DATA COMPRESSION AND COMMENTARIES ON THE QUESTION OF DECODING THE GENETIC CODE

N. Nanobashvili

I. Vekua Institute of Applied Mathematics
I. Javakhishvili Tbilisi State University

The works [1,2] are dedicated to optimality problems of the four-letter code systems from the point of view of data compression, recording and storage. The present paper deals with the reflection of the above mathematical investigation on the question of decoding of biological codes.

One of our scientific interests is establishing such a code system structure which by comparison with other code systems would prove to be more economical and compact from the point of view of recording and storage of information.

On the basis of carrying out mathematical investigation in this context the code system consisted of four-letter alphabet proved to be the most optimal one.

The mathematical investigation will help us to find the rational way of solving some very difficult problems.

From this point of view the results obstained in [1,2] can be transferred and reflected in the next items:

1) The cause of expediency of creation of the four-letter code information system is explained.

It is mathematically proved that the four-letter code system represents that minimum alphabet on the basis of which compression of information is realized (and first of all it means that the four-letter coding system is the most economic one).

2) The principle of intercommunications and unity of the information systems during recording and compressing is shown, which is very actual.

The system takes information from the outside world (in particular from living organisms) in the form of like two-letter alphabet signals as it occurs, on the whole, in the neuron nets: a reaction either takes place or does not. The two-letter code sequence transforms into the four-letter code sequence (It is no matter of principle how it will represent in the electronic form or bio-chemical signals).

According to the sequence and contents of coming information from the outside world (which is two-letter at first), transposition of letters occurs in the fragments consisting of four letters. It is mathematically shown how and in what order the information is recorded and transformed from the two-letter system into the four-letter one.

We think that maybe some unknown to us superintellectual creators of the genetic code followed such a principle of recording information and they have

prepared the genetic code personal variants for every creature living on the Earth.

3) It is necessary for data compression and for its simple decoding later, that the four-letter code would be degenerate. In other words, realization of one and the same instruction is fulfilled by means of a few code fragments.

4) If the information in the code is compressed, then transition even of one letter into another causes distortion and change for the most part of information in the time of restoration (synthesis).

This phenomenon helps us to understand mutation processes very clearly. But the code degeneration is connected not only with its compression and simple decoding.

5) It is mathematically proved in the above-mentioned works that if the four-letter code is degenerative then the code consequences consisting of three letters (triplets) are characterized by a maximum compression coefficient.

In such a case (if the code consists of triplets) information capacity of each of them (taking into account the interconnection between them) equals twenty bits. Theoretically (mathematically) it means that the code fragment consisting of three letters has twenty positions (twenty variants) in the time of recording information. The rest code fragments consisting of three by three letters must be considered in the degeneration context.

We think that the twenty positions (this magic number "twenty") ought to be connected with twenty aminoacids phenomenon.

6) Apart from the above-mentioned results and proofs obtained mathematically there exists a number of theoretically inexplicable experimental results which no improved microscopy and supercomputer technique can completely explain. For this we need not only deep global comprehension of the existing problems, but deep theoretical prediction as well which will connect experimental and theoretical results and explain them as a whole from a single position.

Mathematical modelling, information and coding theory, particularly, data compression can answer many questions in the existing problems. We think that they all are able to make an important contribution to comprehending some fundamental problems of the project "Human genom" more profoundly.

One of the principal counter-arguments of the opponents is that decoding the code is connected with colossal difficulties and that only 10% of the genetic code contains information but the rest 90% are senseless.

The situation is rather puzzling: It comes out that God or some superintellect created a universal information system the genetic code which is so unfit that 90% are senseless and quite useless!

A. Einstein has once said, "The Nature never throws the dice". In other words the nature has rationally provided for evrything and for the genetic code structure too.

If the four-letter system is the most optimal (economical) one, then the assertion, that a text consisting of it is extremely noneconomical is very illogical.

7) Improvidence for compression phenomenon is the main cause of misunderstanding and difficulties.

It is necessary for simple restoration of compressed data to introduce additional symbols (at first sight it seems senseless) in the primary code sequence.

The more the primary data is compressed the more the additional number of letters increases in order that code should be restored after that.

8) In the context of "Human genom" (in the genetic code) the value of data compression coefficient is more than a milliard. Receiving data compression coefficient of such a large value by means of only onecyclic compression process, is frankly speaking, nothing but utopia. It is possible to get it only in the case of multicyclic compression process. As the compression cyclical processes are interconnected, this factor matters very much to the problem of decoding the genetic code in full.

Compression reverse process, i.e., the process of information restoration is characterized by analogous cycles. The process of the man's evolution and growth from the time of his conception to his birth lasts nine months. This time is used for restoration of compressed data stored in the genetic code and for programme realization of ablumens synthesis according to this restored information. The number of cyclical processes in the time of restoration and the time from conception to the birth must be in direct connection with the data compression coefficient value accordingly.

9) The four-letter code systems (when the alphabet consists of four letters) are extremely interesting from the point of view of recording information and compression. The genetic code itself is only one of variants among the codes which can be created in the four-letter systems. In other words, the genetic code is such a four-letter code system which is adapted as much as possible well to the vital and biological processes and speeds of reactions going on the Earth. The phrase "biochemical speeds" is not evident here. It means, that it is possible to create the four-letter system which will be from the point of view of information restoration rapidity really more effective than the genetic code. That is why we consider that the genetic code is only a private case of those common normalities which are the basis of modifications of recording information and compression phenomenon in the four-letter systems. And only such a common approach will make progress in the genetic code decoding.

Otherwise, it will take a whole century to establish interrelation between different parts of the genetic code. It is impossible to understand the cyclic processes of compression and restoration only by experimental way (whatever technical equipment would be).

Making the structural map of the genetic code in the "Human genom" project, as they have been announcing lately, is a very great achievement, but it is only the first step for making progress in the great problem of its study.

It will give us an opportunity to grow teeth and hair and to improve a work of different organs regeneration, but it will not give us an opportunity to solve a number of global problems at all, for example, the problem of life

prolongation.

10) The genetic mechanism which regulates and fixes a time-limit of life lenght in the living organism, as a whole, has an influence on each organ viability (if we do not take into account some possible anomalous genetic deviations). It is impossible to affect this mechanism even in the embryonic stage for the information of life lenght time - limit in the code. That is why it is necessary for solving this problem to comprehend how recording of the initial information about life lenght time-limit and after that its memorizing and translating into four-letter code fragments goes on.

In the end, every attempt of solving this problem will lead us to the necessity to establish the schemes and mechanisms of interconnection and unity of different kinds of information.

Otherwise, the existent experimental results whatever brilliant and great they could be, are only the part of solution of the most complicated problem and remind us the attemps of great researchers of the past to find and establish the all - powerful nonexisting "philosophical stone".

11) The results described in [1,2] and obtained in the following years as well can be briefly formulated in this way:

1a) In the time of recording information by the transposition law information compression begins from the four-letter code system.

2a) It is necessary to take into account the principle of information unity in the process of compression and memorizing. Exactly: the information initially is presented in the universal two-letter code (as in neurons: the reaction either occurs or does not) and then is translated into four-letter code.

We consider that diverse biochemical modifications and realization forms must exist (for example, as the ferment systems and so on). It is to be noted what way recording information from the outside world goes on and how it is reflected in the four-letter code system. We think that in such a way programmes of protein synthesis, every organ evolution, and a time-limit of its life length and so on are recorded.

One of the possible variants (apart from the other variants obtained by us), of recording information, its transformation into the four-letter system after that and its compression are given and considered in these articles.

3a) It is necessary for information compression and for the compressed information simple decoding the four-letter code to be degenerate.

4a) From the mathematical point of view various variants of the degenerate code can point out that different organisms can have various alternative variants of the genetic code for every organism accordingly.

5a) It is proved (mathematically) that if the four-letter code system is degenerate, the triplet (a fragment consisting of three letters) is characterized by a maximum compression coefficient.

6a) The degeneration coefficient of the four-letter code system presented as a triplet equals three (3). It means, that every instruction is manifested on the average in three triplets.

7a) The information capacity of the degenerate code presented as a triplet equals twenty (20). It means that the number of insrtuction equals twenty (20). This number coincides with the number of aminoacids in the genetic code as well.

8a) The meaningless letters do not exist in the genetic code sequence. Rule which restores the compressed information is recorded in meaningless letters. The restoration of information consists of numerous interconnected cyclic processes.

9a) From the point of view of compression the triplet (codon) constitutes an interesting information model. The size of the triplet equals six bits. But in the static state (that is when information is not recorded) it contains only one bit of information. From the point of view of compression this fact seems meaningless (illogical).

But in dynamics under which cyclic processes of recording information are meant, as it was noted above, more than six bits of information quantity are recorded in triplets.

Interconnection of such cyclic processes and their organization are the basis of the graat phenomenon which is the condition of every living organism existing on the Earth and not only on the Earth.

10a) The genetic code was created in the process of recording information. But we do not think the hypothesis that some four-letter sequence existed beforehand and the letters permutation was realized as a result of the outside influence is right and if the variants of genetic codes of the living organisms were obtained in this way. But in reality, in our opinion, the next phenomenon must have taken place: In some material medium (for example, in some organic broth) in the time of recording (under outside impulses influence) four-letter code sequences arise.

We think that the form and rule of recording in accordingly selected material medium realizes, on the whole, four-letter code sequences outcome (creation of the four-letter code sequences from the two-letter systems is shown in the articles metioned above, which at the same time points out an example of realization of the unity principle of information systems).

Under another influence (mechanical, radiation, light and so on) some number of letters can be liquidated, but it will be impossible to record a new information on a global scale then. In other word, we shall not be able to get from the man's genetic code the variants of the genetic codes of a whale or a cockroach.

12) One extremely interesting question is connected with the items mentioned above: how the living system sorts, develops and perceives at different levels the information obtained from the outside world or the information which is already recorded in it, for example, in the form of the genetic code.

A living organism realizes sorting, development and perception of information according to one and the same principle, i.e., the principle of hierarchical architecture. It is a universal principle on the basis of which the information,

having got to an organism in various forms, develops in accordance with one and the same fundamental coding scheme. Any signals are translated in the form of a four-letter coding system.

The appointed principle is the basis of functioning such a system as the eyesight system (the eye retina), hearing organ (ear) and sense of smell.

This universal principle gives an opportunity to mathematically simply (multilevel coding) and physically in full (clearly) describe and explain some phenomenon, which has not been explained yet. It is especially to be noted here that by means of it it is possible to explain how the eye (or the ear) carries out wholeness perception of a form (Gestalt). No computer and no other technologies can be compared with this principle in effectiveness. Any supercomputer system can not realize whole perception of form.

This principle is the basis of recording and compression of information when forming the genetic code beginning from conception to maturity. The principle mentioned above is the basis of account of sequences of cyclic processes. It is the basis of programme control according to established hierarchy also.

It is especially to be noted that data compression encreases effectiveness of this principle. Data compression phenomenon is fundamentally united with multilevel (hierarchical) coding.

13) We take the liberty of considering one of the most important issues. Establishment of the structural map of the "Human genom", as it was introduced above, is a significant step forward on the great way which leads us to decoding (solving) of a great cosmic mystery.

Every new successful step forward should make the scientists stop on their way for a little in order to answer the next question: is it worth going on?

Is it possible that as a result of decoding humanity will hang the second sword of Damocles over its head? Humanity has already hung the first sword of Damocles over its head long ago using the atomic weapon.

Then at what level must the mankind, rise so that every next step forward on the above way would not be dangerous and would not cause disturbance?

To answer all these questions we can cite the expression that in spite of its remoteness is actual today too: "The mind and the reason" (common sense) are two different notion. The human mind is able to find ways in order to turn the impossible into the possible while the reason asks it (the mind) the question: is there any sense in it?

The author of this expression, the famous scientist Max Born, worried that the mind often gains the upper hand over the reason instead of their concerted actions and correcting one another. "Every man is pleased with his mind, but he is not satisfied with his position". This English proverb points out one of the principal reasons why the mind having turned the impossible into the possible can not often control any more how far it has withdrawn from the reason (from common sense).

It is known from the time of Dr. Faust already that one can sell his soul to the devil even at home. There were some intellectuals always who were

ready to sell their souls. Establishing the scientific truth was a matter of minor importance for them. Historic dialectics is the basis of every important scientific discovery. Every new word in science should commence goodness, but this goodness, for its part, may initiate new evil. He who doesn't know history is under risk of its repetition. But before saying something concrete about danger let us note global-scale goodness which may be brought by the mentioned scientific novelty.

a) The genetic code decoding in information context is of the greatest intellectual importance. From this point of view we can point out one factor at least: the man has not created yet such universal mathematical apparatus which will be able from single position to describe mathematically in common as well as concretely all the processes going on in us and in the world around us.

But the common and the concrete, the alive and the developing, the movable and the immovable, the similar and the approximate, propensity and behaviour, future danger and pathology and so on, each of them is recorded in the genetic code in the equal way and simplicity. It is universal apparatus in which a programme is recorded how everyone will live and die on the Earth and not only on the Earth.

One of the principal merits of the genetic code is its extremely effective ability for data compression. From the point of view of compression it is a unique information code system and the compression phenomenon will become one of the basic trends in mathematics of the XXI century. If it has not happened in the XX century yet, it was for the simple reason that data compression phenomenon from the point of view of mathematics is a phenomenon shrouded in great mystery, which is (directly) evident in the genetic code problems. To its scale and applied significance it (compression phenomenon) will be varied and grandiose.

The compression phenomenon belongs to the cosmological scale problems as well, and it is able to explain a lot of grand processes going on in cosmos which have not been explained to this day.

That is why the genetic code decoding in the information context will be a matter of great importance not only for biologists. It will be able to cause the intellectual level growth of researchers in any field of science and hasten research processes.

It should be specially noted that mathematics will gain a new formal apparatus which will be able to describe in the equal way and simplicity micro and macro processes taking place on the earth as well as in cosmos and realize out their mathematical modelling. The most highly effective period of research and cognized of the world will come in the history of mankind.

b) God cognized all the things as a result of immortality. As a result of immortality he takes position of the highest mastery of the matter control. The genetic code decoding in the information context gives an opportunity to alter life length determined by nature. This fact will cause grand change in

people's life and especially in their interrelation.

As the scientists will be able to record a new information in the genetic code before the embryonic stage of evolution, so so-called "life prolongation and renewal code fragments" will be introduced in the code. As the result of influence on the code fragment the hands of the biological clock of man's life prolongation can be set in tens of years after his birth. Each point of this biological clock will equal 10-15 years.

Natural death that has never distinguished one man from another has been the most democratic phenomenon in the history of mankind existence. Changing natural death date and the attempt of its abolition will cause the reappraisal of existing values and their annulment. It will become the measure of appreciation of the man's behavior and his way of life. The man's service before mankind will be appreciated by life prolongation according to a fixed size, but his offence will be followed by his life reduction according to a fixed size. In other words, the resolutions of separate organs and courts about how the deposition of the biological clock hand of this or that person must be changed will replace orders, medals and other encouraging means, on the one hand, and prisons and other punishment means, on the other hand. Influence on people's fate will occur on a global scale. In this time a group of scientists and intellectuals, having sold their souls to the devil, will become one of the principal parts of the ideological mechanism of the tyrannical and plutocratic states. They will try to justify the state interference in people's fate on the pretext of their race and generation improvement. In that case they will expres an opinions similar in contents, but different in form which will very much resemble the statment by the biologist, academician Lisenko's expression: "In our Soviet Union people are not born. Only organisms are born. We make them people afterwards."

It is clear, before turning an organism to the human, these organisms should be sorted. They will be turned into merciless brutal executioners, slaves, spies, adventurers, fanatics, terrorists and what is the main thing, into lots of soldiers. This process will be called a most modern form of revolution. They who have sold their soles to the devil will find a suitable justification even for such revolutions. Which one? Well, may be like that which Alexander Block, the famous Russian poet, found for the bloody revolution in Russia: "Russia will get out of it (of the revolution) in a new fashion great." (This phrase is cited from the article "Intelligentsia and revolution," published in 1918 in the newspaper "The flag of labour").

But the effective forces of resistance will appear and oppose the dangerous forms of "the modern revolution." On the whole, two means of resistance are to be taken into consideration. The first one is connected with the existence of such states where justice, defence of human rights, the high intellect, the common sense and power will help one another as five fingers do when making a fist. Probably, it is clear what countries one can mean in this context and from this point of view, which of them must be the leader (USA).

A factor of great hopes which will be connected with life prolongation, that is with the results of decoding the genetic code, will give rise to the second type of resistance.

Naturally the question is raised: how shall we understand this last idea?

The death phenomenon which is connected with a heavy heart of the dearest and closest man's loss casts doubts on the sense of the man's birth and on the whole, of his existence.

I. Whatever considerations people created in order to ease these feelings! Some popular version inspires people that "two worlds exist. In one of them we are only guests and must leave it very soon and will come back. And if for world it is necessary that power, riches, greatness are requisite, there is only hatred to all hatred with respect to all of them in the second world. It is our choice..." (Labrier "Characters").

Great Dante's vision went much farther. In his opinion, as a butterfly leaves its cocoon and flies out of it, so the human's soul goes out of his body and joins the other world after death.

But in spite of the mentioned version the man always tried to decide and to find an answer to the question: is it worth living? Philosopher Plato gives an original answer to this question from the certain standpoint: "The body creates a lot of obstacles to us as it makes us look for nutrition for it and if some desease is added with this, the seekers of the truth remain empty-handed... That is why we shall never come to the truth..."

But in this expression Plato kept himself from calling the cause of causes and shifted the blame on the body vagaries. It is clear, that the principal reason of this is the short duration of the human life. In the days of Plato emphasizing this reason meant blasphemy.

The short duration of human life has always been one of the principal disasters of mankind.

No epidemics, no floods, no earthquake and no other natural disasters taken all together can not compare with the great damages which the short duration of human life brought to mankind.

The concise time-limit of life looks like the Dzigits trick riding. For the most part the next motives are condition of the desire for winning this race and finishing first.

a. Desire to live longer than others;

b. Desire for greatness;

c. Passion for occupying a high post in the official hierarchy;

d. Wish for growing rich.

The impetuous desire to acquire the mentioned values brought about loss of the spiritual values. It was expressed, mainly, in selfishness, treachery, mercilessness, wars and all kinds of massacre. The history of mankind, as it has been noted right by one English historian, is nothing more, but a register of men's stupid actions, offences and errors.

And if you want to reduce the number of criminal facts commited by people

and registered in the mentioned register, you should know, that it is necessary to reduce the number of unlucky, disillusioned and embittered persons. That is why, as it occurs in sport, one must have once more and again once a good chance of trying his fate in order to achieve a success. But how can one try his fate once more if he has already come to the end of his life?

Aristotel has already proved in his "Ethics" that two main sorts of crimes exist. The first one and it is the most prevalent is the crime brought out by economic motives. The second type of crimes, and the most heavy, is not brought out by economic motives (for example, tyranny). There is something in the man that conditions the second type of crimes. This type can be partly explained with the help of the works of Schopenhauer, Nietzsche, Steiner, Spinoza, Lombrozo and so on.

But it is impossible to receive a concrete and real answer without fragmentary-information research of the genetic code. And what is more: the fragments which are responsible in the code for development of the vicious characteristics of the man will be fixed. At the same time those necessary practices of structural information changes will be determined in the code which will eradicate the disposition to crimes and will protect people from them (noneconomic type).

It can be said, that two fortunate ways will open within the man's life already on the basis of bringing the genetic code into a healthy state:

1. There will occur intervention in the code itself before the embryonic stage so that the disposition to the vicious principles may be annihilated.

2. There will occur changing the points of a biological clock (life time-limit increase), so that unlucky and embittered people's number on the earth should decrease (as the result of the new chance of reaching succes given them).

For realizing these two points it is necessary that people little by little would get ready in order to enter this worthy epoch of the human society.

Special institutes to study the public opinion will be founded, in other words, filters which will count if the people selflessly fighting for goodness and justice will be worthy of immortality.

It is not clear, that it is not easy at first sight to decide negatively the question of life prolongation for some of people being in identical and similar conditions with others. But there exists a certain category of people which are characterized by indifference and that is already evidence of some progress. Just as Dante Alighieri characterized them: "The hottest place in the hell, the ninth circle, is for such people who in the time of struggle between good and evil keep neutrality". And such "neutrals" are the majority of the human race. They always patiently wait untill goodness have been rendered lifeless in the struggle with evil in order to become masters of the situation. The "neutrals" have diverse masks almost in all spheres of people's activities in private and social interrelations, in policy, science, education, art and so on.

Life prolongation phenomenon may become a remedy for bringing the human race into a healthy state. There will come into being and turn into

majority noble "mutants" and they will not be petty plungerers, egoists, liers and above all, malicious. Their nobility will make the people all around them noble. It will happen as Shakespeare wrote: "There is a lofty spirit in the lofty temple".

14) True, they have been comparing lately the progress in the problems of the genetic code structural decoding by its importance with the flight on the Moon, but the structural decoding can be declared the dawn of the second, greater stage. In other words, the genetic code decoding in the information context is at present and will be in the future the most decisive factor.

From this point of view data expression phenomenon deserves special attention, when it is determined in full

a) how the cyclis processes of decoding are connected with each other on a real scale of time;

b) how these processes are reflected on the processes of protein synthesis.

We think that just then one can say that mankind comes pretty near to the end of the solution of this great problem.

REFERENCES

[1] Nanobashvili N.D., On the Question of the Optimality of Discrete Data Compression in a Four-letter Encoding System. Trudy Tbiliss. Univ. 212, 1980, 74-92 (Russian)

[2] Nanobashvili N.D., Data Compression in Assembling Monotonic Pairs of Code Sequences. Bulletin of the Academy of Sciences of the Georgian SSR, 93, No.2, 1979, 313-315 (Russian)

Recieved September 2, 2002; revised November 4, 2002; accepted December 9, 2002.