

Table 7.1. Simple illustration of the binary search (Section 7.2.2). This method reduces the search space by half for each iteration, checking whether a search word taken from I is in the remaining first or second half of a word list corresponding to J. Query sequence I is shown at the top. 4-words contained in sequence J are listed below in alphabetical order. Numbers indicate the position of each word sequence J.

I = GGAATAGCT	
J (4-word, position in sequence)	
AAAT 13 AATA 21 AATG 14 ACAA 19 ACTG 3 AGCC 9 AGCT 24 ATAG 22 ATGG 15 CAAA 12 CAAT 20 CCAA 11 CTAC 26	CTAG 7 CTGC 4 GACA 18 GCCA 10 GCTA 6, 25 GGAC 17 GTAC 1 TACA 27 TACT 2 TAGC 8, 23 TGGA 16 TGCT 5

finding the entry should have required ten steps (nine-letter word, $2^9 = 512$, and $2^{10} = 1024$ —nine “splits” are not enough since $864 > 512$).

7.2.3 Rare Words and Sequence Similarity

For the method described in Section 7.2.1, if k is large the table size can be enormous, and it will be mostly empty. For large k , another method for detecting sequence similarity is to put the k -words in an ordered list.

To find k -word matches between I and J, first break I down into a list of $n - k + 1$ k -words and J into a list of $m - k + 1$ k -words. Then put the words in each list in order, from AA...A to TT...T. This takes time $n \log(n)$ and $m \log(m)$ by standard methods which are routinely available but too advanced to present here. Let's index the list by $(W(i), Pw(i))$, $i = 1, \dots, n - k + 1$ and $(V(j), Pv(j))$, $j = 1, \dots, m - k + 1$, where, for example, $W(i)$ is the i th word in the ordered list and $Pw(i)$ is the position that word had in I.

We discover k -word matches by the following algorithm, which merges two ordered lists into one long ordered list. Start at the beginning of one list. Successively compare elements in that list with elements in the second list. If the element in the first list is smaller, include it in the merged list and continue. If not, switch to the other list. Proceed until reaching the end of