Department of Computer Methods
Nicolaus Copernicus University

Grudziądzka 5, 87-100 Toruń, Poland

Syntactic and semantic information in finite systems

WŁODZISŁAW DUCH

ABSTRACT

A new measure of complexity or information content for finite systems is proposed. For systems composed from a number of interacting substructures the size of the minimal graph representing all possible structures is taken as its complexity. This type of complexity measure may also be used in a knowledge-based system to measure the semantic information. An algorithm for finding the minimal graph is given. Examples of applications include complex systems such as genes, proteins, language dictionaries and games.

UMK-KMK-Technical Report-1-1993

1. Introduction: information and complex systems

What is a complex system? A working definition, given at a recent conference on the subject [1], is: "... systems that exhibit complicated behavior but for which there is some hope that the underlying structure is simple in the sense of being governed by a small number of degrees of freedom".

Another definition recently given is: "A system is loosely defined as complex if it is composed of a large number of elements, interacting with each other, and the emergent global dynamics is qualitatively different from the dynamics of each one of the parts" [2]. Fractals and cellular automata are perhaps the simplest systems, in which almost infinite complexity is generated from extremely simple dynamics. In many-body physics few-body structures cannot be analyzed in details using mathematical models and may exhibit complex behavior. On the other hand this definition seems to be too restrictive. What about those complex systems for which there is no simple underlying structure? There are complex physical structures, like proteins and other biomolecules, which cannot be analyzed using theoretical methods of quantum mechanics because they are too complicated - the number of degrees of freedom is certainly not small. The size may not be that important. Large molecular aggregates, like crystals, may have high degree of symmetry which simplifies theoretical

analysis. Large carbon complexes, like fullerenes, are highly symmetric. However, such molecules are exceptional among about 10 million of chemical compounds synthesized so far. Complex structures are too small and irregular for statistical mechanics and too large for fundamental theories to tackle. Another example of complex system not covered by the quoted definition is the structure of natural language. Words and sentences have some regularity but it is very hard to find the "deep grammatical structure" that will allow us to parse complex sentences. Vocabularies are complex, relatively large systems of information hard to analyze via mathematical means and apparently without simple underlying mechanism that could generate them. Problem solving in artificial intelligence leads to the representation spaces and decision trees that show combinatorial explosion, thus leading to complex behavior or complex structure of their solution spaces.

In some sense “complex system” means “irregular” and large enough to defy analysis by fundamental theories. Whenever complexity of a system or of behavior of a system is generated by some simple underlying mechanism we will say that it is a complex system of the first kind. Whenever complexity is more fundamental and cannot be reduced or explained by some simpler structures we will say that we have a complex system of the second kind. Of course it may be that only trivially complex systems of the first kind exist in nature and the essentially complex systems of the second kind are just artificial constructions of the human mind. Nevertheless, at the present stage of scientific inquiry it seems appropriate to develop also an approach that should allow for characterization of complex systems of the second kind.

One source of inspiration for the theory of essentially complex systems could come from the theory of information, devised by Shannon (Shannon 1948) to measure the amount of information in an arbitrary data system. Different definitions of information exist now, including axiomatic definition (Ingarden and Urbanik 1961, 1962), algorithmic information (Chaitin 1987) and pragmatic information (von Weizsäcker 1974). In particular various algorithms for data compression use this kind of analysis, creating “hashing tables” for commonly encountered groups of symbols. One could define the amount of information in the complex system as infimum over all compression methods. This definition gives us at least an upper bound to the amount of information in the system, for example, if we have a text data we can compress it using different algorithms and claim, that it contains at least as much information as the number of bits in the smallest compressed file we obtain.

$$I_C = \log_2 N$$

More than 40 years after the definition of information appeared in the landmark paper of Claude Shannon (Shannon 1948) we still do not have a satisfactory definition of information that would be in accord with our intuition and that could unambiguously be applied to such concepts as biological information or linguistic information. We will propose here a practical way of measuring the syntactic as well as semantic information content in the finite systems which is very much in accord with the idea of pragmatic information. This new concept of information will lead to interesting ways of representing information and thus new computational techniques. Syntactic information is concerned with the amount of bits needed to store a certain data structure. Semantic information refers to the meaning of the data, an elusive concept that we shall try to formalize.

2. What is wrong with the conventional definitions of information?

The first approach to the quantitative definition of information is based on combinatorics (Kolmogorov 1968). If a variable x belongs to the set of N elements giving it a definitive value $x=a$ an "**combinatorial information**" is transmitted. For example, in a list of all k-digit binary strings $N=2^k$ and the number of bits of information gained by sending one such string is k. This measure of information simply shows, how many binary digits one has to use to code an information. Numbering the set of arbitrary N elements I_C binary digits are sufficient to distinguish each element. The total combinatorial information contained in a list of all binary strings is therefore $kN = k2^k$, i.e. it gets larger with the length of the strings. If this is a full list we can sum the information with one sentence: all k-digit binary numbers. This length of this statement is independent of k or N. If there are a few exceptions we would keep a list of the "exceptional" strings, and only in the case when a completely unrelated list of strings is given $k2^k$ digits will be kept. Combinatorial information evidently refers to the rather uninteresting situation when the list of N elements has no internal structure.

Interesting applications of combinatorial information to the estimation of “entropy of a language” have been reported (Kolmogorov 1968). The entropy of words in a dictionary is considerably higher than the entropy of words in a literary text, indicating that there are some constrains (grammatical and stylistic) in literary texts.

The second approach is based on the probability. It was introduced in the theory of information transmission by C. Shannon (1949). The concept of **probabilistic information** is based on Shannon's formula

$$I_P = -\sum_i p_i \log p_i$$

requiring specification of probabilities p_i of events i. It is identical (except for the units) with the physical entropy of the system. If we know in which state the system is, i.e. all p_i =0 except for one p_k=1, we gain no information observing the system, i.e. I_P=0. If all N states are equally probable the information gain is the largest and equal to I_P= $\lg_2 N$.

Let as suppose that we want to find the amount of syntactic information contained in a dictionary of words, constructed from an alphabet using some grammatical rules. In contrast with the artificial grammars and alphabets natural language contains only a very small fraction of all possible words, that could be constructed according to the grammatical rules. Shannon information is directed rather at transmission of signals than of larger units. In a larger dictionary or other lexicographical structure the probabilities of letters are close to estimated probabilities for the language. If all 27 letters of English alphabet (including space) were equiprobable in a dictionary containing N words with N_{av} length there are $N(N_{\mathrm{av}}+1)$ letters and spaces, therefore information contained in it is equal to $N(N_{\mathrm{av}}+1)1/27 \lg_2 27 \approx 0.176\ N(N_{\mathrm{av}}+1)$ bits, so for a dictionary of 100000 words of length 10 it is 193 kbits. Taking into account the unequal probability of letters this number may be on a factor 3-4 smaller. Unfortunately this information does not give any measure of complexity of the data involved; the amount of information for complex as well as for simple repetitive structure is exactly the same.

Asymptotic behavior of the I_P information measure also does not seem satisfactory. It assigns the largest information to the data structures that are uniform. Suppose that we have a list of n-digit binary numbers. If there is a complex algorithm producing those numbers with unequal frequencies we have a chance of obtaining rather low value of I_S because for transmission a coding of frequently appearing binary digits in a short binary string will

significantly reduce the amount of data for transmission. On the other hand if all numbers between $(0..2^n-1)$ appear only once on the list Shannon information is largest and is equal to 1 bit per digit, or n bits per word, a total of $n2^n$ bits (since all $p_i = 1/2^n$). Because bits appearing on such a list are statistically independent higher order correlations are all powers of 1/2 and the second-order and higher entropies per word are all equal to n bits. It is easier, however, to specify that all digits are present than to give a list of 10 or more randomly selected binary strings. Moreover, if we remove from the list one string the information will hardly change; for large n the change in the Shannon information may be in fact arbitrarily small. Again our intuition tells us that we need more information to specify that one string is missing if the string is longer. Probabilistic measure of information is not that much a complexity measure of a system as it is the measure of uncertainty of a given state or of a surprise at an event.

Another measure of information, called algorithmic or Chaitin-Kolmogorov information is in use in computer science (Kolmogorov 1965,1968, Chaitin 1966,1990). **Algorithmic information** or the relative complexity of an object y with a given object x is defined as the minimal length of the program p for obtaining y from x. Algorithmic information captures some intuitive features of information: a binary string obtained by truly random process cannot be compressed and carries the amount of information equal to the number of its digits. An apparently random string may, however, be easily computable from some short algorithm, for square root or the value of π but because of Gödel and related theorems it is not possible to decide whether the string is random or simple to compute. Although this definition of information has more intuitive features it is not practical either, referring to the concept of universal computer. It is usually very hard to compute the amount of algorithmic information in any non-trivial structure. The amount of computations needed to restore object y from x, called “computational depth” in the literature, is a separate issue - there are cases when simple structures of small complexity require an almost infinite time to restore them from other structures. Algorithmic complexity has found interesting applications in theoretical computer science to estimate the number of steps necessary to solve certain classes of mathematical problems. Some attempts were made to apply this concept to molecular biology, but in a recent book “Information theory and molecular biology” (Yockey 1992) there are no applications of this concept.

Lloyd and Pagels (1988) introduced an interesting measure of complexity in statistical physics, based on the concept of the “depth of a state”. Their formulation stresses the importance of evolution of a system and leads to rather paradoxical and counter-intuitive results. Cyclomatic information measure has been introduced recently in computer science by McCabe [16]. The concept of cyclomatic information is very useful in practice, enabling evaluation of the complexity of software.

Below we are going to present yet another definition of information, free from the difficulties connected with the algorithmic information and more in line with natural intuitions about information. This definition refers to the actual complexity of finite systems and is not related to the difficulty of creating them. Our intuition tells us, that the amount of information in a data structure should be proportional to the size of this data structure and the complexity of the data. It is not enough to look at the number of data bytes, as is quite common among computer users. Much better characterization of information contents is afforded by various algorithms for data compression, creating "hashing tables" for commonly encountered groups of symbols. One could define the amount of information in the complex system as infimum over all compression methods. This definition gives us at least an upper bound to the amount of information in the system, for example, if we have a text data we can compress it using

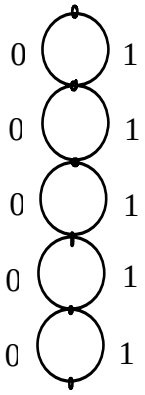
different algorithms and claim, that it contains at least as much information as the number of bits in the smallest compressed file we obtain. Although it is possible to formalize this notion it is as hard to use it in practice as the Kolmogorov-Chaitin information measure. Moreover, we do not see any natural extensions towards the semantic information measures.

3. Syntactic information in practice

Let us first define a measure of syntactic information. We can use various equivalent languages to formulate it: automata theory, statistical correlation matrices or graph theory. Let us first make a simple example to show, how to express the same fact in these 3 formalisms. Imagine that our dictionary contains all n-digit binary numbers, from (0000...0) to (1111...1). Since this is a regular structure a simple automata is sufficient to recreate it:

$$A_i(x) = 0,1; \quad \text{i=1..}n$$

Graph G_0 representing this situation is given below.

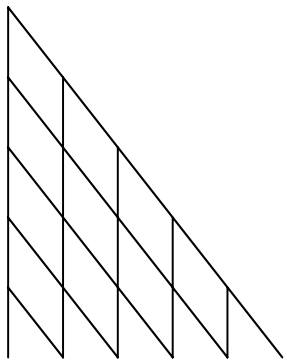


We can also specify a 2 by n matrix giving a probability of finding 0 or 1 at each position in the string:

$$M_{i,x_i} = \frac{1}{2}; \quad i = 1..n, \quad x_i = 0,1$$

with constant values of all higher-order correlation matrices. Thus if we have completely regular structure it may be specified by a short algorithm and thus its algorithmic information will be small.

For structures that are less regular algorithmic description will be longer. We shall present the results only in the forms of graphs from now on, although an equivalent description may be obtained by specifying corresponding automata. To get more information we can use somehow more elaborate graphical structure G_1.

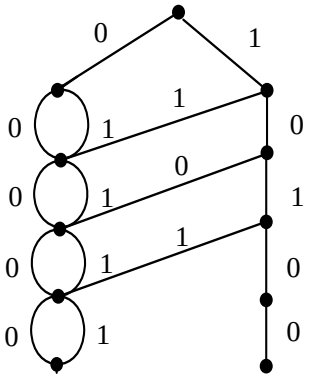


1 5 10 10 5 1

Here each node of a graph gives us an information about the number of 1's and 0's in the binary string passing through the node. Removing a node or an arc in such a graph gives us slightly less regular structure that cannot be represented by the G_0 graph. Minimal graphical structure which has still some regularity may be represented by combining the G_0 and G_1 graphs. A set of randomly chosen binary strings may not be representable by such a minimal structure, but it can always be represented by a tree-like data structure G_2. Number of nodes in this type of graph is 2^{n+1}-1.

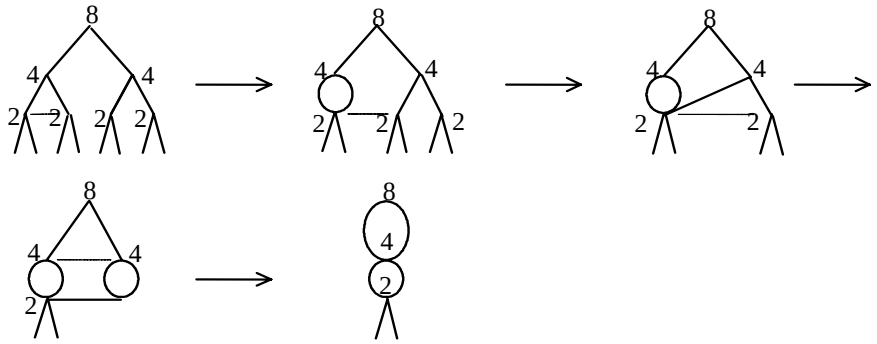
Although in the example presented above all 3 graphs are equivalent in general this is not the case. The tree G_2 may represent an arbitrary sequence of binary strings or an arbitrary dictionary of words. Each node in a tree may be reached in one way only, thus the tree contains full information about the path reaching it. In an n-level binary tree there are 2^{n+1}-1 nodes. G_2 graph contains the same information, however, there is a number of path reaching each node, the only information that we can find is that these path have fixed number of 0's and 1's. G_1 graph does not allow even that.

There are two extremes: if the dictionary is completely regular it may be presented in the form of a compact graph, rather than a tree. If it is completely irregular it corresponds to a tree with every subbranch unique. An intermediate example is shown below: one path is removed from the regular graph.



In the example given above for n-digit strings the graph for all binary strings has $2n$ arcs, while the graph for all strings with one string removed has at most $4n$-3 arcs, with 2 strings removed at most $6n$-8, ect, so that the difference is proportional to the number of digits in a string as one should expect, since the number of digits one has to write in order to specify the exception to the rule is proportional to n.

What is the amount of Shannon information in all binary numbers of n digits? There are 3 characters only, 0, 1 and space. Total number of characters is $N_t=2^n\,(n+1)$ and the probability



of each is $p_0 = p_1 = 2^n\, n/2N_t = 1/2$, $p_s = 2^n\, /N_t = 1/(n+1)$, so the amount of information is around $2^n\, n/2+2^n\, n/2+1 = n2^n$ bits, i.e. is proportional to the number of bits on the list. This is certainly true if the bit strings were completely random. However, once we know that the list of strings is complete or that it is regular, for example composed to the same string 101010...10 repeated 2^n times our intuitive understanding of what information is tells us that it should differ from the random list. Removing one or more binary strings from a long sequence of strings changes this number insignificantly.

In most cases there is some additional structure in such trees, they can be simplified folding certain parts of the subbranches into single nodes. In particular various algorithms for data compression use this kind of analysis, creating "hashing tables" for commonly encountered groups of symbols. One could define the amount of information in the complex system as infimum over all compression methods. This definition gives us at least an upper bound to the amount of information in the system, for example, if we have a text data we can compress it using different algorithms and claim, that it contains at least as much information as the number of bits in the smallest compressed file we obtain.

Definition: *The information contained in a dictionary is proportional to the size of the minimal graph encoding it.*

Finding the minimal graph may at the first glance look as a NP-complete problem, requiring pattern matching techniques to find subgraphs that are the same in different parts of the graph and than choosing some subgraphs as supernodes and gluing together some other nodes. Fortunately it is not so bad and we shall give an algorithm for finding the minimal graph below.

An important class of problems is addressed if only the head and the tail fragments are allowed to be folded. No supernodes are introduced in this process, only gluing of vertices will be used to reduce the size of the graphs. The algorithm of graph reduction folding the tail patterns is described below.

Preliminary step: using the depth search loop over all nodes in the graph and assign weights to the nodes of the graph, adding to the weight the number of nodes visited below given node.

Reducing the number of vertices: for each non-terminal node:
find a node to the right with the same weight;
check the patterns below; if they are the same glue the nodes

The complexity of this algorithm is of the order of the number of nodes in the graph squared. Applied to the completely regular graph G_3 it is going to transform it to the minimal graph G_1 in a series of steps, illustrated below:

The dotted line shows at each stage which nodes are going to be glued. One can also fold the lowest level by adding an additional, terminal node. If the length of the strings (words) in the graph is not constant such additional level may be useful. The algorithm described above folds the graph in this case to the minimal size.

Another way to fold the graph is based on sorting in the reversed order. We have estimated the complexity of these algorithms to be of the same order. We have implemented (Duch and Jankowski, 1993) these folding algorithms in a computer program, applicable to any lists of words composed of some alphabets.

Fractal images are very interesting example of objects with infinite complexity. However, looking at their pictures we may compare their complexity in an intuitive way. Can the proposed definition capture this intuition? First, we should note that algorithmic information is not useful in this case, because all Julia sets obtained from z2+C iterations have the same complexity. For different complex constant C they are, however, very different. For C=0 we get circle, for C=i rather complex picture. Although the complexity of fractal images is infinite (except for some degenerate cases) we should be able to introduce an isotonic relation (?) that will compare the actual pictures, with finite resolution, as we see them. Each little square of a given color is assigned the color number, 8 bits for 256 colors for example, and each line is a word on the list. The complexity of the list of words is measured by the size of the minimal graph again. Interesting questions that arise are:

Although the complexity of the structure goes to infinite with the resolution of the fractal image going to infinity can we prove that the relation "more complex than" is preserved, at least starting from some resolution smaller than a fixed one?

In general finding the smallest possible representation requires not only folding of the existing nodes in the graph but introducing "supernodes" for subgraphs of the same structure that appear in the few places in the graph. Unfortunately an algorithm for optimal selection of such supernodes is NP-hard, leading to the classical problems of combinatorial optimization, such as the knapsack problem.

It is obvious that the amount of information, in the sense of coding theory, depends on the model of data we have. In particular one can compress information of arbitrary length into just one bit. Suppose that *M* is the given message to be transmitted. The code for this message is composed of bit 1 while the codes for all other messages start with bit 0.

4. Alphabet, words and correlation matrices

There is an analogy between the theory of complex systems and linguistics. If the system is not completely chaotic we can find an alphabet, a list of substructures or elements of behavior, which, due to some interactions, generate complexity. Interactions in this case are analogous to the grammatical rules.

Let us imagine a simple example: 4-letter combinations from a 3-letter alphabet.

letter no.	letter abcd
1	1000
2	0110
3	1000
4	1000

Third order structure, for example XOR problem, contains the following vectors:

(1 1 0), (1 0 1), (0 1 1), (0 0 0)

and gives uniform first and second-order correlation matrices. In this case adding one more fixed bit, like (1 0 1 0), (1 0 0 1), (0 0 1 1), (0 0 0 0), reveals the structure in second order correlation matrices!

It is interesting to look at the minimal graph generated by the grammars of different type (cf. Chomsky 1959). The structure of these graphs reflects the hidden dynamics due to the production rules. In case of artificial languages it is easy to go from the grammatical rules to the minimal graphs showing all possible words constructed using these rules. In case of natural languages the task of linguists is to find the rules from the words in dictionary. Since the rules are only approximate, with a number of exceptions, one can analyze the graph disregarding minor irregularities.

Hierarchy of such correlation matrices may be introduced. The challenge is to find minimal structures which strictly encode the original list! Search in correlation matrices for what is acceptable corresponds directly to a tree structure of alphabetically ordered words. If there are some words created by interference they are required for the structure to be more regular. It should be quite interesting to compare this structure after folding with the structure of the true graph after folding. Here is the scheme:

List => Graph => MinGraph
List => Mcorr => GraphM => MinGraphM

Are the MinGraphM more regular (i.e. smaller number of arcs per word coded) than MinGraph? Connection of minimal graphs and correlation matrices may introduce explicitly allowed versus explicitly forbidden cases. Correlation matrices contain probabilities for Shannon information I_N.

Another area where information measure based on graph theory is useful is the computer science. Imagine that each of the nodes in the graphs presented in Fig. 1-3 represents a fragment of the computer code and the graph itself represents possible flows of computation, from the starting node to one of the final nodes. Each node from which more than one path

follows is equivalent to a piece of sequential code followed by a decision statement. Such a representation of the computer program is called a “flowgraph” and frequently used in computer science. In fact one of the most successful and used commercially approaches to measurements of software complexity, called “cyclomatic complexity” [17], is simply based on the number of decisions in the flowgraph:

$$v(G) = \text{number of decision statements} + 1$$

Correlation between the values of cyclomatic complexity and the number of errors in the program and enabled predictions of performance of programmers on comprehension, modification and debugging of the programs.

5. Semantic information

Once we have accepted the definition of syntactic information given above the definition of semantic information is not much harder. Semantic contents or meaning is relevant only if we have some cognitive system. Words and ideas have different meanings and different information contents for different people therefore it is not possible to give a universal definition. Why is the meaning different for different people? Because their internal representation of the world is different. We must refer to some representation of the world to define semantic information. Suppose that we have build an expert system, based on a set of rules stored in a knowledge base. This knowledge base, together with the rules of inference, define our universum of facts, representing knowledge or some model of the world *M*.

One of the ways of storing such a knowledge base is via semantic networks. We can organize these networks in such a way that the minimal graphs are used to describe them. The information contents of this knowledge base is defined as the size of this minimal graph. Let us add now one new assertion to the knowledge base and try to determine the amount of information that this assertion adds to our knowledge. This will be a measure of “meaning”, or semantic contents, of such an assertion. Adding new rule to the knowledge base requires accommodation of this knowledge with the other knowledge, resolution of possible conflicts ect. This process is analogous to the process of relaxation in the neural network. For example, when a new fact is learned by a human it takes from several seconds to minutes, days or even years before this process of relaxation or accommodation of a new fact takes place and the "meaning" is understood. Years of repetition of basic facts in mathematics and natural sciences are required to really digest the information: once it has been integrated into our "knowledge base" we can review the information contained in the schoolbook in a day.

Comparing the size of the minimal network after adding one rule and accommodating it with the size it had before gives us the measure of semantic information contained in the new rule relatively to a given knowledge base. Thus the same fact has different semantic information contents depending on the knowledge base that is used.

6. Possible Applications and Open Problems

Does it fit to intuitive feeling of semantic information? Rather well. We have developed (W. Duch and N. Jankowski, unpublished) a computer program to measure the amount of syntactic information in complex structures. As a first step it analyses the data to determine

the alphabet; then it generates the graphical structure and makes multiple passes over it to determine minimal structure.

There is one problem left - it does not allow us to define absolute amount of semantic information in the knowledge base. A large number of particular rules, requiring rather large semantic network to represent them, may have a weaker predictive power than one general rule, requiring smaller network and thus, according to our definition of information, containing less information. While the concept of meaning of a single rule seemes to be well captured by the resulting change in the size of the knowledge base data structure the absolute amount of information requires a concept of predictive power (what facts can the knowledge base generate) which is not only dependent on the knowledge but also on the inference mechanisms. Basing on the predictive power we could introduce an ordering relation between 2 knowledge bases, fixing the inference mechanism.

Another problem is how to extend the measure of information described here to infinite systems. An obvious way is to discretize them but there might be better ways of doing it.

References

[1] Chaitin G (1966), On the length of programs for computing finite binary sequences, *J. Assoc. Computing Machinery* ***22***, 329-340

[2] Chaitin G (1987), Algorithmic Information Theory, Cambridge University Press, Cambridge

[3] Hofstadter D. R. and Dennett D. C., The mind's I. Fantasies and Reflections on Self and Soul (Basic Books, New York 1981)

[4] Ingarden R.S. and Urbanik, K (1961) Information as a fundamental notion of Statisitcal Physics. Bull Acad. Sci. Pol, Ser sci. math., 9, 313-316

[5] Ingarden R.S. and Urbanik, K (1962) Information without probability. *Colloquium Mathematicum*, **9**, 131-150

[6] Kirkpatrik S, Gelatt Jr. C.D., Vecchi M.P. (1983) *Optimization by Simulated Annealing*, Science **220**, No. 4598, 671-680

[7] Kolmogorov A.N. (1965) Three approaches to the quantitative definition of information, *Problemy Pieredachi Informatsii,* **1**, 3-11; translation (1968), in: *Internat. Journal of Computer Mathematics,* ***2***, 157-168

[8] Lloyd S, Pagels H. (1988) *Ann. Phys.* **188**, 186-213

[9] Nicolis G, Prigogine I. (1989) *Exploring Complexity,* W. Freeman and Co, New York

[10] Penrose R (1989), *The Emperor's New Mind* (Oxfors University Press, 1989)

[11] Shannon C.E. (1948) A mathematical theory of communication, *Bell System Techn. Journ*. **27**, 379-423; 623-656

[12] Shannon C.E. (1951) Prediction and entropy of printed English, *Bell System Techn. Journ*. **30**, 50-64

[13] Shannon C.E. and Weaver W., *The Mathematical Theory of Communication* (University of Illinois Press, Urbana,1949)

[14] von Weizsäcker E., Erstmaligkeit und Bestätigung als komponenten der pragmatischen Information, in: *Offene Systeme I,* ed. E. von Weizsäcker, Ernest Klett Verlag, Stutgart 1974, pp. 82-113

[15] Yockey H.P., Information theory and molecular biology (Cambridge University Press 1992)

[16] Zurek W.H., *Algorithmic Information Content, Church-Turing Thesis and Physical Entropy,* in: 1989 Lectures in Complex Systems", Ed. E. Jen (Addison-Wesley Publishing Co, 1990), pp. 49-65
[17] McCabe, T. J, Butler, C.W. Comm. of the ACM. 32 (1989) 1415-1425.