

Shortest Superstring via Set Cover

SHORTEST SUPERSTRING VIA SET COVER(S)

- 1 Remove from S any string that's a substring of another string of S .
- 2 Create a universe whose elements are the strings of S .
- 3 For each (i, j, k) such (last k characters of s_i)=(first k characters of s_j):
- 4 Create the set of substrings of $\sigma_{i,j,k}$ =(concatenation of s_i and s_j with overlap k).
- 5 The cost of that set associated is $|\sigma_{i,j,k}|$.
- 6 Solve the instance I of Set Cover thus defined.
- 7 Let s denote the concatenation of the $\sigma_{i,j,k}$'s associated to the sets in the set cover, in arbitrary order.
- 8 **return** s

Theorem 1 *The above is an $2H_n$ -approximation algorithm for Shortest Superstring.*

(Feasibility) By definition of set cover, the output string contains all strings of S so it's a feasible solution to the shortest superstring problem.

(Cost analysis) The length of the output is the sum of the lengths of the concatenated $\sigma_{i,j,k}$'s, which is exactly the cost of the corresponding set cover, so it's at most H_n times the value OPT_S of the optimum set cover of I :

$$\text{cost}(\text{Output}) \leq H_n \cdot \text{OPT}_S.$$

Let s^* denote the shortest superstring and OPT denote its length. We relabel the strings so that their order of first appearance in s^* is s_1, s_2, \dots, s_n . (Since no string of S is a substring of another string of S , the order in which the s_i 's start is also the order in which the s_i 's end, so this is well-defined.)

Let $x_1 = 1$, let y_1 be maximum such that s_{y_1} overlaps s_{x_1} in s^* , and let k_1 be their overlap. Assume x_{i-1} , y_{i-1} , and k_{i-1} have been defined. Let $x_i = y_{i-1} + 1$, let y_i be maximum such that s_{y_i} overlaps s_{x_i} in s^* , and let k_i be their overlap.

Since no string of S is a substring of another string of S , σ_{x_i, y_i, k_i} contains all strings s_j , $x_i \leq j \leq y_i$ as substrings, and so the sets associated to the σ_{x_i, y_i, k_i} 's form a feasible set cover \mathcal{A} . So:

$$\text{OPT}_S \leq \text{cost}(\mathcal{A}).$$

But by construction, σ_{x_i, y_i, k_i} is disjoint from $\sigma_{x_{i+2}, y_{i+2}, k_{i+2}}$ in s^* , and so the σ_{x_i, y_i, k_i} 's for i odd are all disjoint. So: $\sum_i \text{odd} |\sigma_{x_i, y_i, k_i}| \leq |s^*|$. Similarly $\sum_i \text{even} |\sigma_{x_i, y_i, k_i}| \leq |s^*|$. Summing, $\sum_i |\sigma_{x_i, y_i, k_i}| \leq 2|s^*|$. The left hand side is the cost of the set cover \mathcal{A} , so:

$$\text{cost}(\mathcal{A}) \leq 2 \cdot \text{OPT}.$$

Concatenating the three inequalities yields the Theorem.