Characteristic Sets for Unions of Regular Pattern Languages and Compactness

Masako Sato, Yasuhito Mukouchi and Dao Zheng

Department of Mathematics and Information Sciences
College of Integrated Arts and Sciences
Osaka Prefecture University, Sakai, Osaka 599-8531, Japan
e-mail: {sato, mukouchi}@mi.cias.osakafu-u.ac.jp

Abstract. The paper deals with the class $\mathcal{RP}^k$ of sets of at most k regular patterns. A semantics of a set P of regular patterns is a union $L(P)$ of languages defined by patterns in P. A set Q of regular patterns is said to be a more general than P, denoted by $P \sqsubseteq Q$, if for any $p \in P$, there is a more general pattern q in Q than p. It is known that the syntactic containment $P \sqsubseteq Q$ for sets of regular patterns is efficiently computable. We prove that for any sets P and Q in $\mathcal{RP}^k$, (i) $S_2(P) \subseteq L(Q)$, (ii) the syntactic containment $P \sqsubseteq Q$ and (iii) the semantic containment $L(P) \subseteq L(Q)$ are equivalent mutually, provided $\sharp\Sigma \geq 2k-1$, where $S_n(P)$ is the set of strings obtained from P by substituting strings with length at most n for each variable. The result means that $S_2(P)$ is a characteristic set of $L(P)$ within the language class for $\mathcal{RP}^k$ under the condition above. Arimura et al. showed that the class $\mathcal{RP}^k$ has compactness with respect to containment, if $\sharp\Sigma \geq 2k+1$. By the equivalency above, we prove that $\mathcal{RP}^k$ has compactness if and only if $\sharp\Sigma \geq 2k-1$.

The results obtained enable us to design efficient learning algorithms of unions of regular pattern languages such as already presented by Arimura et al. under the assumption of compactness.

1 Introduction

A pattern is a string consisting of constant symbols in a fixed alphabet Σ and variables. For example, $p = axbx$ is a pattern, where a and b are constant symbols, and x is a variable. The language $L(p)$ defined by a pattern p is the set of constant strings obtained from the pattern by substituting nonempty constant strings for variables in p. For example, the language defined by the above pattern is $L(p) = \{awbw \mid w \in \Sigma^+\}$.

The class $\mathcal{PL}$ of pattern languages was introduced by Angluin[1] as a class inductively inferable from positive data based on identification in the limit due to Gold[7]. The class $\mathcal{PL}$ is one of the most basic class in the framework of elementary formal systems which was introduced by Smullyan[13] to develop a new theory of recursive functions, and was proposed as a unifying framework for language learning by Arikawa et al.[3]. That is, an elementary formal system

consisting of only one definite clause defines a pattern language. In some practical applications such as genome informatics, pattern languages are paid much attentions (cf. Arikawa et al.[4]).

Angluin[2] showed that the class $\mathcal{PL}$ has a property of so-called finite thickness. Wright[14] introduced a notion of finite elasticity for a language class, which is a natural extension of that of finite thickness, and showed that a class with finite elasticity is inferable from positive data and moreover, the property is closed under union operation. As a result, it was shown that for any fixed k, the class $\mathcal{PL}^k$ of unions of at most k pattern languages is also inferable from positive data. On the other hand, Sato[11] introduced a notion of finite cross property characterizing a class with finite elasticity. The property of finite cross property is closely related with a characteristic set. A nonempty finite set S of strings is said to be a characteristic set of a language L within a class $\mathcal{L}$, when L is the least language within $\mathcal{L}$ containing the set S. We show that a language L has a finite cross property within $\mathcal{L}$ if and only if there is a characteristic set of L within the class $\mathcal{L}$. Thus if a class $\mathcal{L}$ has a finite elasticity, then for any language $L \in \mathcal{L}$, there is a characteristic set of L within $\mathcal{L}$.

Let $\mathcal{D}$ be a set of descriptions which can be partially ordered by an effectively computable relation $\sqsubseteq$, and let $\mathcal{L} = \{L(P) \mid P \in \mathcal{D}\}$ be the language class defined by descriptions of $\mathcal{D}$. We assume that the syntactic containment $P \sqsubseteq Q$ implies the semantic containment $L(P) \subseteq L(Q)$ for any $P, Q \in \mathcal{D}$. Assume that the class $\mathcal{L}$ has finite elasticity. Thus for any description $P \in \mathcal{D}$, there is a characteristic set of $L(P)$ within $\mathcal{L}$. If the problem of finding one of the characteristic sets and the membership problem for languages in $\mathcal{L}$ are efficiently computable, then the containment for languages of $\mathcal{L}$ is also efficiently computable. Furthermore, if the semantic containment $L(P) \subseteq L(Q)$ implies the syntactic containment $P \sqsubseteq Q$, the containment for languages of $\mathcal{L}$ is efficiently computable.

A pattern is said to be regular, if each variable in the pattern appears at most once. In this paper, we deal with the class $\mathcal{RP}^k$ of sets of at most k regular patterns as a class of descriptions, and develop the above discussion for $\mathcal{RP}^k$. A pattern q is a generalization of a pattern p, denoted by $p \preceq q$, when q is obtained from p by substituting patterns for variables in p. For example, a pattern $q = axy$ is a generalization of a pattern $p = axbx$, i.e., $axbx \preceq axy$. The set $\mathcal{P}$ of patterns is a partially ordered set under the relation $\preceq$, provided we identify patterns obtained by renaming variables. Clearly the syntactic containment $p \preceq q$ implies the semantic containment $L(p) \subseteq L(q)$, but not always the converse. Mukouchi showed that the converse is valid for the class $\mathcal{RP}$ of regular patterns.

A set $P = \{p_1, \cdots, p_n\}$ of regular patterns defines a language $L(P) = L(p_1) \cup \cdots \cup L(p_n)$. Let $\mathcal{RPL}^k$ be the class defined by the description class $\mathcal{RP}^k$. For sets $P, Q \in \mathcal{RP}^k$, we define a relation $\sqsubseteq$ as follows: $P \sqsubseteq Q$ if and only if for any $p \in P$, there is a regular pattern $q \in Q$ such that $p \preceq q$. Clearly $P \sqsubseteq Q$ implies $L(P) \subseteq L(Q)$. The relation $\sqsubseteq$ is an efficiently computable and partially ordered relation by restricting to sets in $\mathcal{RP}^k$ of canonical form (cf. Arimura et al.[5]).

The class $\mathcal{RPL}^k$ as well as the class $\mathcal{PL}$ have finite elasticity. Thus for each $P \in \mathcal{RP}^k$, there is a characteristic set of $L(P)$ within $\mathcal{RPL}^k$. Let $S_n(P) =$

$\bigcup_{p\in P} S_n(p)$, where $S_n(p)$ is the set of strings obtained from p substituting nonempty constant strings with length at most n for each variable. Then there is a positive number n such that $S_n(P)$ is a characteristic set of $L(P)$ within $\mathcal{RPL}^k$. We are interested in the positive number n for given $P \in \mathcal{RP}^k$. We first prove that (i) $S_1(P) \subseteq L(Q)$, (ii) $P \sqsubseteq Q$ and (iii) $L(P) \subseteq L(Q)$ are equivalent mutually, provided $\sharp\Sigma \geq 2k+1$. The result is not always valid, if $\sharp\Sigma \leq 2k$. We show, however, that the above equivalency for (i') $S_2(P) \subseteq L(Q)$ instead of (i) is valid, provided $\sharp\Sigma \geq 2k-1$. Thus $S_2(P)$ is a characteristic set of $L(P)$ within $\mathcal{RPL}^k$. It is known that the membership problem for regular pattern languages is polynomial time computable (cf. Shinohara[12]), although it is NP-complete for general pattern languages (cf. Angluin[1]). Thus the containment for languages of $\mathcal{RPL}^k$ is efficiently computable.

On the other hand, Arimura et al.[5] gave an efficient algorithm of languages in $\mathcal{RPL}^k$ under the condition that the class has compactness with respect to containment. The class $\mathcal{RP}^k$ has compactness with respect to containment, if $L(P) \subseteq L(Q)$ implies $P \sqsubseteq Q$ for $P, Q \in \mathcal{RP}^k$. Arimura and Shinohara[6] showed the compactness of $\mathcal{RP}^k$, if $\sharp\Sigma \geq 2k+1$. In terms of the above equivalency, it can be shown that $\mathcal{RP}^k$ has compactness w.r.t. containment, if $\sharp\Sigma \geq 2k-1$. Moreover, a counter-example is given so that $\mathcal{RP}^k$ does not have compactness w.r.t. containment, if $\sharp\Sigma < 2k-1$. Consequently, the containment for languages of $\mathcal{RPL}^k$ reduces to that for $\mathcal{RP}^k$, and thus it is efficiently computable.

2 Regular Pattern Languages

Let Σ be a finite set of *constant* symbols containing at least two symbols, and $X = \{x, y, x_1, x_2, \cdots\}$ be a countable set of *variable* symbols. We assume $\Sigma \cap X = \phi$.

A *pattern* is a string in $(\Sigma \cup X)^*$. Note that we consider the empty string ε as a pattern, for convenience. By $\mathcal{P}$ we denote the set of all patterns. The *length* of a pattern p, denoted by $|p|$, is just the number of symbols composing it. A *substitution* θ is a homomorphism from patterns to patterns that maps every constant to itself. For a pattern p and a substitution θ, we denote by $p\theta$ the image of p by θ. A pattern q is a *generalization* of a pattern p, or p is an *instance* of q, denoted by $p \preceq q$, if there is a substitution θ such that $p = q\theta$. For two patterns p and q, if $p \preceq q$ and $q \preceq p$, then p equals q, denoted by $p \equiv q$, except for labeling variables in them. The set $(\mathcal{P}, \preceq)$ constitutes a partial ordering set with respect to $\equiv$.

The language defined by a pattern p is the set $L(p) = \{w \in \Sigma^* \mid w \preceq p\}$. Clearly if $p \equiv q$, then $L(p) = L(q)$. A language L over Σ is a *pattern language*, if $L = L(p)$ for some pattern p. We denote by $\mathcal{PL}$ the class of all pattern languages.

In this paper, we are especially concerned with a subclass of $\mathcal{P}$. A pattern p is *regular*, if each variable appears at most once in p. A *regular pattern language* is a pattern language defined by a regular pattern. We denote by $\mathcal{RP}$ the set of all regular patterns, and by $\mathcal{RPL}$ the set of all regular pattern languages.

Concerning regular patterns, the next fundamental result has been shown:

Lemma 1 (Mukouchi[10]). *Let p and q be regular patterns. Then $p \preceq q$ if and only if $L(p) \subseteq L(q)$.*

Note that "if" part of the lemma above is not always valid for general patterns, although "only if" part is always valid.

By the result above, the containment problem for regular pattern languages is reduced to the decision problem of partial ordering for regular patterns, which is polynomial time computable (cf. Shinohara[12]).

Now we consider unions of languages defined by patterns. By $\mathcal{P}^+$ we denote the class of all nonempty finite subsets of $\mathcal{P}$. For $k \geq 1$, let $\mathcal{P}^k$ be the class of sets consisting of at most k patterns. By $\mathcal{PL}^k$ we denote the class of unions of at most k pattern languages, that is,

$$\mathcal{PL}^k = \{L(P) \mid P \in \mathcal{P}^k\},$$

where $L(P) = \bigcup_{p\in P} L(p)$. In a similar way, we also define $\mathcal{RP}^+$, $\mathcal{RP}^k$ and $\mathcal{RPL}^k$, respectively.

For $P, Q \in \mathcal{P}^+$, we define the binary relation $P \sqsubseteq Q$ as follows: $P \sqsubseteq Q$ if and only if for any $p \in P$, there is $q \in Q$ such that $p \preceq q$. It is easy to see that $P \sqsubseteq Q$ implies $L(P) \subseteq L(Q)$. However the converse is not valid in general.

Definition 2. A class $\mathcal{C} \subseteq \mathcal{P}^+$ has *compactness with respect to containment*, if for any pattern $p \in \mathcal{P}$ and any set $Q \in \mathcal{C}$, $L(p) \subseteq L(Q)$ implies $L(p) \subseteq L(q)$ for some $q \in Q$.

In a similar way, we also define compactness for a class $\mathcal{C} \subseteq \mathcal{RP}^+$.

For a class $\mathcal{C} \subseteq \mathcal{RP}^+$ with compactness, it is easy to see by Lemma 1 that for any $P, Q \in \mathcal{C}$, $P \sqsubseteq Q$ if and only if $L(P) \subseteq L(Q)$.

In this paper, we show the compactness of the class $\mathcal{RP}^k$ as a corollary of stronger property than the compactness as follows: For some particular finite subset S of $L(p)$, $S \subseteq L(Q)$ implies $L(p) \subseteq L(q)$ for some $q \in Q$. Note that $S \subseteq L(Q)$ implies also $L(p) \subseteq L(Q)$. Such a set S is called a characteristic set for $L(p)$, which is defined as follows:

Definition 3. Let $\mathcal{L}$ be a class of languages and L be a language. A set $S \subseteq \Sigma^+$ is a *characteristic set* for L within $\mathcal{L}$, if S is a finite subset of L and for any $L' \in \mathcal{L}$, $S \subseteq L'$ implies $L \subseteq L'$.

If S is a characteristic set for $L \in \mathcal{L}$, L is the least language among $\mathcal{L}$ containing S in the set-containment ordering, and any finite superset of S contained in L is also a characteristic set for L. Furthermore a finite language $L \in \mathcal{L}$ is a characteristic set for itself.

The notion of a characteristic set has very closed relation with that of finite elasticity due to Wright[14] as well as that of finite cross property due to Sato[11] defined as follows:

Definition 4 (Wright[14] and Motoki et al.[9]). A class $\mathcal{L}$ of languages has *finite elasticity*, if there does not exist an infinite sequence $(w_i)_{i\geq 0}$ of strings and an infinite sequence $(L_i)_{i\geq 1}$ of languages in $\mathcal{L}$ such that for any $i \geq 1$,

$$\{w_0, \cdots, w_{i-1}\} \subseteq L_i, \quad \text{but} \quad w_i \notin L_i.$$

A condition for a class to have finite elasticity is characterized by the notion of finite cross property of a language as follows:

Definition 5 (Sato[11]). Let $\mathcal{L}$ be a class of languages. A language L has *finite cross property within* $\mathcal{L}$, if there does not exist an infinite sequence $(T_n)_{n\geq 1}$ of finite sets of strings and an infinite sequence $(L_i)_{i\geq 1}$ of languages in $\mathcal{L}$ such that (i) $T_1 \subsetneq T_2 \subsetneq \cdots$, (ii) $\bigcup_{i=1}^{\infty} T_i = L$, (iii) $T_i \subseteq L_i$, but $T_{i+1} \not\subseteq L_i$ $(i \geq 1)$.

Lemma 6 (Sato[11]). *A class $\mathcal{L}$ of languages has finite elasticity if and only if every language L has finite cross property within $\mathcal{L}$.*

Furthermore, by their definitions, it is easy to see that the following lemma is valid:

Lemma 7. *Let $\mathcal{L}$ be a class of languages and L be a language. Then L has finite cross property within $\mathcal{L}$ if and only if there is a characteristic set for L within $\mathcal{L}$.*

By Lemmas 6 and 7, we see that a class $\mathcal{L}$ has finite elasticity if and only if for any language L, there is a characteristic set for L within $\mathcal{L}$. Note that this result has already shown in Kobayashi and Yokomori[8].

Wright[14] showed that the class $\mathcal{PL}^k$ has finite elasticity, and so is the subclass $\mathcal{RPL}^k$. Thus, by the lemmas above, we see that for any language $L \in \mathcal{RPL}^k$, there is a characteristic set for L within $\mathcal{RPL}^k$.

Now we define a particular finite subset of a regular pattern language which plays an important role in our paper. For a regular pattern p with just m variables $x_1, \cdots, x_m$ and for $n \geq 1$, we define a finite subset $S_n(p)$ of $L(p)$ as follows: Let $S_n(p)$ be the set of all strings obtained from p by substituting strings in Σ^+ with length at most n for each variable in p.

For a nonempty finite set P of regular patterns, we define

$$S_n(P) = \bigcup_{p\in P} S_n(p).$$

Clearly $S_n(P) \subseteq S_{n+1}(P) \subseteq L(P)$ for any $n \geq 1$.

Since a characteristic set for $L(P)$ is a finite set, we have the following theorem:

Theorem 8. *For any $P \in \mathcal{RP}^k$, there is $n \geq 1$ such that $S_n(P)$ is a characteristic set for $L(P)$ within $\mathcal{RPL}^k$.*

In Section 4, we will show that 2 is sufficient for the number n in the theorem above, under the assumption that the number of constants is not less than $2k-1$.

3 $S_1(P)$ as a Characteristic Set

In this section, we will give some simple characteristic set for each language in $\mathcal{RPL}^k$. The key is the set $S_1(p)$ of strings with the shortest length for a regular pattern p.

Let $p_1rp_2 \preceq q$ for regular patterns p_1, r, p_2 and q, and let $x_1, \cdots, x_n$ be variables appearing in q. The subpattern r in p_1rp_2 is *generated from q by variable substitution*, if there exists a variable x_i in q and a substitution $\theta = \{x_1 := r_1, \cdots, x_i := r'rr'', \cdots, x_n := r_n\}$ such that $p_1 = (q_1\theta)r'$, $p_2 = r''(q_2\theta)$ for $q = q_1x_iq_2$. Note that if the pattern r in p_1rp_2 is generated from q by variable substitution, clearly $p_1xp_2 \preceq q$ holds. In particular, if $p_1ap_2 \preceq q$ but $p_1xp_2 \not\preceq q$ for some $a \in \Sigma$, then the constant a in p_1ap_2 is not generated from q by variable substitution, and moreover $q = q_1aq_2$ holds for some q_1 and q_2 such that $p_j \preceq q_j$ $(j = 1, 2)$. Furthermore, if $p_1xp_2 \preceq q$, then the variable x in p_1xp_2 is always generated from q by variable substitution.

For a pattern p, by head(p) and tail(p) we denote the first symbol and the last symbol of p, respectively.

We first give two fundamental lemmas useful in this paper.

Lemma 9. *Let $p = p_1xp_2$ and $q = q_1q_2q_3$, where p, q, p_1, p_2, q_1, q_2 and q_3 are regular patterns and x is a variable. Then if $p_1 \preceq q_1q_2$, $p_2 \preceq q_2q_3$ and q_2 contains some variables, then $p \preceq q$ holds.*

Proof. Let y be any fixed variable appearing in q_2 and $q_2 = q_2'yq_2''$ for some q_2' and q_2''. By $p_1 \preceq q_1(q_2'yq_2'')$, we can put $p_1 = p_1'p_1''$ for some p_1' and p_1'' such that $p_1' \preceq q_1q_2'$ and $p_1'' \preceq yq_2''$. Similarly, by $p_2 \preceq (q_2'yq_2'')q_3$, we can put $p_2 = p_2'p_2''$ for some p_2' and p_2'' such that $p_2' \preceq q_2'y$ and $p_2'' \preceq q_2''q_3$. Now we consider a substitution $\theta = \{y := p_1''xp_2'\}$. Then we have $p = p_1xp_2 = p_1'(p_1''xp_2')p_2'' \preceq q_1q_2'(p_1''xp_2')q_2''q_3 = q\theta \preceq q$. □

By the result above, if $p_1 \preceq q_1q_2$ and $p_2 \preceq q_2q_3$ but $p \not\preceq q$, then q_2 contains no variable, i.e., $q_2 \in \Sigma^*$.

Lemma 10. *Suppose $\sharp\Sigma \geq 3$. Let p and q be regular patterns. Then if $p\{x := a\} \preceq q$, $p\{x := b\} \preceq q$ and $p\{x := c\} \preceq q$ for distinct constants a, b and c, then $p \preceq q$ holds.*

Proof. If p does not contain the variable x, then it is clearly true. Thus we consider $p = p_1xp_2$ for some regular patterns p_1 and p_2.

Suppose $p \not\preceq q$. As mentioned above, if the constant a in p_1ap_2 is generated by variable substitution from q, then by $p_1ap_2 \preceq q$, $p \preceq q$ holds, which is a contradiction. Thus we can put $q = q_a^{(1)}aq_a^{(2)}$ for some $q_a^{(1)}$ and $q_a^{(2)}$ such that $p_j \preceq q_a^{(j)}$ $(j = 1, 2)$. Similarly, from $p_1bp_2 \preceq q$ and $p_1cp_2 \preceq q$, we can put $q = q_b^{(1)}bq_b^{(2)} = q_c^{(1)}cq_c^{(2)}$ for some $q_b^{(1)}, q_b^{(2)}, q_c^{(1)}$ and $q_c^{(2)}$ such that $p_j \preceq q_b^{(j)}$ and $p_j \preceq q_c^{(j)}$ $(j = 1, 2)$.

Without loss of generality, we can put $q = q_1aq_2bq_3cq_4$, where

$$\begin{array}{ll}(1)\ p_1 \preceq q_a^{(1)} = q_1, & (1')\ p_2 \preceq q_a^{(2)} = q_2bq_3cq_4,\\ (2)\ p_1 \preceq q_b^{(1)} = q_1aq_2, & (2')\ p_2 \preceq q_b^{(2)} = q_3cq_4,\\ (3)\ p_1 \preceq q_c^{(1)} = q_1aq_2bq_3, & (3')\ p_2 \preceq q_c^{(2)} = q_4.\end{array}$$

As easily seen, both q_2 and q_3 contain no variable. In fact, by (2) and (1'), $p_1 \preceq (q_1a)q_2$ and $p_2 \preceq q_2(bq_3cq_4)$. Thus if q_2 contains some variables, it implies by Lemma 9 that $p \preceq q$, which contradicts the assumption. Therefore q_2 contains no variable. Similarly we can show that q_3 contains no variable.

Put $w = q_2$ and $w' = q_3$. By (2) and (3), both aw and $awbw'$ are suffixes of p_1. Therefore if $|w| = |w'|$, then $aw = bw'$ holds, which contradicts the assumption that $a \neq b$.

Assume $|w| < |w'|$. Then aw is a suffix of w', so $w' = w_1aw$ for some $w_1 \in \Sigma^*$. Similarly, by (1') and (2'), both $wbw'c$ and $w'c$ are prefixes of p_2. Thus wb is a prefix of w', so $w' = wbw_2$ for some $w_2 \in \Sigma^*$, and thus $|w_1| = |w_2|$. This implies $a = c$, because both $wbw'c = wbw_1awc$ and $w'c = wbw_2c$ are prefixes of p_2, which contradicts the assumption that $a \neq c$.

We can also show a contradiction similarly for the case of $|w'| < |w|$. This completes our proof. □

Theorem 11. *Suppose $\sharp\Sigma \geq 2k+1$. Let $P \in \mathcal{RP}^+$ and $Q \in \mathcal{RP}^k$. Then the following three propositions are equivalent:*
(i) $S_1(P) \subseteq L(Q)$, (ii) $P \sqsubseteq Q$, (iii) $L(P) \subseteq L(Q)$.

Proof. Clearly (ii) implies (iii) and (iii) implies (i). Now we prove (i) implies (ii). It suffices to show that for any regular pattern p, $S_1(p) \subseteq L(Q)$ implies $p \preceq q$ for some $q \in Q$.

The proof is done by a mathematical induction on the number n of variables in p. In case $n = 0$, $p \in L(Q)$, and so $p \in L(q)$ for some $q \in Q$. Let $n \geq 0$ and assume that it is valid for any regular pattern with n variables. Let p be a regular pattern with $(n+1)$ variables such that $S_1(p) \subseteq L(Q)$, and let x be any fixed variable in p. Put $p_a = p\{x := a\}$ for each $a \in \Sigma$. Note that p_a has just n variables and $S_1(p_a) \subseteq L(Q)$ holds. Thus by the induction hypothesis, $p_a \preceq q_a$ for some $q_a \in Q$. Since $\sharp\Sigma \geq 2k+1$ and $\sharp Q \leq k$, there exists at least one regular pattern $q \in Q$ such that $p_{a_j} \preceq q$ for some distinct constants $a_j \in \Sigma$ $(j = 1,2,3)$. By Lemma 10, it implies $p \preceq q$. □

As a direct corollary of this theorem, we have:

Corollary 12. *Suppose $\sharp\Sigma \geq 3$. Let p and q be regular patterns. Then the following three propositions are equivalent:*
(i) $S_1(p) \subseteq L(q)$, (ii) $p \preceq q$, (iii) $L(p) \subseteq L(q)$.

Note that Theorem 11 is not valid in general if $\sharp\Sigma \leq 2k$. Before illustrating a counter-example, we give the following lemma:

Lemma 13. *Suppose $\sharp\Sigma \geq 3$. Let p and q be regular patterns. Then if $p\{x := a\} \preceq q$ and $p\{x := b\} \preceq q$ for distinct constants a and b but $p \npreceq q$, then there exist regular patterns p_1, p_2, q_1 and q_2 and a string $w \in \Sigma^*$ such that*

$$\begin{array}{lll} p = p_1AwxwBp_2, & q = q_1AwBq_2, & p_j \preceq q_j\ (j=1,2),\\ p_1Aw \preceq q_1, & wBp_2 \preceq q_2, & \end{array}$$

where $A = a$, $B = b$ or $A = b$, $B = a$.

Proof. Clearly p contains the variable x. Let $p = p_1'xp_2'$ for some regular patterns p_1' and p_2'. Similarly to the proof of Lemma 10, we can show that there exist regular patterns q_1 and q_2 and a string $w \in \Sigma^*$ such that $q = q_1AwBq_2$, $p_j' \preceq q_j$ $(j = 1,2)$, $p_1' \preceq q_1Aw$ and $p_2' \preceq wBq_2$. Hence we can put $p_1' = p_1Aw$ and $p_2' = wBp_2$ for some p_1 and p_2 such that $p_j \preceq q_j$ $(j = 1,2)$. It implies $p = p_1AwxwBp_2$. □

By the result above, we can construct the following counter-example for Theorem 11:

Example 1. Let $\Sigma = \{a_1, \cdots, a_k, b_1, \cdots, b_k\}$ be an alphabet with just $2k$ constants. We consider a regular pattern p and a set $Q = \{q_1, \cdots, q_k\} \in \mathcal{RP}^k$ given by

$$p = x_1a_1w_1xw_1b_1x_2, \quad q_i = x_1a_iw_ib_ix_2\ (i = 1,2,\cdots,k),$$

where $w_1, \cdots, w_k$ are defined recursively as follows:

$$w_i = w_{i+1}b_{i+1}a_{i+1}w_{i+1}\ (i = 1,2,\cdots,k-1), \quad w_k = \varepsilon.$$

For instance, in case $k = 3$, $w_3 = \varepsilon, w_2 = b_3a_3$ and $w_1 = (b_3a_3)b_2a_2(b_3a_3)$,

$$\begin{array}{l} p = x_1a_1((b_3a_3)b_2a_2(b_3a_3))x((b_3a_3)b_2a_2(b_3a_3))b_1x_2,\\ q_1 = x_1a_1((b_3a_3)b_2a_2(b_3a_3))b_1x_2, \quad q_2 = x_1a_2(b_3a_3)b_2x_2, \quad q_3 = x_1a_3b_3x_2. \end{array}$$

We will show that $p\{x := a_i\} \preceq q_i$ and $p\{x := b_i\} \preceq q_i$ $(i = 1,2,\cdots,k)$. For $i = 1$, we have $p\{x := a_1\} = (x_1a_1w_1)a_1(w_1b_1x_2) = q_1\{x_1 := x_1a_1w_1\} \preceq q_1$ and, similarly, $p\{x := b_1\} = q_1\{x_2 := w_1b_1x_2\} \preceq q_1$.

Next for $i \geq 2$, as easily seen, by the definition of w_i, we can put $w_1 = (w_ib_i)w^{(i)} = w'^{(i)}(a_iw_i)$ for some strings $w^{(i)}$ and $w'^{(i)}$. Thus for each $i \geq 2$,

$$\begin{aligned} p\{x := a_i\} &= (x_1a_1w_1)a_i(w_1b_1x_2) = (x_1a_1w_1)a_i(w_ib_iw^{(i)})b_1x_2\\ &= (x_1a_1w_1)(a_iw_ib_i)(w^{(i)}b_1x_2)\\ &= q_i\{x_1 := x_1a_1w_1,\ x_2 := w^{(i)}b_1x_2\}\\ &\preceq q_i, \quad \text{similarly,}\\ p\{x := b_i\} &= q_i\{x_1 := x_1a_1w'^{(i)},\ x_2 := w_1b_1x_2\}\\ &\preceq q_i. \end{aligned}$$

Hence $S_1(p) \subseteq L(Q)$. On the other hand, clearly $p \npreceq q_i$, and so $L(p) \not\subseteq L(q_i)$ $(i = 1, \cdots, k)$. □

4 $S_2(P)$ and Compactness

In this section, we show that $S_2(P)$ is a characteristic set for $L(P)$ within $\mathcal{RPL}^k$, under the condition $\sharp\Sigma \geq 2k-1$. As a result, the class $\mathcal{RP}^k$ has compactness with respect to containment.

Lemma 14. *Suppose $\sharp\Sigma \geq 3$. Let p and q be regular patterns. Then if $p\{x := r\} \preceq q$ for any $r \in D$, then $p\{x := xy\} \preceq q$ holds, where D is either one of the followings:*
(i) $D = \{ay, by\}$, or $D = \{ya, yb\}$, where $a \neq b$,
(ii) $D = \{a_1b_1, a_2b_2, a_3b_3\}$, where $a_i \neq a_j$ and $b_i \neq b_j$ for $i \neq j$.

Proof. If p does not contain the variable x, then it is clearly true. Thus we consider $p = p_1xp_2$ for some regular patterns p_1 and p_2. We prove only for the case (i) that if $p_1ayp_2 \preceq q$ and $p_1byp_2 \preceq q$, then $p_1xyp_2 \preceq q$ holds. We can prove for other cases similarly.

Assume $p_1xyp_2 \not\preceq q$. Let us put $p_2' = yp_2$ and $p' = p_1xyp_2 = p_1xp_2'$. Since $p'\{x := a\} \preceq q$, $p'\{x := b\} \preceq q$ but $p' \not\preceq q$, by Lemma 13, there exist regular patterns p_1'', p_2'', q_1 and q_2 and a string $w \in \Sigma^*$ such that $p_1 = p_1''Aw$, $p_2' = wBp_2''$ and $q = q_1AwBq_2$, where $\{A, B\} = \{a, b\}$. By $p_2' = wBp_2''$, head(p_2') must be a constant, which contradicts that $p_2' = yp_2$. □

Lemma 15. *Suppose $\sharp\Sigma \geq 3$. Let p and q be regular patterns. Then if $p\{x := a\} \preceq q$ for some $a \in \Sigma$ and $p\{x := xy\} \preceq q$, then $p \preceq q$ holds, where y is a variable not appearing in p.*

Proof. If p does not contain the variable x, then it is clearly true. Thus we consider $p = p_1xp_2$ for some regular patterns p_1 and p_2.

Assume the converse that $p_1ap_2 \preceq q$ and $p_1xyp_2 \preceq q$ but $p_1xp_2 \not\preceq q$. Similarly to the proof of Lemma 10, we can show that there exist regular patterns q_1 and q_2 and a string $w \in \Sigma^*$ such that

$$q = q_1AwBq_2, \quad p_j \preceq q_j \ (j = 1, 2), \quad p_1 \preceq q_1Aw, \quad p_2 \preceq wBq_2,$$

where $\{A, B\} = \{a, xy\}$.

Let $A = a$ and $B = xy$. By $p_1 \preceq q_1aw$ and $p_2 \preceq wx(yq_2)$, we can put $p_1 = p_1'aw$ and $p_2 = wp_2'p_2''$ for some p_1', p_2' and p_2'' such that $p_1' \preceq q_1$, $p_2' \preceq x$ and $p_2'' \preceq yq_2$. Consider a substitution $\theta = \{x := xwp_2'\}$. Then $p_1xp_2 = p_1'awxwp_2'p_2'' \preceq (q_1)awxwp_2'(yq_2) = ((q_1aw)x(yq_2))\theta = q\theta \preceq q$, which contradicts the assumption.

Similarly, we can show a contradiction for the case of $A = xy$ and $B = a$. □

For a nonempty finite set D of regular patterns, we denote

$$\text{head}(D) = \{\text{head}(p) \mid p \in D\}, \quad \text{tail}(D) = \{\text{tail}(p) \mid p \in D\}.$$

Lemma 16. *Suppose $\sharp\Sigma \geq 3$. Let p and q be regular patterns. Then if $p\{x := r\} \preceq q$ for any $r \in D$, then $p \preceq q$ holds, where D is either one of the followings:*
(i) $D = \{a, b, cy\}$, or $D = \{a, b, yc\}$, where $a \neq b$,
(ii) $D = \{a, by, cy\}$, or $D = \{a, yb, yc\}$, where $b \neq c$.

Proof. If p does not contain the variable x, then it is clearly true. Thus we consider $p = p_1xp_2$ for some regular patterns p_1 and p_2.

For the case (ii), by Lemma 14 (i), we have $p_1xyp_2 \preceq q$. Since $p_1ap_2 \preceq q$, we have $p \preceq q$ by Lemma 15.

We prove for $D = \{a, b, cy\}$ of the case (i). Assume the converse that $p_1ap_2 \preceq q$, $p_1bp_2 \preceq q$ and $p_1cyp_2 \preceq q$ but $p_1xp_2 \not\preceq q$. Then clearly the constant a in p_1ap_2 is not generated by variable substitution from q, and so are b in p_1bp_2 and cy in p_1cyp_2. If $p_1xyp_2 \preceq q$, since $p_1ap_2 \preceq q$, it follows by Lemma 15 that $p \preceq q$, and a contradiction. Thus we have $p_1xyp_2 \not\preceq q$. Similarly to the proof of Lemma 10, we can show that there exist regular patterns q_1 and q_2 and strings w and w' such that

(1) $q = q_1AwBw'Cq_2$, $p_i \preceq q_i$ $(i = 1, 2)$,
(2) $p_1 \preceq q_1Aw$, $p_1 \preceq q_1AwBw'$,
(3) $p_2 \preceq wBw'Cq_2$, $p_2 \preceq w'Cq_2$,

where $D = \{A, B, C\} = \{a, b, cy\}$. Note that head$(D) = \{a, b, c\}$ and tail$(D) = \{a, b, y\}$ (possibly $c = a$ or $c = b$, but $a \neq b$). Since Aw is a suffix of $AwBw'$ by (2) and $w'C$ is a prefix of $wBw'C$ by (3), it follows that $|w| \neq |w'|$ and $w, w' \neq \varepsilon$. Assume $|w| < |w'|$. Then there exist strings $w_1, w_2 \in \Sigma^+$ such that $w' = w_1w = ww_2$, and so A is a suffix of $AwBw_1$ and w_2C is a prefix of Bw_1wC.

If $A = cy$, then tail(w_1) = tail$(A) = y$, which contradicts that w_1 is a constant string.

If $B = cy$, then w_2C contains the variable y, because B is a prefix of w_2C. In this case, $C = a$ or b, and thus w_2C is a constant string. It is a contradiction.

Finally, if $C = cy$, then w_2cy is a prefix of the constant string Bw_1w, and a contradiction.

We can prove for the case of $|w| > |w'|$ similarly. □

Now we present the main theorem in this paper.

Theorem 17. *Suppose $k \geq 3$ and $\sharp\Sigma \geq 2k-1$. Let $P \in \mathcal{RP}^+$ and $Q \in \mathcal{RP}^k$. Then the following three propositions are equivalent:*
(i) $S_2(P) \subseteq L(Q)$, (ii) $P \sqsubseteq Q$, (iii) $L(P) \subseteq L(Q)$.

Proof. It suffices to show the case of $\sharp Q = k$ and $\sharp\Sigma = 2k-1$ and that of $\sharp Q = k$ and $\sharp\Sigma = 2k$. Other cases can be reduced to Theorem 11.

We show that for any regular pattern p, $S_2(p) \subseteq L(Q)$ implies $p \preceq q$ for some $q \in Q$, when $\sharp Q = k$ and $\sharp\Sigma = 2k-1$. The case of $\sharp Q = k$ and $\sharp\Sigma = 2k$ can be shown similarly. Put $Q = \{q_1, \cdots, q_k\}$.

The proof is done by a mathematical induction on the number n of variables in p. In case $n = 0$, $S_2(p) = \{p\}$, and so $p \in L(Q)$. Hence $p \preceq q$ for some $q \in Q$. Let $n \geq 0$ and assume that it is valid for any regular pattern with n variables. Let p be a regular pattern with $(n+1)$ variables such that $S_2(p) \subseteq L(Q)$.

Assume $p \not\preceq q_i$ $(i = 1, \cdots, k)$. Let x be any fixed variable in p and $p = p_1xp_2$ for some p_1 and p_2. For $a, b \in \Sigma$, put $p_a = p\{x := a\}$ and $p_{ab} = p\{x := ab\}$. Note that both p_a and p_{ab} contain just n variables and that $S_2(p_a) \subseteq L(Q)$ and

$S_2(p_{ab}) \subseteq L(Q)$ hold. By the induction hypothesis, for any $a, b \in \Sigma$, there exist $i, i' \leq k$ such that $p_a \preceq q_i$ and $p_{ab} \preceq q_{i'}$.

For each $i \leq k$, put $D_i = \{a \in \Sigma \mid p_a \preceq q_i\}$ and define a bigraph $G_i = (V, E_i)$, where the set V of vertices consists of two sets Σ and $\overline{\Sigma} = \{\overline{a} \mid a \in \Sigma\}$ and the set E_i of edges is defined by $E_i = \{(a, \overline{b}) \mid p_{ab} \preceq q_i\}$. Note that any cycle in a bigraph has even length. For each $a, b \in \Sigma$, $\deg_i(a)$ (resp., $\deg_i(\overline{b})$) means the number of b's (resp., a's) such that $p_{ab} \preceq q_i$. We note that, as easily seen, $\bigcup_{i=1}^{k} D_i = \Sigma$ and $\bigcup_{i=1}^{k} E_i = \Sigma \times \overline{\Sigma}$ hold.

If $\sharp D_i \geq 3$ for some i, then $p \preceq q_i$ by Lemma 10, and a contradiction. Thus $\sharp D_i \leq 2$ for any $i \leq k$. Moreover, since $\sharp\Sigma = 2k - 1$ and $\bigcup_{i=1}^{k} D_i = \Sigma$, it follows that $\sharp D_i \geq 1$ for any $i \leq k$.

Here we note that for the case of $\sharp D_i = 2$, by Lemma 16 (i), $p_1 a y p_2 \not\preceq q_i$ and $p_1 y b p_2 \not\preceq q_i$ hold for any $a, b \in \Sigma$. Therefore, by Lemma 10, there exist neither distinct constants a_j ($j = 1, 2, 3$) nor b_j ($j = 1, 2, 3$) such that $p_{ab_j} \preceq q_i$ and $p_{a_j b} \preceq q_i$, and thus $\deg_i(a) \leq 2$ and $\deg_i(\overline{b}) \leq 2$ for any $a, b \in \Sigma$. These mean that any connected component of the bigraph G_i is a cycle with even length or a chain.

For the case of $\sharp D_i = 1$, by Lemma 16 (ii), the following four cases are possible:

(1) $p_1 a y p_2 \not\preceq q_i$ and $p_1 y b p_2 \not\preceq q_i$ for any $a, b \in \Sigma$,
(2) $p_1 a y p_2 \not\preceq q_i$ and $p_1 y b p_2 \not\preceq q_i$ for any $a \in \Sigma - \{a_0\}$, $b \in \Sigma - \{b_0\}$,
(3) $p_1 a y p_2 \not\preceq q_i$ and $p_1 y b p_2 \not\preceq q_i$ for any $a \in \Sigma - \{a_0\}$, $b \in \Sigma$,
(4) $p_1 a y p_2 \not\preceq q_i$ and $p_1 y b p_2 \not\preceq q_i$ for any $a \in \Sigma$, $b \in \Sigma - \{b_0\}$,

where $a_0, b_0 \in \Sigma$ are some constant symbols.

For the case (1), any connected component of G_i is similarly shown to be a cycle with even length or a chain. For the case (2), we consider a bigraph G'_i obtained from the bigraph G_i by deleting the vertices a_0 and $\overline{b}_0$. Clearly $\deg_i(a) \leq 2$ and $\deg_i(\overline{b}) \leq 2$ in the bigraph G'_i for any $a \in \Sigma - \{a_0\}$ and any $b \in \Sigma - \{b_0\}$. Thus any connected component of the bigraph is a cycle with even length or a chain. For the cases (3) and (4), we can similarly get subbigraphs of G_i whose connected components are all cycles with even lengths or chains.

Hereafter, we prove the following claim:

Claim: For some $i_0 \leq k$, the bigraph G_{i_0} contains at least three edges $(a_j, \overline{b}_j)$, $j = 1, 2, 3$ such that $a_j \neq a_{j'}$, $\overline{b}_j \neq \overline{b}_{j'}$ for $j \neq j'$.

Proof of the claim. Since $\sharp\Sigma = 2k - 1$ and $\bigcup_{i=1}^{k} D_i = \Sigma$, it follows that $\sharp D_i = 2$ for at least $(k-1)$ i's, say, $1, 2, \cdots, k-1$, and $\sharp D_k = 1$ or 2.

(i) In case $\sharp D_i = 2$ for any $i \leq k$. As noted above, for any $i \leq k$, all connected components of the bigraph G_i are cycles with even lengths and chains. As mentioned above, for any $a, b \in \Sigma$, there exists $i \leq k$ such that $p_{ab} \preceq q_i$, and thus for any edge $(a, \overline{b})$, there exists $i \leq k$ such that $(a, \overline{b}) \in E_i$. Since there are $(2k-1)^2$ possible edges, there exists $i_0 \leq k$ such that the bigraph G_{i_0} contains at least $(4k-3)$ $(\geq (2k-1)^2/k)$ edges. Since $\deg_{i_0}(a) \leq 2$ for any $a \in \Sigma$, it means that $\deg_{i_0}(a) = 2$ for at least $(2k-2)$ a's, and $\deg_{i_0}(a) \leq 1$ for at most one a. Hence the bigraph G_{i_0} consists of some cycles with even lengths and at most one chain. Note that a cycle with length $2l$ contains distinct l edges which are not adjacent mutually. As easily seen, G_{i_0} contains a set of edges $(a_j, \overline{b}_j)$, $j = 1, \cdots, 2k-1$ which are not adjacent mutually. Since $k \geq 3$, we have $2k - 1 \geq 5$, and thus our claim is valid.

(ii) In case $\sharp D_i = 2$ for any $i < k$ and $\sharp D_k = 1$. For the case (1) mentioned above, we can show similarly to (i). Let us consider the case (2) above. For any $i \leq k$, let G'_i be a subbigraph obtained from G_i by deleting the vertices a_0 and $\overline{b}_0$. Then for any $i \leq k$, any connected component of the bigraph G'_i is a cycle with even length or a chain. Similarly to (i), since there are $(2k-2)^2$ possible edges $(a, \overline{b})$ each of which is contained in at least one bigraph considered, there exists $i_0 \leq k$ such that the bigraph G'_{i_0} contains at least $4k - 7$ $(\geq (2k-2)^2/k)$ edges. It implies that $\deg_{i_0}(a) = 2$ for at least $(2k-5)$ a's and $\deg_{i_0}(a) \leq 1$ for at most three a's. In particular, if $\deg_{i_0}(a) = 2$ for just $(2k-5)$ a's, $\deg_{i_0}(a) = 1$ for the other three a's. Moreover, if $\deg_{i_0}(a) = 2$ for just $(2k-4)$ a's, $\deg_{i_0}(a) = 1$ for at least one a. In any case, G_{i_0} contains a set of edges $(a_j, \overline{b}_j)$, $j = 1, \cdots, 2k-3$, which are not adjacent mutually. Since $k \geq 3$, we have $2k - 3 \geq 3$, and thus our claim is valid.

We can prove our claim for the cases (3) and (4) similarly. ■

Appealing to Lemma 14, we have $p_1 x y p_2 \preceq q_{i_0}$, and thus $p \preceq q_i$ by Lemma 15. This contradicts our assumption. □

As a direct corollary of this theorem, we have:

Corollary 18. *Suppose $k \geq 3$ and $\sharp\Sigma \geq 2k - 1$. Let $P \in \mathcal{RP}^+$. Then $S_2(P)$ is a characteristic set for the language $L(P)$ within $\mathcal{RPL}^k$.*

Lemma 19. *If $\sharp\Sigma \leq 2k-2$, then the class $\mathcal{RP}^k$ does not have compactness with respect to containment.*

Proof. Let $\Sigma = \{a_1, \cdots, a_{k-1}, b_1, \cdots, b_{k-1}\}$ be an alphabet with just $(2k-2)$ constants. Let p, q_i and w_i ($i = 1, \cdots, k-1$) be regular patterns and strings defined in Example 1, where $w_{k-1} = \varepsilon$. Then let $q_k = x_1 a_1 w_1 x y w_1 b_1 x_2$.

As shown in Example 1, $p\{x := a_i\} \preceq q_i$ and $p\{x := b_i\} \preceq q_i$ for $i = 1, 2, \cdots, k-1$, and thus $S_1(p) \subseteq \bigcup_{i=1}^{k-1} L(q_i)$. On the other hand, clearly, for any string w with $|w| \geq 2$, $p\{x := w\} \preceq q_k$. These mean $L(p) \subseteq L(Q)$. However, clearly, $p \not\preceq q_i$, and so $L(p) \not\subseteq L(q_i)$ $(i = 1, 2, \cdots, k)$. Therefore $\mathcal{RP}^k$ does not have compactness w.r.t. containment. □

By Theorem 17 and Lemma 19, we have the following theorem:

Theorem 20. *Suppose $k \geq 3$. Then the class $\mathcal{RP}^k$ has compactness with respect to containment if and only if $\sharp\Sigma \geq 2k - 1$.*

Note that, independent of ours, Arimura and Shinohara[6] showed that if $\sharp\Sigma \geq 2k + 1$, then the class $\mathcal{RP}^k$ has compactness w.r.t. containment, and so is not if $\sharp\Sigma = k + 1$. Theorem 20 completely fills the gap on the number of constant symbols.

The following example is a counter-example for Theorem 17 in case $k = 2$.

Example 2. Let $\Sigma = \{a, b, c\}$ be an alphabet with just 3 constants. We consider regular patterns p, q_1 and q_2 given by

$$p = x_1 axbx_2, \quad q_1 = x_1 abx_2, \quad q_2 = x_1 cx_2.$$

For any $w \in \Sigma^+$, if c appears in w, then $p\{x := w\} \preceq q_2$ holds. Otherwise, as easily seen, $p\{x := w\} \preceq q_1$ holds. It implies that $L(p) \subseteq L(q_1) \cup L(q_2)$, but clearly $p \not\preceq q_i$ $(i = 1, 2)$. □

Thus Theorem 17 is not always valid for $k = 2$. However, we obtain the following result for the case of $k = 2$.

Theorem 21. *Suppose* $\sharp\Sigma \geq 4$. *Let* $P \in \mathcal{RP}^+$ *and* $Q \in \mathcal{RP}^2$. *Then the following three propositions are equivalent:*
(i) $S_2(P) \subseteq L(Q)$, *(ii)* $P \sqsubseteq Q$, *(iii)* $L(P) \subseteq L(Q)$.

Proof. We can prove similarly to the proof of Theorem 17. □

As direct corollaries of this theorem, we have:

Corollary 22. *Suppose* $\sharp\Sigma \geq 4$. *Let* $P \in \mathcal{RP}^+$. *Then* $S_2(P)$ *is a characteristic set for the language* $L(P)$ *within* $\mathcal{RPL}^2$.

Corollary 23. *Suppose* $\sharp\Sigma \geq 4$. *Then the class* $\mathcal{RP}^2$ *has compactness with respect to containment.*

The following corollary would be very useful in the theory of inductive inference of recursive languages from positive data from the view point of efficiency:

Corollary 24. *Suppose* $k \geq 3$ *and* $\sharp\Sigma \geq 2k - 1$. *Let* $P \in \mathcal{RP}^+$ *and* $Q \in \mathcal{RP}^k$. *Then for any subset* S *of* $L(Q)$, *if* $S_2(P) \subseteq S$, *then* $P \sqsubseteq Q$ *holds.*

Furthermore, because for any regular patterns p and q, whether or not $p \preceq q$ is computable in time polynomial of the sum of lengths of p and q (cf. Shinohara[12]), it follows that containment problem for unions are efficiently computable, that is, we have the following corollary:

Corollary 25. *Suppose* $k \geq 3$ *and* $\sharp\Sigma \geq 2k - 1$. *For any* $P \in \mathcal{RP}^+$ *and* $Q \in \mathcal{RP}^k$, *whether or not* $L(P) \subseteq L(Q)$ *is computable in time polynomial of the total length of patterns appearing in* P *and* Q.

References

1. Angluin, D.: *Finding patterns common to a set of strings*, in Proceedings of the 11th Annual Symposium on Theory of Computing (1979) 130–141.
2. Angluin, D.: *Inductive inference of formal languages from positive data*, Information and Control **45** (1980) 117–135.

3. Arikawa, S., Shinohara, T. and Yamamoto, A.: *Elementary formal systems as a unifying framework for language learning*, in Proceedings of the 2nd Annual ACM Workshop on Computational Learning Theory (1989) 312–327.
4. Arikawa, S., Kuhara, S., Miyano S., Mukouchi, Y., Shinohara, A. and Shinohara, T.: *A machine discovery from amino acid sequences by decision trees over regular patterns*, in Proceedings of the International Conference on Fifth Generation Computer Systems (1992) 618–625 (also in New Generation Computing **11**(3, 4) (1993) 361–375).
5. Arimura, H., Shinohara, T. and Otsuki, S.: *Finding minimal generalizations for unions of pattern languages and its application to inductive inference from positive data*, in Proceedings of the 11th Symposium on Theoretical Aspects of Computer Science, Lecture Notes in Computer Science **775** (1994) 646–660.
6. Arimura, H. and Shinohara, T.: *Compactness for unions of regular pattern languages*, in Proceedings of the Symposium on Language and Automaton, Research on Computational Models and Complexity, RIMS Koukyuroku **950** (1996) 246–249 (in Japanese).
7. Gold, E.M.: *Language identification in the limit*, Information and Control **10** (1967) 447–474.
8. Kobayashi, S. and Yokomori, T.: *Identifiability of subspaces and homomorphic images of zero-reversible languages*, in Proceedings of the 8th International Workshop on Algorithmic Learning Theory, Lecture Notes in Artificial Intelligence **1316** (1997) 48–61.
9. Motoki, T., Shinohara, T. and Wright, K.: *The correct definition of finite elasticity: corrigendum to identification of unions*, in Proceedings of the 4th Annual ACM Workshop on Computational Learning Theory (1991) 375–375.
10. Mukouchi, Y.: *Characterization of pattern languages*, in Proceedings of the 2nd Workshop on Algorithmic Learning Theory (1991) 93–104.
11. Sato, M.: *Inductive inference of formal languages*, Bulletin of Informatics and Cybernetics **27**(1) (1995) 85–106.
12. Shinohara, T.: *Polynomial time inference of pattern languages and its applications*, in Proceedings of the 7th IBM Symposium on Mathematical Foundations of Computer Science (1982) 191–209.
13. Smullyan, R.M.: "Theory of formal systems," Princeton University Press, 1961.
14. Wright, K.: *Identification of unions of languages drawn from an identifiable class*, in Proceedings of the 2nd Annual Workshop on Computational Learning Theory (1989) 328–333.