tripleC 2(1): 20-27, 2004
ISSN 1726-670X
http://tripleC.uti.at

tripleC cognition communication co-operation

A Rational Model In Theoretical Genetics

Karl Javorszky

Institut fuer angewandte Statistik, Landhausgasse 4/23, A-1010 Wien, javorszky@eunet.at

Abstract: This model connects information processing in biological organisms with methods and concepts used in classical, technical information processing. The central concept shows copying and regulatory interaction between a logical sequence consisting of triplets and the amount of constituents of a set. The basic mathematical model of information processing within a biological cell has been worked out. The cell in the model copies its present state into a sequence and reads it off the sequence. The sequence comes in triplets and is not one sequence but appears in two almost identical varieties.

We treat consecutive and contemporary assemblies of information carrying media as equally suited to contain information. Methods used so far utilised the consecutive property of media, while in biology one observes the concurrent existence of specific realisations of possibilities. Genetics connects the two approaches by using an interplay between consecutively (sequentially) ordered logical markers (the DNA) and the state of the set engulfing the DNA.

Several mathematical tools have been evolved to assemble an interface between sequentially ordered carriers and the same number of carriers if they arrive contemporaneously. Using linguistic theory and formal logic one concludes that measurement(s) on a cell are a (set of) logical sentence(s) relating to an assembly of n objects with group structures among each other. We linearise and count all possible group relations on a set of n objects and introduce the concept of multidimensional partitions hitherto left undefined. We introduce the concept of a maximally structured set by establishing an upper limit to the information carrying capacity of n objects used commutatively and sequentially at the same time (like genetics does). The copying and re-copying mechanism which is the core matter with genetics appears in the model as differing transmission efficiency coefficients of media if the media are used once sequentially and once commutatively. In dependence of the strength of the assembly, one can transmit up to several hundred % more information using n carriers if the carriers are used commutatively, as long as the number of the objects remains within some limits.

Keywords: Self-organizing and self-replicating systems, Information processing in biosystems, Multidimensional partitions, Interaction between sequences and mixtures, Number theory of biology, Linearisation of structured sets

1. Introduction

In this paper, we present a combinatorial approach to theoretical genetics. The model allows conceptualising a mechanism which shows the interrelations between ways of information storage. We have found a bridge connecting information processing in biologic organisms with methods and concepts used in classical – mechanical, computerised, binary, technical – information processing.

We discuss the mechanisms of life in this paper from a new and unusual viewpoint: as an information theoretical problem. We target the central concept of a well-balanced system, which keeps repeating itself while maintaining a natural tendency towards change and innovation. That is, we discuss a concept of *homeostasis* or an ideal system which keeps recreating itself while balancing aspects of “*same*” and “*different*” across generations and individual growth phases.

The concept is presented in this paper as a number theoretical problem. That is, we discuss the relations of numbers among each other and state that this discussion has also an immediate, practical relevance.

Fundaments of approach

The model sets off where Frege, Carnap and Chomsky have finished (Wittgenstein 1973, Carnap 1937, 1954). It utilises a logico-linguistic method by formalising every scientific sentence that can be said about an abstract entity. A sentence of science is understood to restrict possibilities for being otherwise, and at the same time pointing out some rational relation among parts of a whole. One disregards the specifics of the results of the persons who look thru the microscope and investigates the structure of the sentences they say. Whatever the colleagues from the empirical departments can say about a cell is brought into a uniform form and each and every sentence biology can say about a cell of whichever size is generated. The new approach can be summarised as follows: *about an assembly of a limited size only a limited number of distinct sentences can be said.* As the cell is of a limited size, this principle will apply. The main inroads are: “After everything will have been said about the cell, how many distinct sentences shall we have heard?” and “Is the validity of these sentences all-at-the-same-time or one-after-the-other or a combination of both?”

Additions to accepted wisdom

We extend classical set theory by introducing the concept of the maximally structured set. This is the assembly about which everything that can be said has been said. This translates to the tool of *multidimensional partitions*. These have so far not been defined, although they exist very well. They translate Chomsky’s semantic markers into mathematics. (Congruence of domains of symbols, e.g. “brown, has 4 legs, feeds on plants, sleeps in winter”) These informations structure the logical space, until the object(s) is (are) pointed out. A second tool is a double sequence of triplets one finds when investigating recombinational possibilities of sets that grow or shrink. This is a tool very well suited to picture the DNA with. To combine the sequential way of storing information with the cross-sectional way of describing a complex material, we have introduced a funny tool that describes the recombinational possibilities of sets.

New ways of treating established concepts

- We use partitions as a probability density.
- We use a middle thing between object and logical relation by discussing the expected number of relations per object and of course the fraction of an object per relation. This is like using a matter-and-energy-equivalent. We count in “matches” or “congruences” between the sequential and the commutative descriptions of the state of one and the same structured set.
- We extend the concept of logical .t. by using .2., .3., etc. truth-values: this goes back to congruences between descriptions.
- We start off from the maximally likely truth extent of the most probable congruence between commutative and sequential descriptions of the state of one and the same set. Relative to this, one encounters under- and overdetermined facts.
- We re-extrapolate the number of objects that are most probably “in use” to carry the number of congruent logical relations from the number of logical relations that can be carried by n objects.

The general idea, explained simply

We discuss how a long sequence interacts with a broad presence. The DNA is after all something where one logical entry comes *after* the previous and *before* the next, while the cell is in a different fashion full of neighbourhood relations, there all constituents are there *at the same time* but are varied in quality. In a simple example: one looks at a hand of cards (a: how many hands of cards are possible with *n* cards?) and discusses the number of ways this hand can be played (b: playing a right card at a right moment is a capacity not beyond humans. Count and discuss). It is obvious that it is not only *what* cards one has but also *which sequence* one plays them. Contrary to intuitive belief, there are some slight differences between the number of *contemporary collections* and the number of *longitudinal sequences*. It appears that Nature makes use of a subtle number theoretical detail by continuously switching between <*A* follows from *B*> and <*B* follows from *A*>, where *A* is the set of distinct sentences that can be said about an assembly *as a contemporary collection* and *B* is the set of distinct sentences that can be said about an assembly *as a sequenced collection*. This interface works best on assemblies that grow and shrink; it is a rather dynamic flip-flop. The numbers fit very well, e.g. optimised for transmission sizes 3*4 and 65. Switching between sentences about an assembly as a contemporary collection, and sentences about it as a sequenced collection is like switching from the present progressive tense to the present perfect (Matsuno/Salthe 2002). It is of course very important to point out that this model is a mathematical, theoretical, abstract model and has as much similarity to actual genetics as trigonometry has to a church tower.

2. Linguistic Analysis of Scientific Sentences

We apply the thoughts of Wittgenstein and Carnap in the approach presented here (Wittgenstein 1973). The main idea is that assertions in any scientific investigations are logical sentences; that logical sentences have formal properties; and that logical sentences with formal properties can be schematized, brought into a general, uniform form --and thus can be systematically generated. Having generated all *idealized* logical sentences we shall have generated all formal sentences science can ever deal with.

In an ideal language, every result of a sentence refers to a logical entity. Why not give this entity a representation in N by a natural number? We simply *enumerate the concepts* and *state everything* possible about them. In practice, this means that we generate *all partitions* of *each natural number* (Javorszky 1985).

The concept of a numeric representation of a logical sentence goes back to the idea of Carnap that a formal language is used to speak about those states of the world about which we can speak correctly (Carnap 1937, 1954). If the thing we speak about is communicable at all, the words of the language we speak about it in will have a formal relation to the thing. In a semantic interpretation: if we say about something that it is made up of some parts, which have specific relations among each other, then we find such an element of N that the specific relations of the parts add up to a concept of the thing enumerated correctly. We simply select that natural number for which the stated relation of the parts will hold true.

Investigating each possible way for a set to be a collection of parts will give us the whole body of all possible sentences of science. We do not care, which of the combinations of the arguments will be useful, technically elegant or needed at all; that task is a question of application of the language while we discuss the grammar of the language. As we shall have created all sentences that can be said, the useful sentences will be among these.

Specifically, when talking about a cell, empirically oriented scientists (biologists, geneticists, etc.) say:

<n_1 parts of the cell> have property <s_1>;
<n_2 parts of the cell> have property <s_2>;
<n_3 parts of the cell> have property <s_3>;
….
<n_k parts of the cell> have property <s_k>;

where again

$n_1 + n_2 + n_3 + \ldots + n_k = n$

That is, insofar applied science talks correctly about a cell, it assigns some symbols to some constituents of the cell. We do not care *which* symbol to *how many* elements of the cell, we just note that each element has a symbol attached to it. The “meaning” of the symbol is of course relevant in the application of the findings, but not in the linguistic analysis

3. Concurrent Categorisations of Objects

We allow concurrent categorisations of objects of the set (Javorszky/Steidl 1998). We generate overlaps among the domains for which the validity of a symbol will hold true. That is, we generate sentences which will hold true in combinations with each other. The form of the sentences will get slightly more complicated as we allow concurrent assignments of symbols. We assume a vector or a matrix of symbols, among which applied science shall pick any combination it will have observed. The “columns” of the matrix of the symbols may refer to a property like “number of C atoms in the molecule” or “number of –

OH radicals in the molecule" or anything else applied science may find useful. The "rows" of the symbol matrix will contain the attributes within the category.[1]

The depth of the abstraction is of no relevance for the linguistic analysis. If the applied science uses the level of "subsystems" of the cell, then it will say

<n_1 subsystems of the cell> have property <s_1>;
<n_2 subsystems of the cell> have property <s_2>;
<n_3 subsystems of the cell> have property <s_3>;
….
<n_k subsystems of the cell> have property <s_k>;

where again

$n_1 + n_2 + n_3 + \ldots + n_k = n$

The picture will be the same if the applied science talks about "atoms" or "molecules" or "minimal units". The main point is, that applied science will assign symbols to objects, and this repeatedly. Each run of assignments (observations, for the applied science) will partition the set into subsets.

4. Multidimensional Partitions

Multidimensional partitions have been left undefined by mathematics so far. We propose to use the idea of concurrent assignments of symbols in order to be able to model the information carrying capacity of elements of a set in a commutative way of transmission of messages. Each "run" of assignment of symbols partitions the set into subsets that are mutually disjunct with respect to that "run". A different run – a "run" means one partition of the set – may or may not generate a collection of subsets that are entirely disjunct relative to a different run. Nonredundant runs are those where at least one subset will be generated that is disjunct relative to the other runs. *It follows that on a finite set only a finite number of nonredundant runs can be applied.*

5. Structured Set

The collection of partitions on a set shall be called the *structure of the set*. On a finite set only a finite number of distinct structures can exist. Structures can be distinguished among each other and this without regard to the nature of the symbols applied. The number of distinct structures on a finite set is dependent on the cardinality of the set. The upper limit for the number of distinct structures on a finite set with cardinality n is given by

n? = E(n)ln(E(n))**

where E(n) is the number of one-dimensional partitions of the natural number n (Javorszky 1995). By means of n objects carrying symbols one can transmit up to n? distinct messages as one can distinguish up to n? distinct structures on a set of n elements. The objects of the set need not be enumerated consecutively (this means that they need not to come in a sequence, that is, they may come all at once, "commutative information transmission").

[1] In an example, this matrix of symbols can be compared to the grades in subjects describing a school of n pupils. There are k courses taught at that school. Each individual pupil will receive a note in k subjects ("columns" of the symbol matrix) and shall possess 1 attribute in each category. The description of a pupil is a vector of attributes from k categories. The results in each course partition the total number n of students. We see several concurrent, (contemporal) partitions conducted on the set of students.

6. Sequential Information Storage

In case the n objects of the set are enumerated consecutively (this means they come in a sequence, the classical way of technical information transmission) they can carry up to n distinct symbols. In this case the information is contained in the sequence of the symbols. A set of n elements if enumerated consecutively can be in any of up to n! distinct states, where **n! = 2 * 3 * … * (n-1) * n**. This is the way information is transmitted in a technical environment and also by means of the DNA.

7. Commutative Information Storage

In case the n objects of the set are not enumerated consecutively (this means they do not come in a sequence, but come all at once, commutatively, in a cross-sectional way of information transmission) there is no restriction on the number nor on the kind of distinct symbols they can carry. In this case the information is contained in the structure of the set, generated by nonredundant symbols.

A set of n elements if enumerated commutatively can be in any of up to n? distinct states, where **n? = E(n)**ln(E(n))**. A symbol is redundant if it points out the same subset of elements a different symbol points out.[2]

8. Interplay between Sequential and Commutative Logical Sentences

A set can be enumerated both commutatively and consecutively at the same time. This appears to be the case in biologic information transfer. The same information is pointed out by a sequence (the DNA) and by a collection of objects which are commutative (happen in the same moment, are there all at the same time, are a temporal cross-section. This is the physiologic melée surrounding the DNA, the cell). One specific arrangement of symbols points out at the same time one specific realisation of up to n? potential cross-sectional arrangements (assemblies, structures, cells) while that same specific arrangement of symbols also points out one specific realisation of up to n! potential longitudinal arrangements (sequences, DNA-s). The relation between injective and surjective pictures between the enumerations (between the complete and the non-complete enumerations: that is, between the sequential and the commutative way of information storage) is dependent on the cardinality property of the set. That is, if the "cell grows", rather, the set increases its cardinality property, there will appear a restriction on the number of possible complete enumerations (sequences, DNAs) while if the "cell grows further" or "shrinks or splits", that is, the set's cardinality is outside some defined bounds, the restriction will apply to the number of possible incomplete enumerations (kinds of cells, physiological melées). This because of the numeric relation between n? and n!, which is as follows:

n	*max(n!, n?)*
1 <= n <= 31	n! > n?;
n = 32	n! ~ n?;
33 <= n <= 95	n! < n?;
n = 96	n! ~ n?;
97 <= n	n! > n?.

Table 1: Numeric relation between n? and n!

[2] If each pupil receives the same marks in subjects Q and W, then either Q or W are redundant (superfluous) as a progress indicator of the student.

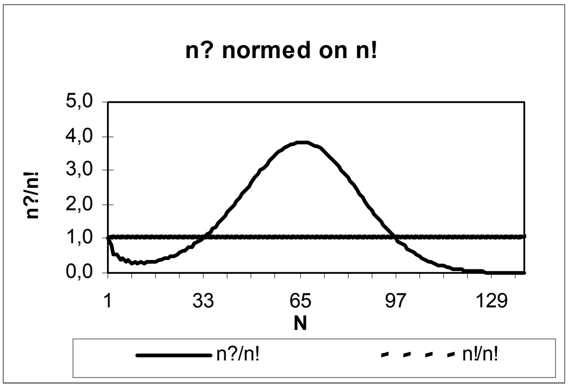

Figure 1. Number of distinct commutative states per number of distinct sequences

The maximum of n?/n! appears near cardinality 66 and reaches ~ 3.8. This means that one may transmit information up to 380% more efficiently (efficiency of information transmission: number of messages per number of carrier objects) relative to the Shannon algorithm if one uses the carrier objects commutatively. Further efficiency increases may be conceptualised by switching to the other extreme, near cardinality 11, where n!/n? is ~ 3.6. This appears to allow a flip-flop mechanism which packages and unpacks info with an elegance which far surpasses any technical approach known so far. One will remark also, that if the cardinality of the set is 135 or slightly above, the relative error between $n?/n!_{135}$ and $n?/n!_{136}$ allows for the "disappearance" of a carrier object. This means that one can keep as many congruence relations between complete and incomplete enumerations using 135 carrier objects as one has using 136 objects. This means that one has *either* one more carrier objects *or* one has a higher density of congruence relations. That the existence and the density of logical relations translates into the appearance or disappearance of a "physical" object may turn out to be a substantial contribution of information theory towards understanding physical and chemical phenomena (Javorszky 2000).

9. Measures of Diversity

Information transmission in biology appears to take place without too much regard to the absolute size of a system. So, the same techniques apply to systems small or big, growing or shrinking. We have shown that the cardinality property of a set is what determines whether the sequential or the contemporary assembly is constrained by the other one. As the copying to-and-fro happens both ways, we need a tool that works irrespective of absolute sizes. A tool fulfilling this requirement is our invention of System M, a recoding of the set N according to following rule[3]:

N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
M	0	-1	0	1	-1	0	1	-2	-1	0	1	2	-2	-1	0

We assign a specific M value to each element of N. Each M value can represent an infinite number of distinct elements of N. The translation of N data in M values serves the goal of making fragmentation information transportable across sizes. Due to the extremely asymmetric nature of the partition distribution, only specific combinations between the properties "number of subsets" and "cardinality" of the set have meaningful probabilities. Generally, throughout this model, we let loose of the "absolute size" (or cardinality) property of the set under discussion and focus rather on the recombinational probabilities of the subsets. This is like discussing the chances of winning in any card game, which can be established

[3] *function m(n); local integer tmp, first, len, zero, m; tmp = int(sqrt(n/2)); first = 2*tmp**2; len = 4*tmp + 2; zero = if(n-first < len/2, first + tmp, first + len/2 + tmp); m = n – zero; return(m).*

also rather independently of the number of the actual, physical cards used in each kind of games. One will note that any element of N has a maximal positive extent on M which is a mod(4) expression. This follows by setting n4 as m1, that is, the maximal (positive) extent of any N value is n/4 on M. The M → N translation brings us in a different way into contact with *four* logical markers (as used by the DNA): the interpretation of an M value on N allows us to choose among *four* series, of which *two and two* belong together, if we disregard the sign change.[4]

Linguistic analysis will show that in transmitting a message we point out two aspects of what is the case, namely a. how that what is the case is similar within itself and with respect to other things, and b. how that what is the case is dissimilar within itself and with respect to other things.

The invention of System M allows discussing doubly .t. sentences, as roughly 10 % of all partitions are .t. in M too. Each partition is of course .t., as it is of the form

$$n = \sum n_i.$$

If the rule

$$m(n) = \sum m(n_i)$$

is concurrently .t., we speak of .d. (doubly .t.) sentences[5].

Multidimensional partitions built on .2. sentences are of course .3., .4., etc. true. There is a convergence of the expected extent of .i. on the set of .d. (.t. level > .1., “doubly” or more .t.) sentences. Relative to the most likely extent of .i., logical sentences appear over- or underdetermined.

10. Double Sequence of Triplets

Regarding the DNA in an information-theoretical, or linguistic analysis context, one will conclude as follows: “the sequenced way of describing what is the case uses *four distinct* logical arguments” (Javorszky 1999). Finding now ways of building logical sentences which contain four distinct arguments, one investigates the table of partitions of n into 4 distinct summands. There, one finds when considering the jump in the increase E(n+1,4)-E(n,4) a sequence which runs like “333444555666….”. We propose to call this sequence H. This sequence can be split up into 2 interlocking parallel sequences, wherein H1 runs like 334556778… and H2 like 344566788… .

We interpret Sequence H as giving the position of a defined extent of an increase in recombinational probabilities within a set while this set is growing. It appears a formidable task to find a model that optimises information storage and transmission by using four distinct arguments. Possible approaches include: a) using the concept of a maximal positive extent of an N-expression in M-translation. This would allow treating any N-expression as a quasi-mod(4) extent by its maximal size on M; b) re-translating into N the M-translation of the structure of the set appears to offer *four* distinct logical groups of possible arguments. The M-form of a logical statement is invariant with respect to the absolute cardinality of the set in N; therefore one can discuss the same “form” of a structure of a set without being fixed to a specific “size” of the assembly. This means that irrespective of the size of the assembly, its information theoretical content remains virtually the same. It is just the same but bigger.

11. Applications

We propose three ways of a constructive outlook being made possible by understanding the interplay between transversal and longitudinal information storage.

- *efficiency increases* in information transmission in a mechanical, technical environment. Using pre-defined message carriers in clusters of ~ 66 objects that are treated as structure, and using the same assembly as ~ 6 sequences of 11 individual objects each should provide efficiency boosts of a significant magnitude. Sub-applications involve cryptography and dealing with alternatives (“thinking”, AI);

[4] E.g. $|m2|$: {8,19,34,53,… 12,23,38,57,…13,26,43,64,…, 17,30,47,68,…}

[5] It is interesting to work on the set of $m(n^2) = \sum m(n^2_i)$ sentences.

- *building a ganglion*, at first as a theoretical gadget. Having non-matching neighbourhoods creates (by properties of the M-translation) spikes of tension. The dissimilarity of subsets translates thus in a pattern of discharges. It appears possible to deduct from the rhythm of the discharges, what the ganglion has encountered. Maybe, one could understand the electric discharges of the nerve cell and the underlying biochemical processes as time-dependent and state-oriented polarisations of the same congruence;
- *logical archetypes* condense out from expected minimal densities of subsets of a structured set (Javorszky 2003). The combinations of properties of combinations of partitions can be seen as possessing differing probabilities of being present in a set. Those objects that are certainly there in a normally (averagely) structured set shall be termed multidimensional truth values, or logical archetypes. These are pictures of the entries in the periodic table of elements known from physics and chemistry. Building up a collection of subsets with defined logical properties allows assembling molecules.

References

Wittgenstein, L.udwig (1973) *Tractatus logico-philosophicus*. Frankfurt/Main. Suhrkamp.

Carnap, Rudolf (1937/1968) *Logische Syntax der Sprache*. Wien/New York. Springer, ^{2}1968; [English: *The Logical Syntax of Language*, London, 1937]

Carnap, Rudolf (1954) *Einfuehrung in die symbolische Logik mit besonderer Beruecksichtigung ihrer Anwendungen*. Wien. Springer [English: I*ntroduction to Symbolic Logic and Its Application*, Dover, New York, 1958]

Matsuno, Koichiro/Salthe, Stanley (2002) *The Origin and Development of Time*. In: International Journal of General Systems, 31 (4), pp. 377-393.

Javorszky, Karl (1985) *Biocybernetics: A Mathematical Model of the Memory*. Vienna. Eigenverlag.

Javorszky, Karl (1995) *Zaragoza Lectures on Granularity Algebra*. Salzburg. Mackinger-Verlag.

Javorszky, Karl (2000) *Interaction Between Sequences and Mixtures*. In: Journal of Theoretical Biology, 2000, 205, pp. 663-665.

Javorszky, Karl (1999) *Double Sequence of Triplet*s, www.bio.vu.nl/tmbm99/contributions.html

Javorszky, Karl/Steidl, Renate (1998) *Messages Transmission by Means of Counting States of Set,* fis.iguw.tuwien.ac.at/resources/preprints/karl_javorszky/MTCSS3a.doc

Javorszky, Karl (2003) *Information Processing in Auto-regulated Systems*. In: Entropy, 5(1), pp. 161-192.