Journal of Statistical Physics, Vol. 12, No. 3, 1975

Transmission of Information in Nucleotide Pools: Application of the Statistical Thermodynamic Formalism

Alexandr Křemen[1]

Received July 8, 1974

Starting from information theory ideas, the probability distribution of nucleotide molecules in a pool is derived, using the Kullback information measure. A statistical thermodynamic formalism leads to analogs of thermodynamic functions like entropy and Helmholtz free energy, and to equations describing their changes. If information transmission is a maximum, these analogs have certain interesting properties. The general case is investigated, when both the actual and the prior distributions change.

KEY WORDS: Nucleotide pools; information theory; Kullback information measure; maximum information transmission; statistical thermodynamic formalism; change of prior distribution.

1. INTRODUCTION

The creation and maintenance of organization in living matter are processes generally requiring the supply of both energy and information. For many reasons, including historical ones, the energetic aspect has drawn much more attention. While it is true that for some processes the informational aspect is

[1] Institute of Microbiology, Czechoslovak Academy of Sciences, Prague, Czechoslovakia.

less important, there are, however, certain very important processes where it should not be neglected. The role of nucleotide pools (the explanation of this concept is given later in this section) in energy utilization in living systems and the use of nucleotides as building units for certain important molecules such as nucleic acids and for some other structures in the cell make these pools objects of primary interest from both the energetic and informational points of view. In this paper, an attempt is made to describe these pools in terms of a statistical thermodynamic formalism based on information theory ideas. The description applies to more general systems of this kind as well.

The use of a statistical thermodynamic formalism in the solution of information theory problems is due to Reiss and Huang.[1] In contrast to their work, Kullback's information measure is used here rather than Shannon's. This is necessary in applications to living systems, as will be shown in the next section. The applications are discussed elsewhere.[2]

For readers not familiar with the concept of a nucleotide pool, an explanation is given here. Consider a system of particles (molecules), each of which can occupy any of a finite number s of energy levels. The number of particles occupying any level is not limited. Transitions between the levels result from some kind of interaction, as, for example, chemical reactions. The system is open, since the particles may be added to or withdrawn from the system, generally regardless of which level is involved. In nucleotide pools, $s = 3$, and the particles occupying the three levels are called (in the order of increasing energy) monophosphates, diphosphates, and triphosphates. There are several pools working in living matter, for example, the adenylates, guanylates, etc. Of these, the adenylate pool is the most universal, the other pools having more or less specialized functions. Briefly, they all serve as media for storage and distribution of energy, and supply molecules for synthesis of certain important, more complicated molecules. The population of the energy levels is usually far from thermodynamic equilibrium. The actual state of any pool depends on the physiological conditions of the living system of interest.

2. KULLBACK INFORMATION MEASURE

In this section, we give arguments in support of the use of the Kullback information measure. For any pool, the population of the energy levels can be expressed as a probability distribution $\{p_i\}$, which in turn represents some degree of organization. There are two elements of asymmetry in the time development of this organization. First, the least organized state of the pool is achieved at thermodynamic equilibrium, at which, of course, the probability distribution $\{p_i^{(0)}\}$ may be different from the set $\{1/3, 1/3, 1/3\}$ for which the Shannon entropy is maximum. Second, the most organized state is, due to the

energy differences between the levels, that corresponding to all nucleotide molecules excited to the highest—triphosphate—level. The corresponding probability distribution is $\{p_i^{(m)}\} = \{0, 0, 1\}$. If the role of the pools in living matter is taken into account, it is reasonable to require that no information be conveyed by a pool in thermodynamic equilibrium and maximum information be conveyed in the most excited state. These two elements of asymmetry lead to the use of the Kullback information measure, the amount of conveyed information being identical with the information gain in the usual terminology. The corresponding formulas are[3]

$$I(p; p^{(0)}) = \kappa \sum_i p_i \log(p_i/p_i^{0)}) \tag{1}$$

for the information gain, and

$$U = I(p^{(m)}; p^{(0)}) - I(p; p^{(0)}) \tag{2}$$

for the entropy or uncertainty. In the following, $\{p^{(0)}\}$ will be called the prior distribution, in agreement with common usage, although a name like reference distribution would be perhaps more appropriate here.

3. PROBABILITY DISTRIBUTION

The concept of information transmission is a basic one in the present paper. For this reason, the derivation of the probability distribution in the pool starts from the concepts of a message, a source, and a channel, in relation to the role of nucleotide pools. The line of reasoning resembles very closely that of the work by Reiss and Huang.[1]

Messages are defined as sequences of symbols, which, in this case, are molecules of the pool. Subsequent symbols in a message may be correlated; the messages, on the contrary, are supposed long enough to be substantially uncorrelated. The pool is thus considered as a source emitting messages, the kth message with probability P_k. The processes in which the molecules of the pool take part comprise a channel (or a set of parallel channels), the properties of which may be identified with a coding procedure. For the kth message they define a characteristic quantity T_k (or a set of quantities T_k, W_k, ...) and a "channel" probability P_k'. In the following, only one characteristic quantity T_k will be considered and interpreted as the transmission time of the kth message, in agreement with the work cited.[1] In principle, the characteristic quantities may have certain other physical dimensions as well. The probabilities P_k and P_k' are generally different; however, the laws of chemical kinetics, of transport, etc., provide for mutual matching of the source and the channel, so that $P_k = P_k'$, and the source works most efficiently in the sense

that the most typical message composed by the source is also the most typical message composed by the channel.

If source and channel are matched, the probability distribution of the messages can be found by maximizing the entropy (2) subject to certain constraints.[3] These are two in this case: A given quantity $\mathscr{T}$ must be composed of an integer number of the T_k's, that is,

$$\mathscr{T} = \sum_k N_k T_k$$

where N_k is the number of messages of the kth type. Another constraint, which turns out to limit information transmission, requires that the total number of messages $\mathscr{N}$ fitting into $\mathscr{T}$ be fixed,

$$\mathscr{N} = \sum_k N_k$$

Without this constraint, an equal or greater number of messages fits into $\mathscr{T}$. The absence of this constraint therefore corresponds to maximum transmission of information. Thus one looks for the extremum of the expression

$$\phi = \sum_k N_k \Bigg[I_M(m;0) - \kappa \sum_n \frac{N_n}{\sum_k N_k} \log\left(\frac{N_n}{\sum_k N_k} \Big/ P_0^{(n)} \right) - \frac{1}{\kappa\tau} \left(\sum_n \frac{N_n T_n}{\sum_k N_k} - \frac{\mathscr{T}}{\sum_k N_k} \right) + (\Omega + \kappa) \left(\sum_n \frac{N_n}{\sum_k N_k} - \frac{\mathscr{N}}{\sum_k N_k} \right) \Bigg]$$

The last term $(\Omega + \kappa)(...)$ is missing if $\sum_k N_k$ is not fixed. The probabilities P_n were identified with the frequencies,

$$P_n = N_n \Big/ \sum_k N_k$$

The coefficients $(\Omega + \kappa)$ and $1/\kappa\tau$ are the undetermined multipliers; $I_M(m;0)$ denotes the maximum information gain in case of the message distribution. The condition for an extremum is

$$\partial\phi/\partial N_i = 0$$

This is equivalent to

$$\partial\phi/\partial P_i = 0$$

if $\mathscr{N}$ is fixed. If this constraint is lifted,

$$\partial P_n/\partial N_i = (\delta_{in} - P_n) \Big/ \sum_k N_k$$

where

$$\begin{aligned} \delta_{in} &= 1 \quad \text{if} \quad i = n \\ &= 0 \quad \text{otherwise} \end{aligned}$$

Then we get

$$P_k = \frac{P_k^{(0)} \exp(-T_k/\kappa\tau)}{Q}, \qquad Q = \sum_k P_k^{(0)} \exp - \frac{T_k}{\kappa\tau} \tag{3}$$

if $\mathcal{N}$ is fixed, and

$$Q^* = \exp[-I_M(m; 0)/\kappa] \tag{4}$$

if $\sum_k N_k$ is not fixed; it is now seen that this condition determines maximum information transmission at given τ.

Now let all messages be composed of the same number n of symbols, and let the mean characteristic quantities t_i of the ith symbol be the same regardless of the message; then

$$n = \sum_i n_{ik}, \qquad T_k = \sum_i n_{ik} t_i$$

where n_{ik} is the number of times the ith symbol appears in the kth message. In this approximation the symbols are effectively uncorrelated (the source is without memory) and their probabilities can be written as

$$p_i = p_i^{(0)} \exp(-t_i/\kappa\tau)/q, \qquad q = \sum_i p_i^{(0)} \exp(-t_i/\kappa\tau) \tag{5}$$

if $\mathcal{N}$ is fixed, and

$$q^* = p_3^{(0)} \tag{6}$$

if $\sum N_k$ is not fixed [since, for nucleotide pools, only one message consisting exclusively of triphosphates contributes to $I_M(m; 0)$,

$$I_M(m; 0) = -\kappa \log(p_3^{(0)})^n$$

and (6) results by (4) and $Q = q^n$].

4. ANALOGS OF THERMODYNAMIC FUNCTIONS

The formulas (3) and (5) resemble those known from statistical thermodynamics (except for the use of the prior distribution). This is not the only reason analogs of some thermodynamic functions are derived here. These analogs have interesting properties at maximum information transmission. In the following, advantage is taken of the use of Kullback information measure and the general case is treated, when both the actual distribution $\{p_i\}$ and the prior distribution $\{p_i^{(0)}\}$ change. Sporulation (a change from an open to a closed system) can be given as an example for which such a treatment can be useful.

With regard to notation, a symbol in angular brackets denotes the mean value of the corresponding quantity, as, for example,

$$\langle t \rangle = \langle t_i \rangle = \sum_i p_i t_i$$

An asterisk refers to the condition of maximum transmission of information, as in (6) or (4). The symbol $\langle t \rangle'$ denotes the time derivative of $\langle t \rangle$, whereas $\langle t' \rangle = \sum p_i t_i'$ is the mean value of time derivatives of the quantities t_i.

By (2) and (5) or (6), the entropy is

$$U = I(m; 0) + (\langle t \rangle / \tau) + \kappa \log q, \qquad U^* = \langle t \rangle / \tau \tag{7}$$

$I(m; 0)$ is the maximum information gain in case of the probability distribution of the symbols, i.e., in the pool.

The analog of the Helmholtz free energy is

$$F = \langle t \rangle - \tau U = -\tau[I(m; 0) + \kappa \log q] = -\kappa\tau(\log q - \log p_3^{(0)}) \tag{8}$$
$$F^* = 0$$

It is interesting to note that if more characteristic quantities, say t_i, w_i, ..., determine the distribution (together with corresponding τ_t, τ_w, ...), then only in case of maximum information transmission is a simple definition of F possible, and both U and F split into separate parts, each one for one of the variables, as

$$U_t^* = \langle t \rangle / \tau_t, \qquad U_w^* = \langle w \rangle / \tau_w, \ldots$$
$$F_t^* = \langle t \rangle - \tau_t U_t^* = 0, \qquad F_w^* = \langle w \rangle - \tau_w U_w^* = 0, \ldots$$

Considering time changes, it is convenient to derive the relations using the function F, since $F^* = 0$. The function F is a function of the variables q, $p_3^{(0)}$, and τ; q is a function of $p_2^{(0)}$, $p_3^{(0)}$, t_1, t_2, t_3, τ; and each t_i is generally a function of p_2, p_3, $p_2^{(0)}$, $p_3^{(0)}$, τ, and, in addition, of some external parameters x_j. Then

$$F' = \frac{\partial F}{\partial q}\left[\sum_i \frac{\partial q}{\partial t_i}\left(\frac{\partial t_i}{\partial p_2} p_2{}' + \frac{\partial t_i}{\partial p_3} p_3{}' + \frac{\partial t_i}{\partial \tau}\tau' + \sum_j \frac{\partial t_i}{\partial x_j} x_j{}'\right.\right.$$
$$\left.\left. + \frac{\partial t_i}{\partial p_2^{(0)}} p_2^{(0)\prime} + \frac{\partial t_i}{\partial p_3^{(0)}} p_3^{(0)\prime}\right) + \frac{\partial q}{\partial \tau}\tau' + \frac{\partial q}{\partial p_2^{(0)}} p_2^{(0)\prime} + \frac{\partial q}{\partial p_3^{(0)}} p_3^{(0)\prime}\right]$$
$$+ \frac{\partial F}{\partial \tau}\tau' + \frac{\partial F}{\partial p_3^{(0)}} p_3^{(0)\prime} \tag{9}$$

We define analogs of thermodynamic chemical potentials

$$\mu_k - \mu_1 = -\kappa\tau \frac{\partial}{\partial p_k} \log q = \sum p_i \frac{\partial t_i}{\partial p_k} = \left\langle \frac{\partial t}{\partial p_k} \right\rangle, \qquad k = 2, 3 \tag{10}$$

$$\mu_k^{(0)} - \mu_1^{(0)} = \sum p_i \frac{\partial t_i}{\partial p_k^{(0)}} = \left\langle \frac{\partial t}{\partial p_k^{(0)}} \right\rangle, \qquad k = 2, 3 \tag{11}$$

and analogs of pressure

$$\pi_j = \kappa\tau \frac{\partial}{\partial x_j} \log q = -\sum_i p_i \frac{\partial t_i}{\partial x_j} = -\left\langle \frac{\partial t}{\partial x_j} \right\rangle \tag{12}$$

Noting that

$$\frac{\partial F}{\partial q}\left(\sum_i \frac{\partial q}{\partial t_i}\frac{\partial t_i}{\partial \tau} + \frac{\partial q}{\partial \tau}\right) + \frac{\partial F}{\partial \tau}$$

$$= -\kappa(\log q - \log p_3^{(0)}) - \kappa\tau \frac{d}{d\tau} \log q$$

$$= \frac{F}{\tau} + \tau \sum p_i \frac{\partial}{\partial \tau}\left(\frac{t_i}{\tau}\right) = \left\langle \frac{\partial t}{\partial \tau} \right\rangle - U$$

we see that the relation (9) reads

$$F' = (\mu_2 - \mu_1)p_2' + (\mu_3 - \mu_1)p_3' + \left(\left\langle \frac{\partial t}{\partial \tau} \right\rangle - U\right)\tau'$$
$$- \sum_j \pi_j x_j' + (\mu_2^{(0)} - \mu_1^{(0)})p_2^{(0)\prime} + (\mu_3^{(0)} - \mu_1^{(0)})p_3^{(0)\prime}$$
$$- \kappa\tau\left[\left(\frac{p_2}{p_2^{(0)}} - \frac{p_1}{p_1^{(0)}}\right)p_2^{(0)\prime} - \left(\frac{1 - p_3}{p_3^{(0)}} + \frac{p_1}{p_1^{(0)}}\right)p_3^{(0)\prime}\right] \tag{13}$$

If $F' = 0$,

$$(\mu_2 - \mu_1)p_2' + (\mu_3 - \mu_1)p_3' + \left(\left\langle \frac{\partial t}{\partial \tau} \right\rangle - U\right)\tau'$$
$$- \sum_j \pi_j x_j' + (\mu_2^{(0)} - \mu_1^{(0)})p_2^{(0)\prime} + (\mu_3^{(0)} - \mu_1^{(0)})p_3^{(0)\prime}$$
$$- \kappa\tau\left[\left(\frac{p_2}{p_2^{(0)}} - \frac{p_1}{p_1^{(0)}}\right)p_2^{(0)\prime} - \left(\frac{1 - p_3}{p_3^{(0)}} + \frac{p_1}{p_1^{(0)}}\right)p_3^{(0)\prime}\right] = 0 \tag{14}$$

This equation describes time changes of the variables if they are mutually related. If the prior distribution changes independently, (14) splits into two relations

$$(\mu_2 - \mu_1)p_2' + (\mu_3 - \mu_1)p_3' + \left(\left\langle \frac{\partial t}{\partial \tau} \right\rangle - U\right)\tau' - \sum_j \pi_j x_j' = 0$$
$$(\mu_2^{(0)} - \mu_1^{(0)})p_2^{(0)\prime} + (\mu_3^{(0)} - \mu_1^{(0)})p_3^{(0)\prime} \tag{15}$$
$$- \kappa\tau\left[\left(\frac{p_2}{p_2^{(0)}} - \frac{p_1}{p_1^{(0)}}\right)p_2^{(0)\prime} - \left(\frac{1 - p_3}{p_3^{(0)}} + \frac{p_1}{p_1^{(0)}}\right)p_3^{(0)\prime}\right] = 0$$

If the prior distribution changes follow the condition

$$\frac{p_2^{(0)\prime}}{p_3^{(0)\prime}} = \frac{dp_2^{(0)}}{dp_3^{(0)}} = \frac{[(1 - p_3)/p_3^{(0)}] + (p_1/p_1^{(0)})}{(p_2/p_2^{(0)}) - (p_1/p_1^{(0)})} \tag{16}$$

then also

$$\frac{dp_2^{(0)}}{dp_3^{(0)}} = -\frac{\mu_3^{(0)} - \mu_1^{(0)}}{\mu_2^{(0)} - \mu_1^{(0)}} \tag{17}$$

The first equation (15) relates the changes of the actual distribution to changes of τ and of the parameters x_j. If all variables were independent, (15) would imply a steady state.

From (7), by direct computation,

$$\begin{aligned} U' &= \kappa(\log q - \log p_3^{(0)})' + \frac{\langle t\rangle'}{\tau} - \frac{\langle t\rangle}{\tau}\frac{\tau'}{\tau} \\ &= \frac{F}{\tau^2}\tau' - \frac{F'}{\tau} + \frac{\langle t\rangle'}{\tau} - \frac{\langle t\rangle}{\tau}\frac{\tau'}{\tau} \\ &= \kappa\left(\left\langle \frac{p^{(0)\prime}}{p^{(0)}}\right\rangle - \frac{p_3^{(0)\prime}}{p_3^{(0)}}\right) + \frac{\langle t\rangle'}{\tau} - \frac{\langle t'\rangle}{\tau} \\ &= u_0' + \frac{\langle t\rangle'}{\tau} - \frac{\langle t'\rangle}{\tau} \end{aligned} \tag{18}$$

since

$$\begin{aligned} \kappa(\log q)' = \kappa\frac{q'}{q} &= \kappa\frac{\sum p_i^{(0)\prime} e^{-t_i/\kappa\tau}}{q} - \kappa\sum p_i\left(\frac{t_i}{\kappa\tau}\right)' \\ &= \kappa\left\langle \frac{p^{(0)\prime}}{p^{(0)}}\right\rangle - \frac{\langle t'\rangle}{\tau} + \frac{\langle t\rangle}{\tau}\frac{\tau'}{\tau} \end{aligned}$$

In (18),

$$u_0' = \kappa\left[\left(\frac{p_2}{p_2^{(0)}} - \frac{p_1}{p_1^{(0)}}\right)p_2^{(0)\prime} - \left(\frac{1 - p_3}{p_3^{(0)}} + \frac{p_1}{p_1^{(0)}}\right)p_3^{(0)\prime}\right]$$

and (16) is a condition for $u_0' = 0$. With this condition,

$$\tau U' = \langle t\rangle' - \langle t'\rangle \tag{19}$$

Comparison with the equation of thermodynamics representing the combined first and second laws shows that τ is an analog of absolute temperature,[1] $\langle t\rangle$ is an analog of internal energy, and $-\langle dt\rangle$ is an analog of infinitesimal work.

If information transmission is maximum, $F = 0$ and, by (18),

$$\tau U^{*\prime} = \langle t\rangle' - (\langle t\rangle/\tau)\tau' - F' \tag{20}$$

Combining (19) and (20), we obtain

$$\langle t'\rangle = (\langle t\rangle/\tau)\tau' = U^*\tau' + F' \tag{21}$$

Relation (21) is equivalent to

$$\langle (t/\kappa\tau)' \rangle = F'/\kappa\tau$$

Finally, (13) can be written as an analog of the Gibbs equation

$$\langle t \rangle' = \tau U^{*\prime} + \left\langle \frac{\partial t}{\partial \tau} \right\rangle \tau' - \sum_j \pi_j x_j' + (\mu_2 - \mu_1) p_2' + (\mu_3 - \mu_1) p_3' \qquad (22)$$

The main results can be summarized as follows.

1. At maximum information transmission, the analogs of thermodynamic entropy and Helmholtz free energy are, respectively,

$$U = \langle t \rangle / \tau \qquad \text{and} \qquad F = 0$$

2. If the prior distribution changes in such a way that

$$p_3^{(0)\prime}/p_3^{(0)} = \langle p^{(0)\prime}/p^{(0)} \rangle$$

then

$$\tau U' = \langle t \rangle' - \langle t' \rangle$$

is an exact analog of the combined first and second laws of thermodynamics.

3. An analog of the Gibbs equation holds for $\langle t \rangle'$. If the prior distribution changes independently, this analog has the form (22).

5. CONCLUDING REMARKS

It seems reasonable to assume that maximum information transmission represents a favorable condition for living systems. The derivations in the preceding sections were therefore carried out mainly with regard to this condition.

The approximation of a memoryless source renders this treatment applicable to ideal systems only. Here it means that the interactions of the source with the channel are weak enough for the probability distribution to be adjusted to only averaged requirements of the living system. "Averaged" means both with respect to time (the presumed absence of correlation among messages was virtually transferred to individual symbols) and with respect to the various kinds of utilization of the nucleotide pool. The limitation to an ideal system does not make the treatment useless, however. Important and useful results have been obtained by investigating ideal systems in statistical thermodynamics. It is hoped that the analogy with statistical thermodynamics observed in the foregoing sections will extend the applicability of the results presented here.

REFERENCES

1. H. Reiss and C. Huang, *J. Stat. Phys.* **3**:191 (1971).
2. A. Křemen, submitted to *Studia Biophysica.*
3. A. Hobson and Bin-Kang Cheng, *J. Stat. Phys.* **7**:301 (1973).