MIM2 - Algorithmes d'approximation- TD 3

Emmanuelle Lebhar
elebhar@ens-lyon.fr - Bureau 317

Surfacteur minimum
6 Octobre 2003

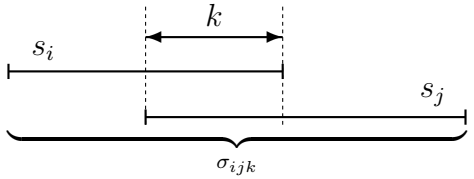
Exercice 1 (Dominant)

On appelle ensemble *dominant*, un ensemble de sommets D d'un graphe $G = (V, E)$ tel que pour tout $v \in V \setminus D$, il existe $d \in D$ avec $(v, d) \in E$.
Le problème de trouver un dominant est-il le même que de trouver une couverture par sommets ? Sinon en quoi sont-ils différents ?

Exercice 2 (Application à la biologie : le surfacteur minimum)

On peut voir l'ADN humain comme un mot très long, sur un alphabet à quatre lettres (A,C,G,T), que les scientifiques essaient de déchiffrer. Comme ce mot est très long, ils n'ont tout d'abord déchiffré que de courts morceaux de ce mot, qui se chevauchent les uns les autres. Bien entendu, les positions des morceaux dans le mot d'ADN original sont inconnues. Une hypothèse est que le mot le plus court qui admet ces morceaux pour *facteur* est une bonne approximation de l'ADN original.
Un mot v est facteur de u si il existe x et y tels que $u = xvy$, on dit alors que u est *surfacteur* de v.
Le problème du surfacteur minimum est la donnée d'un ensemble $U = \{s_1, \ldots, s_n\}$ de n mots sur un alphabet fini Σ, il s'agit de trouver un mot minimum s qui contienne une occurence de chaque s_i.
Sans perte de généralité, nous pouvons supposer qu'aucun des mots s_i n'est surfactceur d'un autre mot s_j de l'ensemble pour $i \neq j$.

Approximation par couverture par ensembles. On considère l'instance suivante de la couverture par ensembles, que nous noterons $\mathcal{C}$. Pour chaque $s_i, s_j \in U$ et $k > 0$, tels que le suffixe de longueur k de s_i est un préfixe de s_j, on note σ_{ijk} le mot obtenu en chevauchant les suffixe et préfixe de longueur k de s_i et s_j respectivement :



Notons M l'ensemble des mots σ_{ijk}, pour tous les choix valides de i, j, k. Pour chaque mot $\pi \in \Sigma^+$, notons $\text{set}(\pi) = \{s \in U \mid s \text{ est un facteur de } \pi\}$. U est l'ensemble univers de l'instance $\mathcal{C}$ de la couverture par ensembles, et $\{\text{set}(\pi) \mid \pi \in U \cup M\}$ est la collection de sous-ensembles associée. Le coût de chaque $\text{set}(\pi)$ est la longueur de π, $|\pi|$.

1. Notons respectivement $\text{OPT}_{\mathcal{C}}$ et OPT le coût d'une solution optimale de $\mathcal{C}$ et la longueur d'un plus court surfacteur de U.
   Montrez que $\text{OPT} \leq \text{OPT}_{\mathcal{C}}$.

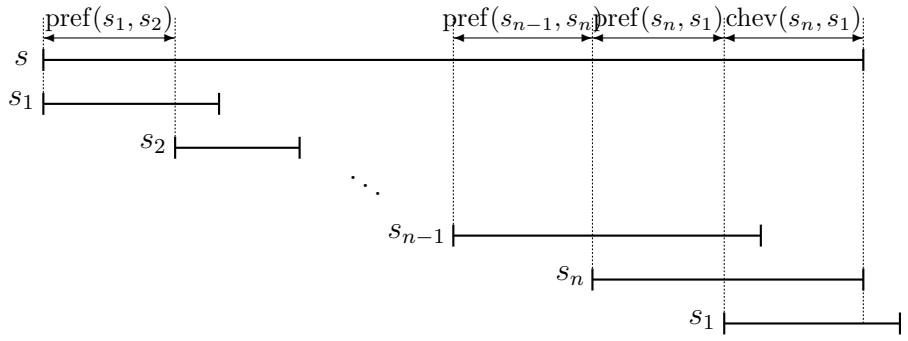
   Algorithme : surfacteur via set cover
   - Utiliser l'algorithme glouton de couverture par ensembles sur l'instance $\mathcal{C}$.
     Soient $\text{set}(\pi_1)$, ..., $\text{set}(\pi_k)$ les ensembles sélectionnés dans cette couverture.
   - Concaténer les mots $\pi_1, \ldots, \pi_k$, dans n'importe quel ordre.
   - Retourner le mot $s = \pi_1\pi_2 \ldots \pi_k$ obtenu.

2. Montrez que $\text{OPT}_{\mathcal{C}} \leq 2 \cdot \text{OPT}$ en exhibant une couverture par ensemble de coût inférieur à $2 \cdot \text{OPT}$. *Indication : on pourra trier les mots $s_1, \ldots, s_n$ dans l'ordre de leur première occurence dans s^*, surfacteur optimal, puis les partitionner en t ensembles $\pi_1, \pi_2, \ldots, \pi_t \in U \cup M$.*
3. Concluez sur le facteur d'approximation de l'algorithme pour le surfacteur.

Approximation directe. On note $\text{pref}(s_i, s_j)$ le préfixe de s_i obtenu en enlevant son chevauchement avec s_j, noté $\text{chev}(s_i, s_j)$.

1. Soit s^* un surfacteur minimum optimal. Par symétrie, on peut renuméroter les mots de U dans leur ordre d'apparition dans s^*. Exprimez OPT en termes de préfixes.

   On considère le graphe orienté complet de sommets $\{1, \ldots, n\}$, et les arêtes $(i \to j)$ ont pour poids $|\text{prefixe}(s_i, s_j)|$. On appelle ce graphe, *graphe des préfixes*.
   À quoi correspond le cycle $1 \to 2 \to \ldots n$ pour ce graphe ? En déduire une première minoration de OPT.



2. Couverture par cycles.
   Le problème de la couverture par cycles de poids minimum est de trouver un ensemble

de cycles de poids minimum dans un graphe tel que tout sommet est couvert. Ce problème est polynomial (cf *question bonus*).
Donnez un second minorant de OPT.

3. Pour tout cycle $c = (i_1 \to i_2 \to \dots i_l \to i_1)$ du graphe des préfixes, on pose

$$\alpha(c) = \text{pref}(s_{i_1}, s_{i_2}).\dots.\text{pref}(s_{i_{l-1}}, s_{i_l}).\text{pref}(s_{i_l}, s_{i_1})$$

On pose $\sigma(c) = \alpha(c).s_{i_1}$ et on appelle s_{i_1} le *mot représentatif* de c.
Que peut-on dire des s_{i_j} par rapport à $\sigma(c)$ et $\alpha(c)$?

4. Nous allons montrer que l'algorithme qui calcule $\mathcal{C} = \{c_1, \dots c_k\}$, une couverture de poids minimum du graphe des préfixes, et retourne le mot $\sigma(c_1).\dots.\sigma(c_k)$ est une 4-approximation du problème du surfacteur minimum.
Si chaque mot représentatif était de longueur inférieure au poids du cycle, que pourrait-on dire du facteur d'approximation ?

5. Soit $U' \subseteq U$, supposons qu'il existe un mot t tel que tout mot de U' soit un facteur de t^∞. Montrez qu'il existe un cycle de poids inférieur à $|t|$, dans le graphe des préfixes, qui couvre exactement les sommets correspondants aux mots de U'.

6. Soient c et c' deux cycles de la couverture par cycles $\mathcal{C}$ et r et r' des mots représentatifs de ces cycles, montrer que $|\text{chevauch}(r, r')| < \text{poids}(c) + \text{poids}(c')$, où $\text{chevauch}(r, r')$ est le chevauchement des deux mots r, r'. *On pourra raisonnner par l'absurde et démontrer que sinon, tous les mots couverts par c et c' seraient facteurs de $\alpha(c)^\infty$. Appliquer la question 5 conduirait alors à une absurdité.*
On pourra étudier α et α', les préfixes de longueurs respectives $\text{poids}(c)$ *et* $\text{poids}(c')$ *du chevauchement.*

7. En notant r_i le mot représentatif du cycle c_i, montrez que :

$$\sum_{i=1}^{k} |\sigma(c_i)| = \text{poids}(\mathcal{C}) + \sum_{i=1}^{k} |r_i|$$

Concluez.

8. *Question bonus :* Montrez que le problème de la couverture par cycles de poids minimum est polynomial.