BIOLOGIJA. 2004. Nr. 4. P. 30–32

# Bioinformation and bioinformatics otherwise: on estimation of information quantity

**D. Kirvelis**

** Vilnius University, Department of Biochemistry and Biophysics, M. K. Èiurlionio 21, LT-03101 Vilnius, Lithuania*

The present study is an attempt to scan through bioinformation procedures in living beings and biosystems from the point of view of information theory, informatics and to discuss the units and methods of quantitative estimation of bioinformation. Using methods of logic analysis and synthesis, an organized functional scheme of animal was composed. It has been used to show that the essence of the living being is the three-level informational flows for coding-decoding and control, implemented by genetic, hormonal (pheromonal) and neural subsystems. It is suggested to apply information theory for studies of these bioinformational flows and use subsystem-specific logarithmic units of information quantity, GITs, PROTs and others, adjusted to the particular information carriers (signals). The quantitative method for investigation of a biosystem in terms of information theory and control theory expands the conception of bioinformatics.

**Key words**: bioinformation, bioinformatics, coding, decoding, bioinformation units

## INTRODUCTION

Bioinformatics is a new branch of rapidly developing biology and biotechnology. In European and Lithuanian Sciences Classification, it is attached to General Biomedical Sciences B 001 as branch B 110 [1, 2]. Bioinformatics is commonly taken for making use of informatics technologies in genomics, proteomics, and data mining. Workshops on bioinformatics held in conjunction with the 13th International Symposium on Methodologies for Intelligent Systems (ISMIS 2002, Lyon, France, 2002) and “Bioinformatics: from inference to predictive models” in Oxford, UK (Gordon research conference, 2003) exemplify the expanded horizon of bioinformatics [3]. Analysis of metabolic systems and regulation, systems biology, neural networks, fuzzy logic, knowledge modeling, and even biosemiotics are embraced by bioinformatics. It becomes necessary to change the common conceptions of bioinformation and bioinformatics [4, 5].

Informatics is a science of informational procedures taking place in computers, communication systems and automats-robots. The point is that bioinformatics could be taken for a science that uses quantitative information units for evaluation and interpretation of informational (bioinformation) procedures taking place in biological systems and living beings. Half a century ago, such a view was called biocybernetics.

Same as physicists seeking for the understanding of nature were “forced” to invent the concept of energy and find methods how to measure it, now specialists in cybernetics and informatics develop the concept of information and create methods how to estimate and measure it since it is generally accepted that quantitatively evaluated information enables the human being to understand himself, society, living nature and produce automats-robots as well as effective informational technologies of control.

The present study is an attempt to scan through bioinformation procedures in living beings and biosystems from the point of view of information theory, informatics and to discuss the units and methods of quantitative estimation of bioinformation. Attention was focused on the information concept, function and estimation.

## METHODS

Methods of logic analysis, deductive synthesis and mathematic analogies were used in the present study. A comparative analysis of informational coding, transcoding and decoding procedures in enginee-

* Address for correspondence. E-mail: dobilas.kirvelis@gf.vu.lt

ring systems, on the one hand, and biotechnological procedures in living systems on the other was carried out.

## RESULTS AND DISCUSSION

### Functional organization of the living beings

A comparison between computer technologies and biotechnologies taking place in living beings resulted in a scheme of the functional structure of animal (Figure) [6]. The scheme was the basis for the bioinformation estimations presented in this study.

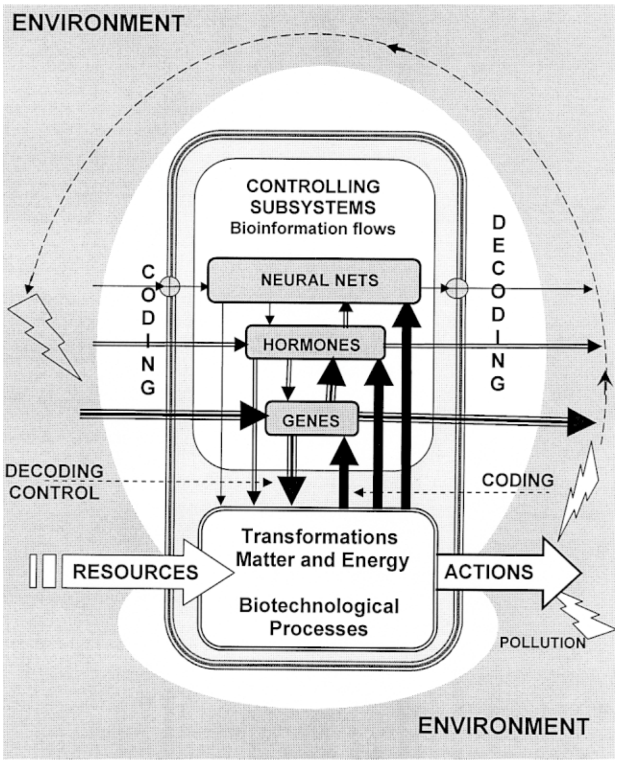

**Figure.** Functional diagram of animal as a hierarchical organized bioinformational system functioning on the principle of closed-loop coding-decoding

A living being as an organized system includes two functionally different subsystems: the controlling one (a controller), which processes the information, and the controlled one, which carries out transformations of matter and energy, *i. e.* biotechnologies for goal-oriented actions. The controller of cell consists of nucleic acids, the controller of a multicellular organism is augmented by hormonal regulation, and the animal's controller has a neural system in addition. In animal, this three-level structure of control linked by internal and external feedbacks to the environment, forms a hierarchically organized closed-loop coding-decoding system. Coding-decoding processes are essential for reproduction of organisms, since the stability of functional structures can be ensured by multiplication of discretely coded genetic projects of the organism [7]. Biotechnology of reproduction becomes a rather steady bioinformation technology.

Many phenomena of living nature could be explained in the best way by using the terms of bioinformation technologies.

### Estimation and measure of information

Information, according to C. E. Shannon is a removed or reduced uncertainty. Computer engineers made their devices from the elements of steady binary states, and for estimation of a potential of the device (the number of the possible states) they introduced a unit to measure the information quantity as the logarithm to base 2, BIT. Logarithmic units are useful for operations with quantities that increase by geometric progression. Logarithms transform multiplication procedures to much simpler addition ones. Now, computer science uses larger units, bytes (1 byte = 8 bits).

Physicists deduce the information quantity from the Boltzmann concept of power entropy according to the equation

$$S = k \cdot \ln \omega,$$

where $S$ is power entropy, $k$ is the Boltzmann constant, $\omega$ is the number of states. Consequently, in this case the information quantity is measured in NATs, the natural logarithm to base *e*. In computer science, other units for measurement of information quantity are used, *e. g.*, DITs, the logarithms to base 10. It is possible to adjust the unit for measurement of information quantity to technological characteristics or symbol systems.

Having in mind that living systems are based on bioinformation coding-decoding procedures using various codes, it is useful to introduce adequate bioinformation units adjusted to the elements of a particular bioinformation technology. For measurement of information quantity contained in nucleic acids, which implement genetic bioinformation functions using four symbols, A, G, C, and T(U), it should be recommended to use a logarithm having the base 4 as the information unit GIT. For measurement of information quantity contained in proteins, which realize genetic information transforming it to the effectory subsystem made of 20 symbols (amino acids), it should be recommended to use the logarithm to base 20 as the information unit PROT.

When an information-carrying structure has $N$ steady states with equal probability, it carries the largest quantity of information and has the maximal uncertainty, information entropy $H_0 = \log N$. The units are determined by the base of the logarithm. Therefore, it is possible to evaluate the quantity of information carried by a single state of the different coding systems:

$$1 \text{ PROT} = 1.3 \text{ DIT} = 2.161 \text{ GIT} = 2.995 \text{ NAT} = 4.322 \text{ BIT}.$$

Using these proportions, it is easy to pass from one unit of the information quantity to another.

The coded information is carried and transmitted by symbols (structural elements), stringing them into particular sequences (words, sentences, etc.). In reality, it never happens that all symbols are used with equal frequency and without correlation. In such cases of informational redundance, when symbols are used with different frequencies, information quantity carried by a single symbol is estimated by the information entropy $H$ acoording to the Shannon equation

$H = -\Sigma p_i \cdot \log p_i$,

where $p_i$ is probability for the $i$ symbol. The quantity of carried information $I$ could be calculated using information entropy of the code, since

$I = n \cdot H$,

where $I$ is the overall information quantity of a sequence, $n$ is the number of elements forming the sequence, $H$ is information entropy of the code. According to M. Yèas' example [8, 9], information entropy of the code of the southern bean mosaic virus RNA is

$\mathrm{H_{RNA}}$ = 0.999 [GIT ] = 1.992 [BIT],

and that of the code of the virus protein is

$\mathrm{H_{PROT}}$ = 3.883 [BIT] = 1.441 [GIT] = 0.898 [PROT].

### On information theory in biosystems

From the point of view of information technologies, it is obvious that the functions of genes and proteins are coding, decoding and molecular informational control of cellular biotechnological processes. Thus, methods of information theory could by applied for estimation of these bioinformational technologies, especially self-correcting coding on the level of genes and proteins, protection from noise, information reliability, etc. Methods of information theory could be applied to biosystems of a higher level, *i.e.* for control of a multicellular organism by hormons and pheromons. Since the species-specific amount of hormons or pheromons $n$ is finite (*e.g.*, 45 in human) and reactions of a biosystem are determined by the existing blend of these hormons or pheromons, uncertainty or information entropy of each blend could be defined by combinations ${}_nC_k$, in the following way:

$H_0 = \log {}_nC_k$,

where $n$ is the overall amount of hormons, $k$ is the amount of hormons in the existing blend ($n > k$). If to admit that the functioning of the organism is determined also by inequalities of concentrations of hormons or pheromons, information entropy could be defined by permutations as follows:

$H_0 = \log k!$

It seems that coding in neural networks could be defined in a similar way. When neurons in excited state transmit the excitation power by spiking frequency and information is processed by parallel functioning of neural networks, the determinant informational role is played by the inequality and order of spiking frequency [10]. In this case, the state of the neural network and the quantity of processed information should be estimated in logarithmic units of the factorial. The problem what base of the logarithm should be selected for the quantitative estimation is open, since it depends on a particular set point.

## CONCLUSIONS

1. Analysis and explanation of the informational processes taking place in molecular biology as well as development of new biotechnologies require application of methods of information theory, including matching quantitative information units: GITs for nucleic acids and PROTs for proteins.

2. The units GITs and PROTs are useful for evaluation of bioinformation quantity and could stimulate the progress of bioinformatics towards the quantitative theory of bioinfotechnological procedures.

Received 21 January 2004
Accepted 25 November 2004

### References

1. Official J Eur Communities 1991; L 189, 34(1): 23–34.
2. Valstybës þinios 1998; (6): 126–37.
3. Kirvelis D, Beitas K. Informatics in Education 2003; 2(1): 39–52.
4. Marijuan CP. BioSystems 2002; 64(1–3): 111–8.
5. Kay LE. Who Wrote the Book of Life? A History of the Genetic Code. Stanford, CA. 2000: 441 p.
6. Kirvelis D. Int J Computing Anticip Syst 2000; 5: 183–98.
7. Kirvelis D. Int J Computing Anticip Syst 2002; 13: 50–61.
8. Yèas M. In: Symposium on Information Theory in Biology. Yockey P, Platzman RL, Quastler H. (Eds). New York–London, 1958; 70–102.
9. Yèas M. The Biological Code. Amsterdam–London, 1969: 360 p.
10. Kirvelis D. Int J Computing Anticip Syst 2000; 7: 263–78.

**D. Kirvelis**

### KITOKIA BIOINFORMACIJA IR BIOINFORMATIKA: APIE INFORMACIJOS KIEKIO VERTINIMÀ

Santrauka

Loginës analizës ir sintezës metodais sudaryta gyvûno funkcinës organizacijos schema. Ji rodo, kad organizmuose slypi bent trijø lygiø informacinës kodavimo-dekodavimo bei valdymo procedûros, kurias realizuoja atitinkamos struktûros: genetinës, hormoninës (feromoninës) ir nervinës. Pasiûlyta ðias bioinformacines procedûras tirti informacijos teorijos poþiûriu, kiekvienai struktûrai panaudojant specifinius logaritminius informacijos kiekio vienetus: GITus, PROTus bei kitus, priderintus prie konkreèiø informacijos neðikliø (signalø). Atkreipiamas dëmesys, kad toks kiekybinis biosistemø tyrimo ir aiðkinimo informacijos bei valdymo teorijø poþiûriu metodas kitaip interpretuoja bioinformacijà ir iðpleèia bioinformatikos sampratà.