

Predicting Protein-Protein Interactions from Protein Domains Using a Set Cover Approach

Theorem 4.1 Relationship between the approximation by *GREEDY_MSSC* and an optimal solution.

Proof. If $m(\mathcal{C}_o) = 0$, it means that all elements in \mathcal{C}_o are subsets of X . *GREEDY_MSSC* cannot choose a set that is not completely in X , because at any given time when $U \neq \emptyset$ there exists a set $S \in \mathcal{C}_o \cap \mathcal{E}$ such that $\frac{|S-X|}{|S \cap U|} = 0$, i.e., $S \subseteq X$. Hence $m(\mathcal{C}_o) = 0$. We assume $m(\mathcal{C}_o) \neq 0$ from this point. Without loss of generality, we can also assume that $|\mathcal{C}_o| \leq |X|$. If $|\mathcal{C}_o| > |X|$, it means that at least one element S in \mathcal{C}_o is redundant to cover X , so we can remove S from \mathcal{C}_o , and the remaining set is still an optimal solution.

$|\mathcal{C}_o| \leq |X|$ implies that $|m(\mathcal{C}_o)| \leq k|X|$. Suppose

$$m(\mathcal{C}_o) = a|X|, \text{ for some value } a, 0 < a \leq k.$$

At a given time, assume that the minimum $\frac{|S-X|}{|S \cap U|}$ is r , where U is defined as in *GREEDY_MSSC*. For any $Z \in \mathcal{C}_o \cap \mathcal{E}$,

$$\frac{|Z-X|}{|Z \cap U|} \geq r,$$

so

$$\frac{|Z \cap U|}{|Z-X|} \leq \frac{1}{r}. \tag{1}$$

We have

$$\begin{aligned} |U| &= \left| \bigcup_{Z \in \mathcal{C}_o \cap \mathcal{E}} Z \cap U \right| \\ &\leq \sum_{Z \in \mathcal{C}_o \cap \mathcal{E}} \frac{|Z-X|}{r}, \text{ by Equation (1)} \\ &\leq \frac{m(\mathcal{C}_o)}{r} \\ &= \frac{a|X|}{r}. \end{aligned}$$

Therefore, there are $|X| - |U| \geq (1 - \frac{a}{r})|X|$ points of X that are already covered when S is the next set to be chosen, i.e., *GREEDY_MSSC* cannot choose a set S with

$$\frac{|S-X|}{|S \cap U|} \geq r$$

until a fraction $(1 - \frac{a}{r})$ of X has been covered. Conversely, if $x = \frac{|X-U|}{|X|} = 1 - \frac{a}{r}$ (the covered part of X), then $r \leq \frac{a}{1-x}$, and for the set S chosen by *GREEDY_MSSC*,

$$f(x) := \frac{|S-X|}{|S \cap U|} = \frac{a}{1-x}.$$

x is increasing from 0 to 1. Every time a new subset S is chosen, x “jumps” to a new value, so $f(x)$ is a step function of x . Since $|S \cap U| \geq 1$ and $|S - X| \leq k - 1$, $f(x) \leq k - 1$. Note that $\frac{a}{1-x} = k - 1$ if and only if $x = 1 - \frac{a}{k-1}$.

When *GREEDY_MSSC* chooses a set S , S covers $|S \cap U| = |X|\Delta x$ more points of X , where $\Delta x = \frac{|S \cap U|}{|X|}$. The contribution of S to $m(\mathcal{C}_a)$ is

$$|S - X| = f(x)|S \cap U| = f(x)|X|\Delta x.$$

Therefore,

$$\begin{aligned} m(\mathcal{C}_a) &= \sum_{S \in \mathcal{C}_a} |S - X| \\ &= \sum_{S \in \mathcal{C}_a} f(x)|X|\Delta x \\ &= |X| \int_0^1 f(x)dx, \quad f(x) \text{ is a step function} \\ &\leq |X| \left[\int_0^{1 - \frac{a}{k-1}} \frac{a}{1-x} dx + \int_{1 - \frac{a}{k-1}}^1 (k-1)dx \right] \\ &= |X|(a \ln(k-1) - a \ln a + a) \\ &\leq a|X|[\ln(k-1) + 1] \\ &= m(\mathcal{C}_o)[\ln(k-1) + 1]. \end{aligned}$$

□