Computing and Informatics, Vol. 27, 2008, 441–451

INFORMATION THEORY OVER MULTISETS

Cosmin Bonchis, Cornel Izbasa

Research Institute "e-Austria" Timişoara, Romania
e-mail: cosmin@ieat.ro, cornel@ieat.ro

Gabriel Ciobanu

"A. I. Cuza" University of Iaşi, Faculty of Computer Science
Blvd. Carol I No. 11, 700506 Iaşi
e-mail: gabriel@info.uaic.ro

Revised manuscript received 3 December 2007

Abstract. Starting from Shannon theory of information, this paper presents the case of producing information in the form of multisets, and encoding information using multisets. We review the entropy rate of a multiset information source and derive a formula for the information content of a multiset. Then we study the encoder and channel part of the system, obtaining some results about multiset encoding length and channel capacity.

1 MOTIVATION

The attempt to study information sources which produce multisets instead of strings and ways to encode information on multisets rather than strings originates in observing new computational models like membrane systems which employ multisets [5]. Membrane systems have been studied extensively and there exist several results regarding their computing power, language hierarchies and complexity. However, while any researcher working with membrane systems (called also P systems) would agree that P systems are processing information, and that living cells and organisms do this too, we are unaware of any attempt to precisely describe natural ways to encode information on multisets or to study sources of information which produce multisets instead of strings. One could argue that, while some of the information in a living organism is encoded in a sequential manner, like in DNA for example,

there might be important molecular information sources which involve multisets (of molecules) in a non-trivial way.

Let us start with a simple question: given a P system with one membrane and, say, 2 objects a and 3 objects b from a known vocabulary V (suppose there are no evolution rules), how much information is present in such a system? Moreover, many examples of P systems perform various computational tasks; these systems encode the input (usually numbers) in various ways, either by superimposing a string-like structure on the membrane system [1], or by using the natural encoding of unary numeral system, that is, the natural number n is represented with n objects, for example, a^n. However, just imagine a gland which uses the bloodstream to send molecules to some tissue which, in turn, sends back some other molecules. There is an energy and information exchange. How can we describe it? Related questions are: what are the natural ways to encode numbers (information) on multisets, and how to measure the encoded information?

If membrane systems, living cells and any other (abstract or concrete) multiset processing machines are understood as information processing machines, then we believe that such questions should be investigated. We start from the idea that a study of multiset information theory might produce useful results at least in systems biology; if we understand the *natural* ways to encode information on multisets, there is a chance that *Nature* might be using similar mechanisms.

Another way in which this investigation seems interesting to us is that there is more challenge in efficiently encoding information on multisets (till now they constitute a poorer encoding media compared to strings). Encoding information on strings or even richer, more organized and complex structures is obviously possible and has been studied. Removing the symbol order or their position in the representation as strings can lead to multisets carrying a certain penalty, which deserves a precise description. Order or position do *not* represent essential aspects for information encoding; symbol multiplicity, a native quality of multisets, is *enough* for many purposes. We focus mainly on such "natural" approaches to information encoding over multisets, and present some advantages they have over approaches which superimpose a string structure on the multiset. Then we encode information using multisets in a similar way as it is done using strings.

There is also a connection between this work and the theory of numeral systems. The study of number encodings using multisets can be seen as a study of a class of purely non-positional numeral systems.

In Section 2 we derive a formula for the information content of a multiset, and in Section 3 we compute the multiset channel capacity.

2 ENTROPY RATE OF AN INFORMATION SOURCE

Shannon's information theory represents one of the great intellectual achievements of the twentieth century. Information theory has had an important and significant

influence on probability theory and ergodic theory, and Shannon's mathematics is a considerable and profound contribution to pure mathematics.

Shannon's important contribution comes from the invention of the source–encoder–channel–decoder–destination model, and from the elegant and general solution of the fundamental problems which he was able to pose in terms of this model. Shannon has provided significant demonstration of the power of coding with delay in a communication system, the separation of the source and channel coding problems, and he has established the fundamental natural limits on communication. As time goes on, the information theoretic concepts introduced by Shannon become more relevant to the increasingly complex process of communication.

2.1 Short Review of Shannon Information Theory

We use the notions defined in the classical paper [6] where Shannon has formulated a general model of a communication system which is tractable to a mathematical treatment.

Consider an information source modelled by a discrete Markov process. For each possible state i of the source there is a set of probabilities $p_i(j)$ associated to the transitions to state j. Each state transition produces a symbol corresponding to the destination state, e.g. if there is a transition from state i to state j, the symbol x_j is produced. Each symbol x_i has an initial probability $p_{i=\overline{1\ldots n}}$ corresponding to the transition probability from the initial state to each state i.

We can also view this as a random variable X with x_i as events with probabilities p_i, $X = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$.

There is an entropy H_i for each state. The entropy rate of the source is defined as the average of these H_i weighted in accordance with the probability P_i of occurrence of the states:

$$H(X) = \sum_i P_i H_i = -\sum_{i,j} P_i p_i(j) \log p_i(j). \tag{1}$$

Suppose there are two symbols x_i, x_j and $p(i,j)$ is the probability of the successive occurrence of x_i and then x_j. The entropy of the joint event is

$$H(i,j) = -\sum_{i,j} p(i,j) \log p(i,j).$$

The probability of symbol x_j to appear after the symbol x_i is the conditional probability $p_i(j)$.

Remark 1. The quantity $H(X)$ is a reasonable measure of choice or information.

String Entropy. Consider an information source X which produces sequences of symbols selected from a set of n independent symbols x_i with probabilities p_i. The

entropy formula for such a source is given in [6]:

$$H(X) = \sum_{i=1}^{n} p_i \log_b \frac{1}{p_i}.$$

2.2 Multiset Entropy

We consider a discrete information source which produces multiset messages (as opposed to string messages). A message is a multiset of symbols. The entropy rate of such a source is proved to be zero in [7]:

$$H(X_{multiset}) = \lim_{n\to\infty} \frac{1}{n} H(\{m_i\}_{i=1}^{n}) = 0.$$

Information Content. Following [4], the *information content* of an outcome x is

$$h(x) = \log \frac{1}{P(x)} \tag{2}$$

where $P(x)$ is the probability of the multiset x.
Let $k \in \mathbb{N}$ and $X = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$ a random variable, and $x = x_1^{m_1} x_2^{m_2} \dots x_n^{m_n}$ a multiset over symbols from X with $\sum_{i=1}^{n} m_i = k$. The probability of the outcome x is given by the multinomial distribution $\begin{pmatrix} k \\ m_1, m_2, \dots, m_n \end{pmatrix} \prod_{i=1}^{n} p_i^{m_i}$:

$$P[x \quad = \quad (m_1, m_2, \dots, m_n)] = \frac{(\sum_{i=1}^{n} m_i)!}{\prod_{i=1}^{n} m_i!} \prod_{i=1}^{n} p_i^{m_i}.$$

So, the information content of the multiset x is:

$$\begin{aligned} h(x = x_1^{m_1} x_2^{m_2} \dots x_n^{m_n}) &= \log \frac{1}{P[x]} = \log \left(1 / \frac{(\sum_{i=1}^{n} m_i)!}{\prod_{i=1}^{n} m_i!} \prod_{i=1}^{n} p_i^{m_i} \right) \\ &= \log \frac{\prod_{i=1}^{n} m_i!}{(\sum_{i=1}^{n} m_i)! \prod_{i=1}^{n} p_i^{m_i}}. \end{aligned}$$

Remark 2. The results and procedures presented in this paper refer mainly to deterministic P systems. A deterministic P system has the entropy rate converging to zero, and the information content of a unique configuration converging to zero.

Example 1. As an example, we consider a P system described by Figure 1. Essentially, a P system is a multiset of objects and a set of rules. By applying the rules, we can generate all the possible configurations (multisets of objects), and their probabilities of being generated at each step of the execution.

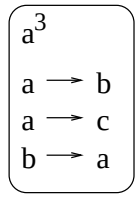

Figure 1. An example of a simple P system

Using all the possible configurations, the information content is computed for each configuration, and then represented in Figure 2. Only the information content of c^3 goes closer to 0; this means that the probability of this configuration goes closer to 1. Therefore c^3 has the highest probability of being the final configuration of the system.

The entropy rate is computed for these configurations at each step of the evolution, and this is represented in Figure 3.

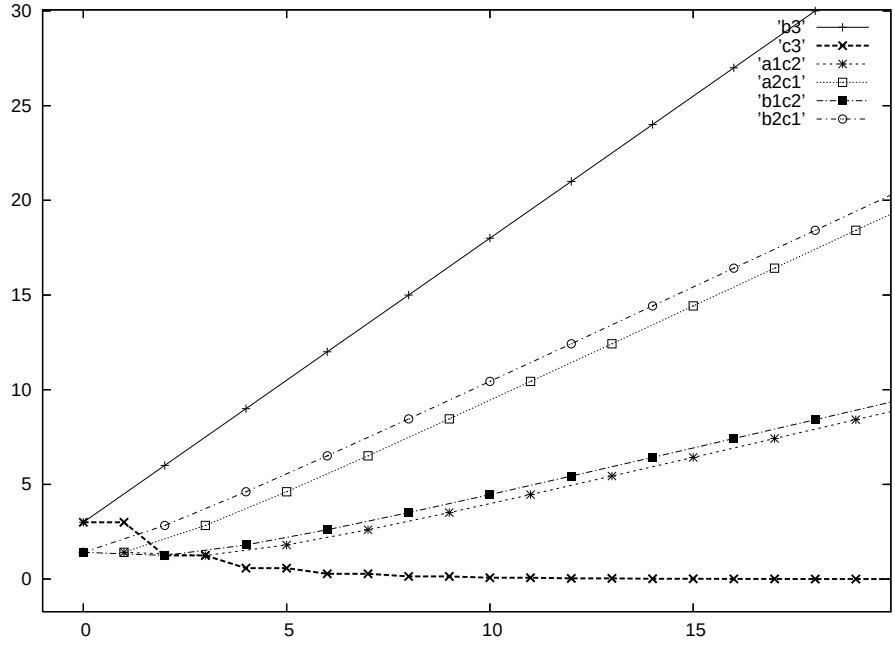

Figure 2. Information content

The entropy converges to 0, meaning that the system is deterministic and in time a configuration will appear with a probability converging to 1. Looking to Figure 2, we can identify c^3 as the (only possible) final result of the evolution.

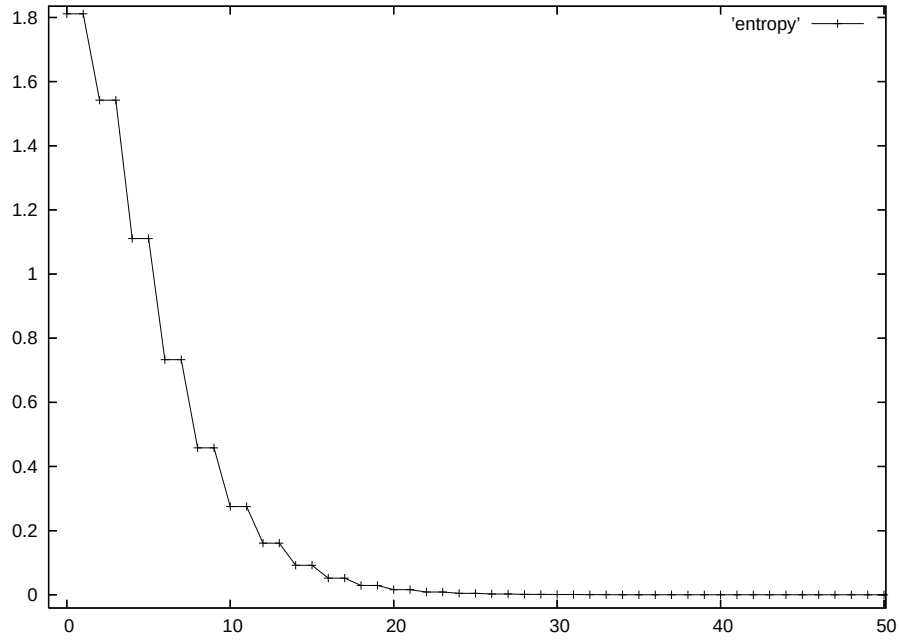

Figure 3. Entropy rate

3 MULTISET ENCODING AND CHANNEL CAPACITY

After exploring the characteristics of a multiset generating information source, we move to the channel part of the communication system. Properties of previously developed multiset encodings are analyzed in [2, 3]. The capacity of multiset communication channel is derived based on Shannon's definition and on the capacity theorem. We can have a multiset information source, and a usual sequence-based encoder and channel. All the following combinations are possible:

Source/Encoder	Sequential	Multiset
Sequential	[6]	this paper
Multiset	this paper	this paper

3.1 String Encoding

We shortly review the results concerning the string encoding.

Encoding Length. We have a set of symbols X to be encoded, and an alphabet A. We consider the uniform encoding. Considering the length l of the encoding, then $X = \{x_i = a_1 a_2 \dots a_l | a_j \in A\}$.

If $p_i = P(x_i) = \frac{1}{n}$, then we have

$$H(X) = \sum_{i=1}^{n} \frac{1}{n} \log_b(n) = \log_b(n) \leq l.$$

It follows that $n \leq b^l$. For $n \in \mathbb{N}$, $n - b^x = 0$ implies $x_0 = \log_b n$ and so $l = \lceil x_0 \rceil = \lceil \log_b n \rceil$.

Channel Capacity.

Definition 1 ([6])**.** The *capacity* C of a discrete channel is given by

$$C = \lim_{T \to \infty} \frac{\log N(T)}{T}$$

where $N(T)$ is the number of allowed signals of duration T.

Theorem 1 ([6])**.** Let $b_{ij}^{(s)}$ be the duration of the s^{th} symbol which is allowable in state i and leads to state j. Then the channel capacity C is equal to $\log W$, where W is the largest real root of the determinant equation:

$$\left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right| = 0$$

and where $\delta_{ij} = 1$ if $i = j$, and zero otherwise.

3.2 Multiset Encoding

We present some results related to the multiset encoding.

Encoding Length. We consider a set X of N symbols, an alphabet A, and the length of encoding l:

$$X = \{x_i = a_1^{n_1} a_2^{n_2} \dots a_b^{n_b} \mid \sum_{j=1}^{b} n_j = l,\ a_j \in A, i = \overline{1..N}\}.$$

Proposition 1. Non-uniform encodings of X over multisets are shorter than uniform encodings of X over multisets.

Proof. Over multisets we consider both uniform and non-uniform encodings in [3].

1. For an *uniform* encoding (where all the encoding representations have the same length l) we have $N \leq N(b,l) = \left\langle \begin{matrix} b \\ l \end{matrix} \right\rangle = \left(\begin{matrix} b+l-1 \\ l \end{matrix} \right) = \frac{(b+l-1)!}{l!(b-1)!} = \frac{\prod_{i=1}^{b-1}(l+i)}{(b-1)!}$. If x_0 is the real root of $N - \frac{\prod_{i=1}^{b-1}(x+i)}{(b-1)!} = 0$, then $l = \lceil x_0 \rceil$.

2. For a *non-uniform* encoding, $N \leq N(b+1, l-1) = \left\langle \begin{array}{c} b+1 \\ l-1 \end{array} \right\rangle =$

$$\left(\begin{array}{c} b+l-1 \\ l-1 \end{array} \right) = \frac{(b+l-1)!}{(l-1)!b!} = \frac{\prod_{i=0}^{b-1}(l+i)}{b!} = \frac{l}{b}\frac{\prod_{i=1}^{b-1}(l+i)}{(b-1)!} = \frac{l}{b}N(b,l).$$

Let x_0' be the real root of $N - \frac{\prod_{i=0}^{b-1}(x+i)}{(b-1)!} = 0$. Then $l' = \lceil x_0' \rceil$.

From $N - N(b, x_0) = 0$ and $N - \frac{x_0'}{b}N(b, x_0') = 0$ we get $N(b, x_0) = \frac{x_0'}{b}N(b, x_0')$. In order to prove $l > l' \iff x_0 > x_0'$, let suppose that $x_0 \leq x_0'$. We have $x_0' > b$ (for sufficiently large numbers), and this implies that $N(b, x_0) \leq N(b, x_0') < \frac{x_0'}{b}N(b, x_0')$. Since this is false, it follows that $x_0 > x_0'$ implies $l \geq l'$. □

Channel Capacity. We consider that a sequence of multisets is transmitted along the channel. The capacity of such a channel is computed for base 4, then some properties of it for any base are presented.

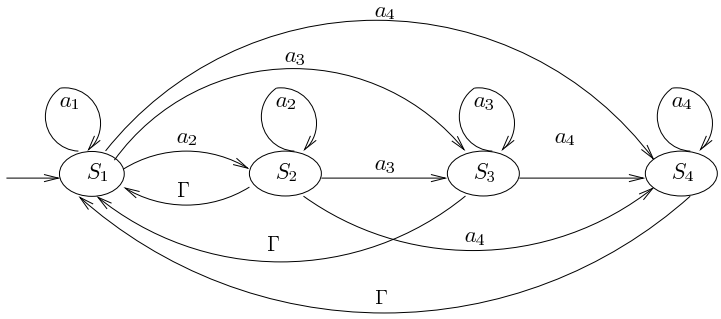

Figure 4. Multiset channel capacity

Multiset channel capacity in base 4. In Figure 4 we have a graph $G(V, E)$ with 4 vertices $V = \{S_1, S_2, S_3, S_4\}$ and $E = \{(i, j) \mid i, j = \overline{1..4},\ i \leq j\} \cup \{(i, j) \mid i = 4,\ j = \overline{1..3}\}$

Using the notation of Theorem 1, we have $b_{ij}^{(a_k)} = t_k$ because we consider that the duration to produce a_k is the same for each $(i, j) \in E$. The determinant equation is

$$\begin{vmatrix} W^{-t_1}-1 & W^{-t_2} & W^{-t_3} & W^{-t_4} \\ 0 & W^{-t_2}-1 & W^{-t_3} & W^{-t_4} \\ 0 & 0 & W^{-t_3}-1 & W^{-t_4} \\ 0 & 0 & 0 & W^{-t_4}-1 \end{vmatrix} = 0.$$

If we consider $t_k = t$, then the equation becomes $\left(1 - \frac{1}{W^t}\right)^4 = 0$, and $W_{real} = 1$. Therefore $C = \log_4 1 = 0$.

Multiset Channel Capacity in Base b.

Theorem 2. The multiset channel capacity is zero, i.e., $C = 0$.

Proof.

First approach. The first method for computing the capacity is using the definition from [6].

$$\begin{aligned} C &= \lim_{T\to\infty} \frac{\log N(T)}{T} = \lim_{T\to\infty} \frac{\log N(b,T)}{T} \\ &= \lim_{T\to\infty} \frac{\log \left\langle \begin{array}{c} b \\ T \end{array} \right\rangle}{T} = \lim_{T\to\infty} \frac{1}{T} \log \frac{(b+T-1)!}{T!(b-1)!} \end{aligned}$$

Using Stirling's approximation

$$\log n! \approx n \log n - n$$

we obtain

$$\begin{aligned} C &= \lim_{T\to\infty} \frac{1}{T} \left(\log(b+T-1)! - \log T! - \log(b-1)!\right) \\ &= \lim_{T\to\infty} \frac{1}{T} \left((b+T-1)\log(b+T-1) - T\log T - (b-1)\log(b-1)\right) \\ &= \lim_{T\to\infty} \frac{b-1}{T} \log\left(1 + \frac{T}{b-1}\right) + \lim_{T\to\infty} \log\left(1 + \frac{b-1}{T}\right) \\ &\quad - \lim_{T\to\infty} \frac{(b-1)\log(b-1)}{T} = 0 \end{aligned}$$

Second approach. Using Theorem 1, the determinant equation for a multiset encoder is:

$$\begin{vmatrix} W^{-t_1}-1 & W^{-t_2} & W^{-t_3} & \cdots & W^{-t_b} \\ 0 & W^{-t_2}-1 & W^{-t_3} & \cdots & W^{-t_b} \\ 0 & 0 & W^{-t_3}-1 & \cdots & W^{-t_b} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & W^{-t_{b-1}}-1 & W^{-t_b} \\ 0 & 0 & 0 & \cdots & W^{-t_b}-1 \end{vmatrix} = 0.$$

Claim 1. If $t_k = t$, then the determinant equation becomes

$$\left(1 - \frac{1}{W^t}\right)^b = 0. \tag{3}$$

The capacity C is given by $C = \log_b W$, where W is the largest real root of the equation (3). Considering $x = W^{-t}$, then we have

$$W = \frac{1}{\sqrt[t]{x}} \Rightarrow C = -\tfrac{1}{t}\log_b x. \tag{4}$$

Since we need the largest real root W, then we should find the smallest positive root x of the equation $(1-x)^b = 0$ which is $x = 1$, and so $C = 0$. □

4 CONCLUSION

Based on Shannon's classical work, we derive a formula for the information content of a multiset. Using the definition and the determinant capacity formula, we compute the multiset channel capacity. As future work we plan to further explore the properties of multiset-based communication systems, and compare these to similar results for string-based communication systems.

Acknowledgements

This work has been partially supported by the research grants CEEX 47/2005 and CNCSIS TD-24/2007.

References

[1] Atanasiu, A.: Arithmetic with Membranes. Workshop on Multiset Processing, Curtea de Argeş, pp. 1–17, 2000.

[2] Bonchiş, C.—Ciobanu, G.—Izbaşa, C.: Encodings and Arithmetic Operations in Membrane Computing. Theory and Applications of Models of Computation, Lecture Notes in Computer Science Vol. 3959, pp. 618–627, Springer, 2006.

[3] Bonchiş, C.—Ciobanu, G.—Izbaşa C.: Number Encodings and Arithmetics over Multisets. SYNASC'06, pp. 354–361, IEEE Computer Society, 2006.

[4] MacKay, D.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, 2003.

[5] Păun, Gh.: Membrane Computing. An Introduction. Springer, 2002.

[6] Shannon, C. E.: A Mathematical Theory of Communication. Bell System Technical Journal, Vol. 27, 1948, pp. 379–423 and pp. 623–656.

[7] Varshney, L. R.—Goyal, V. K.: Toward a Source Coding Theory for Sets. Proceedings of the Data Compression Conference, 2006.

Cosmin Bonchiş is a junior researcher affiliated at Research Institute "eAustria" in Timişoara, and a Ph. D. student at the West University of Timişoara. He is interested in membrane computing and distributed systems.

Cornel Izbaşa is a researcher affiliated at Research Institute "eAustria" in Timişoara, and a Ph. D. student at the West University of Timişoara. He is interested in distributed systems and membrane computing.

Gabriel Ciobanu is a Professor at "A. I. Cuza" University of Iasi, and a senior researcher at the Institute of Computer Science of the Romanian Academy. He has wide-ranging interests in computing including distributed systems and concurrency, computational methods in biology, membrane computing, and theory of programming (semantics, formal methods, logics, verification). He has published over 100 papers in computer science and mathematics, and several volumes. His webpage is `http://www.info.uaic.ro/~gabriel`.