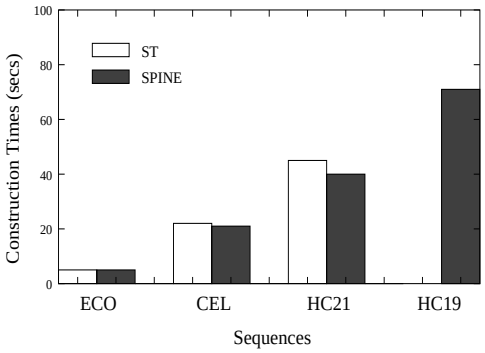

Figure 6. Index Construction Times (In Memory)

Data Seq	Query Seq	ST	SPINE
ECO	CEL	20	16
CEL	HC21	45	31
HC21	CEL	26	17
HC21	HC19	83	54
HC19	HC21	–	30

Table 5. Substring Matching Times (secs)

Data Seq	Query Seq	ST	SPINE
CEL	ECO	3515	2119
HC21	ECO	3514	2163
HC21	CEL	15077	8701

Table 6. Number of Nodes Checked (In 1000s)

6.2. Performance on Disk

We now move on to assessing the performance of SPINE and ST on disk. Note that while SPINE can be expected to have a basic advantage due to its smaller node size, the more important issue here is the *locality* of the accesses made by the index structures.

To study their behavior, we constructed generic SPINE and ST indexes on disk without any extra disk-specific optimization. Further, the indexes were constructed using synchronous I/O (O_SYNC option) for writes to minimize the modulation of the locality behavior by other system factors. Figure 7 shows the time taken to construct the indexes for the various genomes on disk. We see here that SPINE takes almost half the time as required by ST to construct the index on disk. Note that this cannot be attributed solely to the smaller-sized nodes since that would have at best reduced the time by a factor of about 30%. The additional

20% improvement arises due to the better locality exhibited by SPINE.

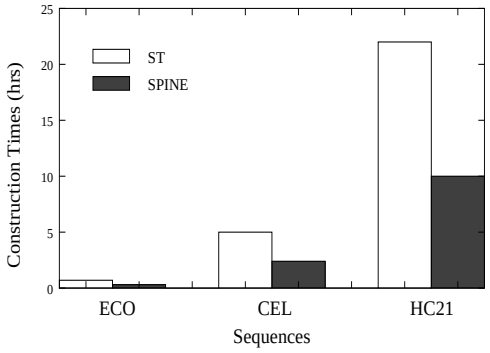

Figure 7. Index Construction Times (On Disk)

We investigated the issue of locality further and an interesting feature that we observed in the SPINE index is that most of the links point to the *upper* nodes in the backbone, and that the number of links pointing to a node keeps monotonically decreasing as we descend the backbone. This is shown quantitatively in Figure 8, which shows the distribution of the link destinations for different data strings. This indicates that while constructing the SPINE, the upstream nodes would be accessed more often than those downstream.

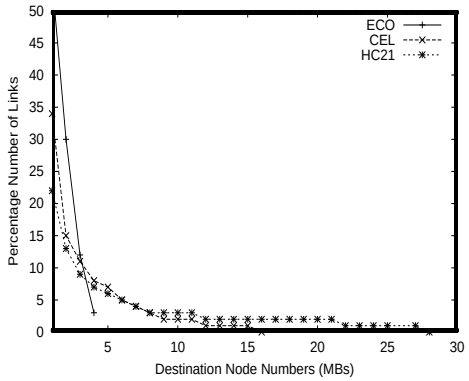

Figure 8. Link Distribution over the Backbone

The above observation suggests a *simple buffering strategy* for SPINE, when sufficient memory is not available: *"Retain as much as possible of the top part of the Link Table in memory"*.

Moving on to index search times, we observed that the time required to obtain all the exactly matching substrings also improved by a factor of two with SPINE as compared to ST. This is explicitly shown in the speedup numbers of Table 7 for the various genome combinations (the large absolute times are due to our synchronous disk-write artifact).

Data Seq	Query Seq	MUMmer	SPINE	Speedup
CEL	ECO	0.98	0.47	52.1%
HC21	ECO	0.97	0.48	49.8%
HC21	CEL	4.30	2.02	52.8%
HC19	HC21	7.92	3.87	51.1%

Table 7. Substring Matching (On Disk, in hours)

Due to space limitations, we do not present performance results for protein strings here, but our experiments with these strings showed that the SPINE construction times for proteins also scaled linearly with the string lengths, and that the search times are independent of the data string length. Overall, SPINE works as well with protein strings as it does with DNA strings.

7. Related Work

A rich body of literature exists with regard to vertically compacted trie indexes such as suffix trees. The primary focus of this research has been on optimizing the space occupied by the tree nodes – for example, an implementation that requires 12.5 bytes per indexed character for DNA strings was proposed in [9]. The point to note, however, is that these optimizations only increase the *maximum length* of the data string that can be hosted in memory but do not improve the construction times and search times of the basic suffix tree.

In fact, some of these optimizations *adversely impact* the performance or the functionality of the tree. For example, an extremely space-efficient implementation, called Lazy Suffix Trees [5], has been recently proposed, taking only 8.5 bytes per indexed character. However, it has constraints on its functionality, including not being online, and not being able to perform approximate and substring matching efficiently due to the absence of suffix links. Similarly, suffix arrays [11] reduce the space requirement to just 6 bytes per indexed character but increase the time complexity from linear to supra-linear. In summary, we can expect that the timing benefits of SPINE with regard to ST, which were demonstrated in this paper, will carry over to ST's different implementation flavors.

In contrast to vertical compaction, there is almost no prior work available with regard to horizontal trie compaction. The only exception that we are aware of in this regard is *DAWGS - Direct Acyclic Word Graphs* [2], which require around 34 bytes per character for DNA strings [9]. A compacted version of DAWGS, called *CDAWGs* [10] was proposed, but that too requires more than 22 bytes per indexed character [9]. Unlike SPINE, they are unable to achieve complete horizontal compaction due to their tech-

nique for eliminating false positives. Further, they lack position information of the matching pattern in the data string because their nodes do not correspond to the character positions in the string.

Recently, in order to make suffix-tree construction on disk efficient, a partition-based technique was proposed in [7]. This algorithm is predicated on dispensing completely with the suffix links that are essential for retaining the linear time construction complexity – as a result, the algorithm in [7] has quadratic complexity.

An elegant two-level search technique called MRS-index was recently proposed in [8], wherein a preprocessing phase using a very small approximate index is used to first filter out those regions of the data string that potentially contain matching entries, and then a seed-based approach is used on the filtered regions. While MRS gives only approximate answers, both SPINE and ST provide exact answers. Further, the performance improvement through complete indexes is typically substantially more, albeit at the cost of increased resource consumption [12].

8. Conclusions

In this paper, we have proposed the SPINE index data structure, which achieves a complete horizontal compaction of the basic trie structure used for indexing long strings, and ensures that the number of nodes in the index is equal to the number of characters in the underlying data string. To the best of our knowledge, this is the first string index with these properties, and is in marked contrast to suffix-trees, the defacto standard string indexing structure. A rich set of forward and backward edges are employed in SPINE to ensure that all suffixes of the data string are captured in the index structure. Further, the false positives that inevitably resulted from the trie compaction were eliminated through a simple but powerful numeric labeling strategy that constrains when the index edges can be traversed. Finally, the SPINE index is prefix-partitionable, a property not shared by suffix-trees.

We provided detailed algorithms for both online construction of the SPINE index as well as for performing complex searching operations on the resulting indexes. A feature of the search algorithm is that it considerably reduces the number of suffixes that have to be examined during the alignment process. While a simplistic implementation of SPINE would have resulted in huge node sizes, we identified and incorporated a variety of structural optimizations that finally resulted in SPINE taking less than 12 bytes per indexed character, comparing favorably with the 17 bytes taken by standard suffix tree implementations.

A performance evaluation of SPINE against ST (SuffixTree) over a variety of very long genetic strings, including human chromosomes, showed that significant speedups were obtained for the searching operations, for both memory-resident and disk-resident scenarios. It was also observed that along with 30 percent lesser index size, SPINE exhibits much higher node locality than ST, resulting in a more efficient disk-based implementation. Finally, it was shown that a very simple buffering strategy was sufficient for SPINE to be able to take advantage of the locality observed in our experiments. In summary, SPINE appears to be a viable alternative to suffix-trees for string indexing.

Acknowledgements This work was supported in part by a Swarnajayanti Fellowship from the Dept. of Science & Technology, Govt. of India, and by a research grant from the Dept. of Bio-technoloy, Govt. of India.

References

[1] S. Altschul and W. Gish, "Basic Local Alignment Search Tool", *J. Mol. Biol.*, 215:403-410, 1990.

[2] A. Blumer, J. Blumer, D. Haussler, A. Ehrenfeucht, M. Chen and J. Seiferas, "The Smallest Automaton Recognizinng the Subwords of a Text", *Theoretical Computer Science*, 40:31-55, 1985.

[3] M. Crochmore and R. Verin, "Direct Construction of Compact Acyclic Word Graphs", *Proc. of Symp. on Combinatorial Pattern Matching*, 1997.

[4] A. Delcher, S. Kasif, R. Fleischmann, J. Peterson, O. White and S. Salzberg, "Alignment of whole genomes", *Nucleic Acids Research*, 27:2369-2376, 1999.

[5] R. Giegerich, S. Kurtz and J. Stoye, "Efficient Implementation of Lazy Suffix Trees", *Proc. of 3rd Workshop On Algorithm Engineering*, July 1999.

[6] D. Gusfield, "Algorithms on Strings, Trees, and Sequences", *Cambridge University Press*, 1997.

[7] E. Hunt, M. Atkinson and R. Irving, "A Database Index to Large Biological Sequences", *Proc. of the 27th VLDB Conference*, 2001.

[8] T. Kahveci and A. Singh, "An Efficient Index Structure for String Databases", *Proc. of the 27th VLDB Conference*, 2001.

[9] S. Kurtz, "Reducing the space requirements of suffix trees", *Software Practice and Experience*, 29:1149-1171, 1999.

[10] S. Inenaga, H. Hoshino, A. Shinohara, M. Takeda, S. Arikawa, G. Mauri and G. Pavesi, "On-Line Construction of Compact Directed Acyclic Word Graphs", *Proc. of Symp. on Combinatorial Pattern Matching*, 2001.

[11] U. Manber and G. Myers, "Suffix Arrays: A New Method for On-line String Searches", *SIAM J. Comput.*, 22(5), 1993.

[12] N. Neelapala, R. Mittal and J. Haritsa, "A Horizontally-Compacted Trie Index for Strings", *Tech. Rep. TR-2003-05*, DSL/SERC, Indian Institute of Science, 2003.

[13] *http://www.nist.gov/dads/HTML/trie.html*

[14] *http://www.tigr.org/software/mummer*