Information Theory over Multisets*

Cosmin Bonchiş[1], Cornel Izbaşa[1], and Gabriel Ciobanu[2]

[1] Research Institute "e-Austria" Timişoara, Romania
cosmin@ieat.ro, cornel@ieat.ro
[2] Romanian Academy, Institute of Computer Science
Blvd. Carol I nr.8, 700505 Iaşi
gabriel@iit.tuiasi.ro

Summary. Starting from Shannon theory of information, we present the case of producing information in the form of multisets, and encoding information using multisets. We rewiew the entropy rate of a multiset information source and we derive a formula for the information content of a multiset. We then 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 are plenty of 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 process 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.

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 that system? Also, many examples of P systems perform various computational tasks. Authors of such systems encode the input (usually numbers) in various ways, some by superimposing a string-like structure on the membrane system [1], some by using the natural encoding or unary numeral system, that is, the natural number n is represented with n objects, for example,

* This work has been partially supported by the research grant CEEX 47/2005

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 for sure an energy and information exchange. How to describe it? Another, more general way to pose that question is: what are the natural ways to encode numbers, and more generally, 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. According to our knowledge, this is the first attempt of such an investigation. We start from the idea that a study of multiset information theory might produce interesting, 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, because they constitute a poorer encoding media compared to strings. Encoding information on strings or even richer, more organized and complex structures are obviously possible and have 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 valid purposes. We focus mainly on such "natural" approaches to information encoding over multisets, and present some advantages they have over approaches that 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.

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 day-to-day more 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 $p_{i\in\overline{1..n}}$ 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 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, and a multiset is a string equivalence class. The entropy rate of such a source is proved to be zero in [7]:

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

Information content

The *information content* of an outcome (multiset) x is $h(x) = \log \frac{1}{P(x)}$. See [4].

Let be $k \in \mathbb{N}$ 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 $\sum_{i=1}^n m_i = k$, then the probability of the outcome x is given by the multinomial distribution $\binom{k}{m_1, m_2, \dots, m_n} \prod_{i=1}^n p_i^{m_i}$:

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

So, the information content of the multiset x is:

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

$$= \log \frac{\prod_{i=1}^n m_i!}{(\sum_{i=1}^n m_i)! \prod_{i=1}^n p_i^{m_i}}$$

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 also on the capacity theorem. Please note that one can have a multiset information source and a usual sequence-based encoder and channel. All the following combinations are possible:

3.1 String Encoding

We shortly review the results concerning the string encoding.

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

Table 1. Source/Encoder types

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:*

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

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, therefore:

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

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

Proof. Over multisets we have

1. for an *uniform* (all the encoding representation have the same length l) encoding: $N \leq N(b,l) = \left\langle {b \atop l} \right\rangle = \binom{b+l-1}{l} = \frac{(b+l-1)!}{l!(b-1)!} = \frac{\prod_{i=1}^{b-1}(l+i)}{(b-1)!}$. 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 *non-uniform* encoding: $N \leq N(b+1, l-1) = \left\langle {b+1 \atop l-1} \right\rangle = \binom{b+l-1}{l-1} = \frac{(b+l-1)!}{(l-1)!b!} = \frac{\prod_{i=0}^{b-1}(l+i)}{b!} = \frac{l}{b}\frac{\prod_{i=1}^{b-1}(l+i)}{(b-1)!} = \frac{l}{b}N(b,l)$. 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.

Multiset channel capacity in base 4

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

In Theorem 1 we get $b_{ij}^{(a_k)} = t_k$ 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 $C = \log_4 1 = 0$.

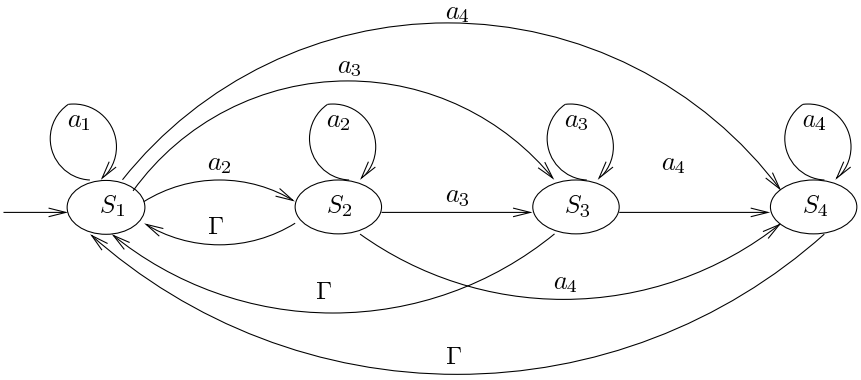

Fig. 1. Multiset channel capacity

Multiset channel capacity in base b

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

Proof. ***First approach***

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

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

Using Stirling's approximation $\log n! \approx n \log n - n$ we obtain

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

Second approach

Using 1, the determinant equation for a multiset encoder is:

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

Proposition 2. *If $t_k = t$, then the determinant equation becomes*

$$\left(1 - \frac{1}{W^t}\right)^b = 0. \tag{2}$$

The capacity C is given by $C = \log_b W$, where W is the largest real root of the equation (2). Considering $x = W^{-t}$, then we have

$$W = \frac{1}{\sqrt[t]{x}} \Rightarrow C = -\frac{1}{t}\log_b x. \tag{3}$$

Since we need the largest real root W then we should find the smallest positive root x of the equation $(1-x)^b = 0 \Rightarrow x = 1 \Rightarrow C = 0$.

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.

References

1. A. Atanasiu. Arithmetic with Membranes, *Pre-Proceedings of the Workshop on Multiset Processing*, Curtea de Argeş, pp.1-17, 2000.
2. C. Bonchiş, G. Ciobanu, C. Izbaşa. Encodings and Arithmetic Operations in Membrane Computing. *Theory and Applications of Models of Computation*, Lecture Notes in Computer Science vol.3959, Springer, pp.618–627, 2006.
3. C. Bonchiş, G. Ciobanu, C. Izbaşa. Number Encodings and Arithmetics over Multisets. *SYNASC'06: 8th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, IEEE Computer Society, pp.354-361, 2006.
4. D. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, Cambridge, England 2003.
5. Gh. Păun. *Membrane Computing. An Introduction*. Springer, 2002.
6. C.E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal* vol.27, pp.379-423 and pp.623-656, 1948.
7. L. R. Varshney, V. K. Goyal, *Toward a Source Coding Theory for Sets*, in Proceedings of the Data Compression Conference (DCC 2006), Snowbird, Utah, 28-30 March 2006.