COUNTING WITH RATIONAL FUNCTIONS

C. CHOFFRUT
Université de ROUEN
Faculté des Sciences
B.P. 67
76130 MONT-SAINT-AIGNAN
M.P. SCHUTZENBERGER
Université Paris 7
U.E.R. de Mathématiques et d'Informatique
2, place Jussieu, Tout 55-56, 1er étage
75251 PARIS CEDEX 05

Abstract : Rational functions of a free monoid A* into the free cyclic monoid t* generated by a unique element t, can be viewed as assigning an integer to every word $u \in A^*$. We investigate those functions which count occurrences of some fixed (and special) subsets $X \subseteq A^*$ in all words of A* and show that they can be characterized in terms of "bounded variation", a notion which is close to continuity of functions.

Résumé : Les fonctions rationnelles d'un monoïde libre A* dans le monoïde libre cyclique t* engendré par l'unique élément t peuvent être considérées comme des applications qui à tout mot $u \in A^*$ associent un entier. Nous étudions plus précisément celles qui peuvent être définies comme comptant les occurrences d'un ensemble fixe (et particulier) $X \subseteq A^*$ dans chaque mot de A*. Nous montrons qu'elles peuvent être caractérisées en termes de "variation bornée" une notion proche de celle de continuité des fonctions.

1. INTRODUCTION.

Rational functions of a free monoid A* into another B* are obtained by providing a finite, not necessarily deterministic automaton $\mathcal{A}$ with an output function, associating thus a word in B* with every transition of $\mathcal{A}$. Morphisms of A* into B* are special cases of such functions.

Numerous areas of computer science are directly dealing with rational functions : codes (encoding and decoding of messages), lexical analysers (assigning a token to some portion of a program), sorting (the Soundex encoding of surnames, cf. [Kn p. 391], defines a special case of rational function [Jo]), text editing (systems like Multics or Unix provide a large range of commands substituting to all occurrences of some rational expression a given word) etc...

The importance of rational functions is also theoretical since they play, with respect to free monoids, a role similar to that of rational fractions with

respect to complex numbers. In both cases these functions are directly defined from the structure on which they act (concatenation in the former case, addition and multiplication in the latter).

We are here concerned with a problem which has long been considered in automata theory where it was asked what automata could possibly count. Refining the notion of threshold and modulo counting, various classes of rational languages were defined. In the present paper we consider rational functions α of the input monoid A^* into the free cyclic monoid t^* generated by the single symbol t. Then for every $u \in A^*$ the length of its image by α, i.e., $|u\alpha|$, is an integer and therefore "counts" something. More precisely, we are interested in characterizing these rational functions which count occurrences of some rational subset H of A^* : we call such functions counting. A nécessary condition on H in order to insure the linear growth of the image $u\alpha$ is that there do not exist an infinite chain (in the ordering "being factor of") of distincts words. This again is equivalent to H being a finite union of rational semaphors i.e. of rational subsets $X \subseteq A^*$ for which two elements may not be a proper factor of each other (cf. [BePer], Chap. II,5).

The characterization requires a notion close to continuity of functions as used in analysis. We say that a function $\alpha : A^* \to t^*$ has bounded variation if there exists an integer $k<0$ such that :

$$-kn \leqslant |u\alpha| - |v\alpha| \leqslant kn$$

holds for all $u,v \in A^*$ where n is the minimum number of letters which have to be erased in u and v in order to obtain a common subword. Then our main result states that (after a possible partition of A^* into finitely many rational subsets, which we ignore here for the simplicity of the exposition) counting and bounded variation are essential the same notions for rational functions.

Basics on rational functions are recalled in Section 2. In section 3 we define various distances on free monoids and the related notions of bounded variation (with respect to one type of distance) by their rational transducers, and focuses on the case where the image free monoid is cyclic. The last section gives a sketch of the proof.

2. PRELIMINARIES.

Let A a non empty finite set - or alphabet, A^* and A^+ respectively the free monoid and semigroup which it generates. An element w of A^* is a word and its length is denoted by $|w|$. The identity of A^* or empty word is denoted by 1 : $A^+ = A^* - \{1\}$.

Given arbitrary sets X and Y, we shall view a relation ρ from X to Y, denoted by $\rho : X \to Y$, as an application of X into the power set of Y. Such a relation is a function if the image of every element $x \in X$ contains at most an element, i.e. if it defines a partial mapping of X into Y. The relation operates to the right, writing thus $x\rho$ instead of $\rho(x)$ and the domain of ρ is denoted by :

$$\mathrm{dom}\rho = \{x \in X \mid \exists y \in Y \ y \in x\rho\}$$

We are mainly concerned with relations ρ from a free monoid A^* into another B^*, and more precisely with those which are rational i.e. whose graph :

$$\rho = \{(u,v) \in A^* \times B^* \mid v \in u\rho\}$$

is a rational subset of the product monoid $A^* \times B^*$ (cf. [Ei] P. 236).

Designating by RatM the semiring of all rational subsets of an arbitrary monoid M, the rational relations are characterized as follows (cf. e.g. [Be] Thm. 7.1. or [Ni]) :

THEOREM 1 :

Let $\rho : A^* \to B^*$ be a relation. The following conditions are equivalent :

1) ρ is rational

2) there exist a finite non empty subset Q and a morphism μ of A^* into the multiplicative monoid of QxQ-matrices with entries in Rat B^*, an Q-row vector λ and a Q columm vector γ such that

$$u\rho = \lambda(u\mu)\gamma$$

holds for all $u \in A^*$.

We say that the triple (λ,μ,γ), or simply μ, is a transducer realizing ρ. The set Q is usually called the set of states and its cardinality n is the dimension of the transducer.

A rational function is simply a rational relation which is a function (notice that this property is decidable, (cf. [Sch 1] p. 245). Then without loss of generality we may assume from now on that all entries in λ,γ and $a\mu$ ($a \in A$) are in $B^* \cup \{\emptyset\}$.

Particularly important both from historical and conceptual points of view are the sequential functions studied by Ginsburg and Rose, cf. [GiRo], which are rational functions realized by transducers where λ has all entries equal to 0 except one equal to 1, γ has all entries equal to 1 and for every $a \in A$, $a\mu$ is row monomial i.e. has at most one non zero entry in each row. The resulting transducer and by extension μ itself is called sequential. In this case, taking advantage of the "monomiality" of the matrices, the following more concise notations are useful. Let $q\_ \in Q$ be the index of the non empty entry of the vector λ. Then for every $q \in Q$ and $u \in A^*$ we set $q.u=q'$ if q' is the unique, if any, element of Q satisfying $u\mu_{qq'} \neq \emptyset$. With our conventions, $q.u=\emptyset$, if the q-row of $u\mu$ has only empty entries. Further, we set $q^*u=u\mu_{qq'}$, where $q'=q.u \in Q$ and $q^*u=\emptyset$ if $q.u=\emptyset$.

Then the following identities are easily verified :

(1) $q.(uv)=(q.u).v$ for all $q \in Q$ $u,v \in A^*$

(2) $q^*uv=(q^*u)((q.u)^*v)$ for all $q \in Q$ $u,v \in A^*$

(3) $q\_^*u=u\alpha$ for all $u \in A^*$.

The subsequential functions have been introduced in [Sch2] where they were shown to satisfy a noticeable functional equation. They generalize the notion of sequential functions in the sense that γ is an arbitrary vector with entries in $B^* \cup \{\emptyset\}$ while λ and μ satisfy the same conditions as for sequential functions.

3. DISTANCES. BOUNDED VARIATIONS.

For i = 1,2,3,4 we define the different distances over A* by setting :

$$d_i(u,v) = |u| + |v| - 2\,L_i(u,v)$$

where :

$L_1(u,v) = \max\{|z| \mid u,v \in zA^*\}$

$L_2(u,v) = \max\{|z| \mid u,v \in A^*z\}$

$L_3(u,v) = \max\{|z| \mid u,v \in A^*zA^*\}$

$L_4(u,v) = \max\{|z| \mid u,v \in A^*z_1A^*\ldots z_rA^* \text{ where } z=z_1\ldots z_r\}$.

These distances are respectively called prefix - suffix - factor - and subword - distance.

Assume A* and B* are respectively equipped with the distances d_i and d_j ,i,j$\in\{1,2,3,4\}$. Then a function $\alpha : A^* \to B^*$ has bounded variation (abbreviated b.v.) for (d_i,d_j) if there exists k>0 such that $d_i(u\alpha,v\alpha) < k\, d_i(u,v)$ holds for all u,v$\in$ dom α.

In the case where i=j we simply way that α has b.v. for d_i - or b.v. without any reference to d_i when the distance is understood.

Some families of rational functions can be characterized in terms of distances. As a striking example let us recall the following result which extends Ginsburg and Rose's theorem on sequential functions (cf. [Ch]).

THEOREM 1

A function $\alpha : A^* \to B^*$ is subsequential iff it satisfies the following conditions :

1) α has bounded variation for the prefix distance

2) for all Y $\in$ Rat B we have*

$$Y\alpha^{-1} = \{x\in A^* \mid x\alpha\in Y\}\in \text{Rat } B^*.$$

It is possible to compare the different notions of bounded variations when A* and B* are arbitrarily equipped with the different distances. As an example we have :

PROPOSITION 2.

Let $\alpha : A^* \to B^*$ be a rational function which has b.v. for the factor distance. Then it has b.v. for the subword distance.

The converse does not hold in general. However, when B is reduced to one single element, then every rational function which has b.v. for the subword

distance, has b.v. for the prefix - suffix - and factor - distances. As a consequence of Theorem 1 we have :

Corollary 3

If B is reduced to one letter, then every rational function $\alpha : A^* \to B^*$ which has b.v. for the subword distance is subsequentifal.

4. TRANSDUCERS OF FUNCTIONS WITH b.v.

In this paragraph we assume A* and b* are equipped with the subword distance.

The functions with b.v. may be characterized by their transducers.

THEOREM 1

Let $\alpha : A^* \to B^*$ be a rational function realized by a transducer μ. Then the following conditions are equivalent :

1) α has bounded variations
2) there exists an integer $k>0$ such that for all $w \in A^*$ and for any two entries $x,x' \in B^*$ of the matrix $w\mu$, we have : $d(x,x')<k$
3) for all $w \in A^*$ and for any two diagonal entries $x = w\mu_{q,q} \in B^*$, $x' = w\mu_{q',q'} \in B^*$ the words x ans x' are conjugate, i.e. $xy=yx'$ for some $y \in B^*$.

The implications 1) $\Rightarrow$ 2) $\Rightarrow$ 3) are proved by standard arguments. The crux of the implication 3) $\Rightarrow$ 1) it the following result which is interesting in its own right.

LEMMA 2

Let S be a semigroup of the product monoïd $B^* \times B^*$ such that u and v are conjugate for all $(u,v) \in S$. Then for some $z \in B^*$ the equality

$$uz = zv$$

holds for all $(u,v) \in S$.

Based on this last technical result and using an iteration property (see e.g. [Sch1]) the proceding result provides an algorithm to decide whether a given rational function has b.v. for d_4.

Let us study more precisely the case where B consists of one single letter t. By Theorem 3.1 and Corollary 3.3 any function $\alpha : A^* \to t^*$ which has b.v. can be

realized by a subsequential transducer (λ,μ,γ), where as we saw, $u\mu$ is a monomial matrix for all $u\in A^*$. This leads to the following definitions.

For any positive integer $n>0$ let $\mathfrak{M}_n$ (or simply $\mathfrak{M}$) denote the monoid of nxn matrices with entries in $\{\emptyset\}\cup t^*$ which are row-monomial.

Given a matrix $m\in\mathfrak{M}$ let $m\rho$ be the longest factor common to all non zero entries of m and $m\pi$ be the unique matrix satisfying :

$$m = m\rho\ m\pi$$

Thus, unless m is the empty matrix, $m\pi$ necessarily has an entry equal to the empty word 1.

The following identities are straightforward :

(1) $m_1 m_2 \pi = (m_1\pi)\ (m_2\pi)\pi$

(2) $m_1 m_2 \rho = m_1\rho m_2\rho (m_1\pi m_2\pi)\rho$.

As a consequence, given any morphism $\mu : A^* \to \mathfrak{M}$, the relation $u\underset{\mu}{\sim}v$ (or more simply $u \sim v$ when μ is understood) defined for all $u,v \in A^*$, such that $u\mu\pi=v\mu\pi$ is a congruence. Then the following is the converse of Corollary 3.3.

PROPOSITION 3

Let $\alpha : A^* \to t^*$ be a rational function realized by a subsequential transducer μ. Then α has b.v. iff the congruence $\underset{\mu}{\sim}$ as finite index.

5 COUNTING FUNCTIONS.

In this paragraph, B is reduced to the letter t.

By a semaphor $H\subseteq A^*$ we mean a subset containing no proper factor :

$$H\cap(A^*HA^+\cup A^+HA^*) = \emptyset$$

A function $\alpha : A^* \to t^*$ with recognizable domain $X=dom\alpha$ is an elementary counting function if there exists a recognizable semaphor $H\in RecA^*$ and two rational numbers $r,s\in Q$ such that :

(1) $$|w\alpha| = r|w|_H + s$$

where $|w|_H$ denotes the number of different occurrences of H in w We say that α counts H.

A counting function is defined as a function $\alpha : A^* \to t^*$ for which there exists a partition of A^* into finitely many recognizable subsets :

$A^* = \bigcup_{1 \leqslant i \leqslant n} X_i$ and for each $1 \leqslant i \leqslant n$ a family $\alpha_{i_1}, \ldots, \alpha_{i_m}$ of elementary counting functions of domain X_i such that :

$$w\alpha = w\alpha_{i_1} \ldots w\alpha_{i_m}$$

holds for all $w \in X_i$.

It is not hard to see that the restrition of α to X_i has b.v. but α itself has not necessarily b.v.

Notice that every rational function with finite image is a counting function (take r=0 in (1)).

It is not difficult to construct a transducer realizing a counting function, i.e. :

PROPOSITION 1

Every counting function is rational.

The main theorem requires a few closure properties of counting functions.

PROPOSITION 2

Let A,B be two alphabets and $\beta : B^* \to t^*$ be a counting function.

Assume $A = A_1 \cup A_2$ where $A_2 \neq \emptyset$ is a partition and let $\gamma : B^* \to A^*$ be a rational substitution such that $\bigcup_{b \in B} b\alpha$ is a partition of $A_1^* A_2$.

Then $\gamma^{-1}\beta : A^* \to t^*$ is a counting function.

Reducing the proposition to the case where β is an elementary counting function counting a semaphor H as in the definition, the proof consists in showing that the function $\gamma^{-1}\beta$ "counts" occurrences of the semaphor :

$$K = A_2(H\alpha) \subseteq A^*$$

PROPOSITION 3

Let $D = A \cup B$ be a partition with $A \neq \emptyset$ and $B \neq \emptyset$ and let $\alpha : A^* \to t^*$ be a counting function. Consider the function $\beta : D^* \to t^*$ defined by :

$$u\beta = \begin{cases} 1 \text{ if } u \in A^* \\ u_0\alpha \ldots u_n\alpha \text{ if } u_0, \ldots, u_n \in A^* \\ \text{and } u = u_0 B u_1 \ldots B u_n \end{cases}$$

Then β is counting function.

As in the preceding Proposition the proof consists in showing that if α

"counts" occurrences of $H \subseteq A^*$ then β "counts" all occurrences of H and of B.

6 PROOF OF THE THEOREM.

The main result of this paper is the following :

Theorem

Let $\alpha : A^* \to t^*$ be a rational function with b.v. Then α is a counting function.

We shall briefly sketch the proof of this result.

Let $\mu : A^* \to \mathbb{N}$ be a sequential transducer of dimension n realizing α. For every word $u \in A^*$ let $u\sigma$ be the support of the matrix $u\mu$, i.e. the boolean matrix of dimension n defined by :

$$u\sigma_{ij} = \begin{cases} 1 \text{ if } u\mu_{ij} \neq \emptyset \\ \emptyset \text{ otherwise} \end{cases}$$

We proceed by induction on the cardinality of the monoid $A^*\sigma = M$. Indeed, by Krohn and Rhodes' classification, M satisfies one of the following conditions (cf.e.g. [La] Lemma 7.2.7) :

1) M is cyclic.
2) M is reduced to one L-class (possibly with an identily).
3) There exists a partition $A = A_1 \cup A_2$ such that $A_1^*\sigma$ and $(A^*A_2)\sigma$ are proper subsemigroups of M.

The main arguments for the first two cases which constitue the basis of the induction, are present in the case where M is a group of permutation (i.e. every letter $a \in A$ defines a permutation of the set of states) which we prove here thoroughly.

Using the notations of section 2, let $\phi : A^* \to \mathbb{Q}$ where $\mathbb{Q}$ is the additive monoid of rational numbers be the function defined by :

$$u\phi = \frac{1}{n} \sum_{q \in Q} |q^*u|$$

Since every $u \in A^*$ defines a permutation on Q, ϕ is a morphism.

Assume $u \in A^*$ acts as the identity on Q. Then because of assertion 3) of Theorem 4.1 we have : $|q^*u| = u\phi$ for all $q \in Q$. More generally the assumption $q.u = q$ implies $|q^*u| = u\phi$.

Let us now prove that for an arbitrary $u \in A^*$, $|u\alpha| - u\phi$ depends only on $q_{\_}.u=q$. Indeed, consider a fixed $w \in A^*$ such tat $q_{\_}.uw=q_{\_}$. Then by definition we have :

$$|uw\alpha| = |u\alpha| + |q^*w| = uw\phi$$

i.e.

$$|u\alpha| - \sum_{a \in A} (a\phi)\ |u|_a = (q^*w + w\phi)$$

which completes the proof, by taking H=A.

The induction itself consists essentially in proving that the restriction α' of α to $A^* A_2$ is a counting function. This is done by associating with every element of $A_1^* A_2$ its class in the congruence $\underset{\mu}{\sim}$. Let B be a finite alphabet in bijection with the set of the classes thus obtained and τ : $B^* \to (A_1^* A_2)^* = 1 + A^* A_2$ the rational substitution naturally defined. Then α may be factrorized as $\alpha' = \tau^{-1} \beta$ where β : $B^* \to t^*$ is a counting function (induction hypothesis). It then suffices to apply Proposition 5.2.

7. REFERENCES.

[Be] **BERSTEL J.**, "Transductions and Context-Free Languages", Teubner, 1979

[Be Per] **BERSTEL J.,& D. PERRIN**, "Theory of codes", 1985, Academic Press.

[Ch] **CHOFFRUT C.**, A generalization of Ginsburg and Rose's characterization of g-s-m mappings. Proceedings of the 6th ICALP Conférence, 1979, p. 88-103.

[Ei] **EILENBERG S.**, "Automata, Languages and Machines", Vol. A. 1974, Academic Press.

[GiRo] **GINSBURG S. & G.F. ROSE**, A characterization of machine mappings, Can. J. of Math., 18, 1986, p. 381-388.

[Jo] **JOHNSON J.H.**, Formal models for string similarity, PhD thesis, University of Waterloo, 1983 (also Research Report C8-83-32).

[Kn] **KNUTH D.E.**, " The Art of Computer Programming", vol. 3., 1973, Addison-Wesley.

[La] **LALLEMENT G.**,"Semigroups and Combinatorial Applications", 1979, Wiley-Interscience.

[Ni] **NIVAT M.**, Transductions des langages de Chomsky, Ann. de l'Inst. Fourier, 18, 1986 p. 339-456.

[Sch 1] **SCHUTZENBERGER M.P.**, Sur les relations rationnelles entre monoïdes libres, Theoret. Comput. Sci., 3, 1976, p. 243-259.

[Sch 2] **SCHUTZENBERGER M.P.**, Sur une variante des fonctions séquentielles, Theoret. Comput. Sci., 4, 1977, p. 243-259.