The Shortest Superstring Problem: The Overlap Lemma

Last updated Nov 30, 2007

Overlap Lemma: Let c and c' be cycles in $\mathcal{C}$ and r, r' their respective representative strings. Then

$$overlap(r, r') < wt(c) + wt(c')$$

where $wt(c)$ is the cost of cycle c.

Cycle Lemma:

If every string in $S' \subseteq S$ is a substring of t^∞ for some string t

$\Rightarrow$there is a cycle of weight at most $|t|$ in the prefix graph covering all of the vertices (corresponding to strings) in S.

Proof:

Let $t = t_1 t_2 \ldots t_k$.

Suppose $S' = \{s_1, \ldots, s_m\}$.

Now let j_i be the starting point of the first occurrence of s_i in t^∞. This must be in the first copy of t (why).

Note that all of the j_i are different from each other since no substring in S is a substring of any other.

Now sort the strings by their starting points and consider the cycle c that visits the vertices corresponding to the strings in the sorted order. This cycle has length at most t so we are done.

<u>GCD Lemma:</u> Let X be a prefix of both α^∞ and β^∞ with $|X| \geq |\alpha| + |\beta|$. Then

1. If $|\alpha| = |\beta|$ then $\alpha = \beta$.
2. If $|\alpha| > |\beta|$ then X is a prefix of γ^∞ where $\gamma = X[1]X[2]\ldots X[|\alpha| - |\beta|]$.

<u>Proof:</u> (i) is obvious. To prove (ii) set $p = |\alpha|$, $q = |\beta|$. By definition, $\forall$, $0 < i \leq q$ **and** $0 < j \leq p$,

$$X[i+p] = X[i] \qquad \textbf{and} \qquad X[j+q] = X[j]$$

We now show that $\forall i$, $0 < i \leq |X| - (p-q)$,

$$X[i + (p-q)] = X[i].$$

First assume that $0 < i \leq q$. Then

$$\begin{aligned} X[i + (p-q)] &= X[i + (p-q) + q] \\ &= X[i+p] = X[i] \end{aligned}$$

Now assume that $q < i \leq |X| - (p-q)$. Then

$$\begin{aligned} X[i + (p-q)] &= X[i + (p-q) - p] \\ &= X[i-q] = X[i] \end{aligned}$$

<u>Corollary:</u> Let X be a prefix of both α^∞ and β^∞ with $|X| \geq |\alpha| + |\beta|$. Then X is a prefix of γ^∞ where $\gamma = X[1]X[2]\ldots X[gcd(|\alpha|, |\beta)|]$. Thus

$$\gamma^\infty = \alpha^\infty = \beta^\infty.$$

Overlap Lemma: Let c and c' be cycles in $\mathcal{C}$ and r, r' their respective representative strings. Then

$$overlap(r, r') < wt(c) + wt(c')$$

where $wt(c)$ is the cost of cycle c.

Proof: Assume the contrary, that

$$overlap(r, r') \geq wt(c) + wt(c')$$

Let α be the prefix of length of $wt(c)$ of $overlap(r, r')$ and α' the prefix of length of $wt(c')$ of $overlap(r, r')$. Notice that

1. **Every string "in" c is a substring of α^∞.**

2. **Every string "in" c' is a substring of $(\alpha')^\infty$.**

3. ***overlap*(r, r') is a prefix of both α^∞ and $(\alpha')^\infty$.**

From the GCD Lemma we know that the string γ containing the first $gcd(wt(c), wt(c'))$ characters of $overlap(r, r')$ satisfies

$$\gamma^\infty = \alpha^\infty = (\alpha')^\infty.$$

We just saw that

$$\gamma^\infty = \alpha^\infty = (\alpha')^\infty$$

so γ^∞ contains every string in c and every string in c'. Furthermore, by construction,

$$|\gamma| = gcd(wt(c), wt(c'))$$

so, from the Cycle Lemma ,we therefore have that there is a cycle of weight at most $gcd(wt(c), wt(c'))$ covering all strings in c and c'.
This contradicts the minimality of $\mathcal{C}$. Thus

$$overlap(r, r') < wt(c) + wt(c')$$