Complexity, Coarse-Graining and Symbolic Description

Bailin Hao

SFI WORKING PAPER: 2006-10-037

SANTA FE INSTITUTE

SFI Working Paper

To appear in: *Crossroads. The Asia Pacific Center for Theoretical Physics (APCTP) Web Journal. http://crossroads.apctp.org*

Complexity, Coarse-Graining and Symbolic Description

Bailin Hao

The T-Life Research Center, Fudan University, Shanghai 200433, China

The Santa Fe Institute, Santa Fe, NM 87501, USA

The Institute of Theoretical Physics,

Academia Sinica, Beijing 100080, China

(Dated: September 20, 2006)

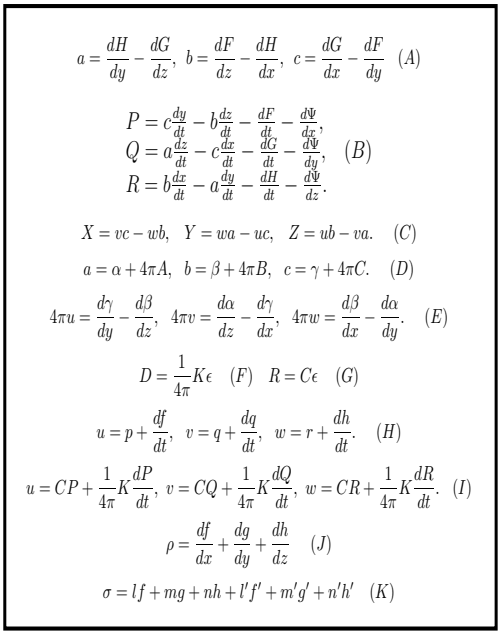
The entirety of human scientific knowledge may be thought as being embodied in an ever expanding spindle-shaped volume. The two sharp tips represent the well recognized frontiers of science – the microscopic world and the cosmos. For the time being, these two extremities span more than 35 orders of magnitude both in space and time ranging from cosmological scales of 10^{17} seconds (time elapsed since the Big Bang) and 10^{26} meters (the size of the observable universe) down to the atomic world measured in attoseconds (10^{-18}) and nanometers (10^{-9}). Scientific research aimed at pushing the two tips farther has become so expensive that it requires progressively more international collaboration; consequently, progressively fewer scientists are fortunate enough to be directly involved in these studies.

On the other hand, people often underestimate or even overlook the most extensive frontier of science, namely, the surface of the spindle far from the tips that mostly corresponds to the macroscopic world of our own size. In fact, human life on Earth depends on production and reproduction in the macroscopic world and the overwhelming majority of scientists work on problems confined to macroscopic space and time scales. If one asks what is the unifying theme of scientific research on macroscopic scales, perhaps, the answer lies in understanding complexity. Scientists from all walks of life agree that complex systems and complex behaviors exist everywhere in the macroscopic world, however, people disagree on the meaning of complexity. Intrinsic to the profound essence of the concept itself, complexity cannot be approached by applying a clear-cut definition. We are suspicious of the existence of a "science of complexity" or a universal measure of complexity. One must bear in mind that complexity goes along with specificity. Without setting a framework from the outset it is impossible to talk about complexity in general terms. This will become clearer when we look at biological symbolic sequences later on.

Before delving into the problem further I would like to make a few points based on common sense. Firstly, how do things get complex? Projection from a higher-dimensional space into a lower-dimensional space makes life more "complex". A simple example is putting a one-dimensional curve in a two-dimensional surface may cause self-intersections that cannot be removed by small deformations, but self-intersections of one-dimensional curves in a three- or higher-dimensional space are incidental and can be eliminated by infinitesimal deformations. In engineering practices one often has to add new coordinates or parameters to make things look simpler in the enlarged space. Repeated use of simple rules may produce complex behaviors or patterns as is evident in the iteration of the simple map $x_{n+1} = 1 - \mu x_n^2$

or applications of some nearest-neighbor rules in one-dimensional cellular automata. The use of the wrong reference system may also lead to a more complicated appearance of the phenomena, a well-known example being the geocentric system of Ptolemeaus versus the heliocentric system of Copernicus.

Thus we see the necessity of distinguishing objective complexity from the complicated way of describing the phenomena. Maxwell's equations of electromagnetism provides a sharp contrast. In his 1865 paper Maxwell introduced 20 equations for 20 variables without using vector notation. In Fig. 1 we copied the equations from Maxwell's *A Treatise on Electricity and Magnetism* (1873) keeping the original labeling by capital letters (all other formulas were simply numbered). Only a true genius could infer from these equations the existence of electromagnetic waves that propagate with the speed of light.



$$a = \frac{dH}{dy} - \frac{dG}{dz},\ b = \frac{dF}{dz} - \frac{dH}{dx},\ c = \frac{dG}{dx} - \frac{dF}{dy} \quad (A)$$

$$\begin{aligned} P &= c\tfrac{dy}{dt} - b\tfrac{dz}{dt} - \tfrac{dF}{dt} - \tfrac{d\Psi}{dx}, \\ Q &= a\tfrac{dz}{dt} - c\tfrac{dx}{dt} - \tfrac{dG}{dt} - \tfrac{d\Psi}{dy}, \\ R &= b\tfrac{dx}{dt} - a\tfrac{dy}{dt} - \tfrac{dH}{dt} - \tfrac{d\Psi}{dz}. \end{aligned} \quad (B)$$

$$X = vc - wb,\quad Y = wa - uc,\quad Z = ub - va. \quad (C)$$

$$a = \alpha + 4\pi A,\quad b = \beta + 4\pi B,\quad c = \gamma + 4\pi C. \quad (D)$$

$$4\pi u = \frac{d\gamma}{dy} - \frac{d\beta}{dz},\quad 4\pi v = \frac{d\alpha}{dz} - \frac{d\gamma}{dx},\quad 4\pi w = \frac{d\beta}{dx} - \frac{d\alpha}{dy}. \quad (E)$$

$$D = \frac{1}{4\pi} K\epsilon \quad (F) \quad R = C\epsilon \quad (G)$$

$$u = p + \frac{df}{dt},\quad v = q + \frac{dq}{dt},\quad w = r + \frac{dh}{dt}. \quad (H)$$

$$u = CP + \frac{1}{4\pi} K\frac{dP}{dt},\ v = CQ + \frac{1}{4\pi} K\frac{dQ}{dt},\ w = CR + \frac{1}{4\pi} K\frac{dR}{dt}. \quad (I)$$

$$\rho = \frac{df}{dx} + \frac{dg}{dy} + \frac{dh}{dz} \quad (J)$$

$$\sigma = lf + mg + nh + l'f' + m'g' + n'h' \quad (K)$$

FIG. 1: The Maxwell equations in 1865.

A hundred years after Maxwell these same equations may be written in one line as

$$\delta d\theta = J.$$

This equation utilizes 5 symbols including the equality sign and looks much "simpler" than the original system of Maxwell. However, in order to understand this relation one has to recognize d as the exterior differential, δ as the co-differential, and θ as a differential 1-form, that is, one must be acquainted with modern differential geometry. Things may look simple

if one stands on high ground. Phenomena that looked complex to primitive human beings may not be so for us. We should avoid any confusion caused by description and concentrate on complexity objectively.

In order to highlight one way of doing this, let us start with an observation.

A high energy physicist recognizes the following lower case letters u, d, c, s, b, t as quark names and associates them with a certain mass, charge and quantum numbers such as "charm" or "flavor". More scientists use the symbols p, n, e to denote proton, neutron and electron each having a certain mass, charge, spin or magnetic moment, but they are not concerned with from which three quarks a proton or neutron is made.

Chemists consider H, C, N, O, P, S to be element names and know their atomic number, ion radius, chemical valence and affinity. Chemical compounds may be denoted by combined use of such symbols as H_2O, NO, CO_2, etc. However, when it comes to writing chemical formulas for the nucleotides and amino acids which are the constituents of DNAs and proteins, there is no need to write down the tens of atomic symbols each time. Biochemists call the nucleotides a, c, g, t and the amino acids $A, C, \cdots, W, Y$. Now all one has to know is c and g are strongly conjugated by three hydrogen bonds while the weak coupling of a and t is made by two hydrogen bonds. Here "strong" and "weak" differ by many orders of magnitudes from that in high energy physics. In a biochemical pathway or a metabolic network, proteins/enzymes are denoted by simple names and there is no need to spell out the amino acids that make the proteins.

This list can be extended further. What is the morale learned from the observation? In describing Nature one cannot grasp the details on all levels at once; one has to concentrate on a particular level at a time treating larger scales as background and reflecting smaller scales in "parameters", etc. For example, in describing the Brownian motion of a pollen the environment at large is represented by a temperature while the friction force is given by using a coefficient of friction. If necessary, one could go down to the molecular level to calculate the coefficient directly. This is called coarse-grained description. Coarse-graining is reached by making "approximations", i.e., by ignoring details on finer scales. Nevertheless, it may lead to rigorous conclusions. Geoffrey West, the President of the Santa Fe Institute mentions that had Galileo be equipped with our high precision measuring instruments he would not be able to discover the law of free falling body and would have to write a 42-volume *Treatise on Falling Bodies*.

Furthermore, coarse-grained description of Nature is always associated with the use of symbols. If one is lucky enough these symbols may form symbolic sequences. Symbolic sequences from biology fall in this class. We understand primarily DNAs and proteins as biological sequences, both being one-dimensional, directed and unbranching hetero-polymers made of four and twenty kinds of monomers/letters, respectively.

Since we have come to the notion of symbolic sequences, let us recollect a basic fact on huge collections of symbolic sequences. In Claude Shannon's seminal 1948 paper that laid the foundation of modern information theory, besides the famous definition of information now familiar to all students, he stated a few other Theorems. Theorem 3 can be roughly interpreted as follows. Given a sequence of length N made of 0's and 1's, there are in total 2^N such sequences. Generally speaking, when N gets very large, these 2^N sequences can be divided into two subsets: a huge subset of "typical" sequences and a small group of "atypical" sequences. The statistical property of a typical sequence resembles that of any other typical sequence or the bulk of the huge group, while the property of any atypical sequence is very specific and has to be scrutinized almost individually. The simplest members of the atypical set are sequences made of N consecutive '1's or '0's as well as various kinds of periodic and quasi-periodic sequences. However, the most significant ones from the atypical set are those with hidden regularities mixed with seemingly random background. These are the true complex sequences we have to characterize.

Biological sequences are the result of several billions years of evolution and natural selection; they must belong to the set of atypical sequences in the space of all possible sequences of similar lengths. Due to the huge volume of data, the inevitably noisy background, and experimental errors, statistical tools should be invoked in the beginning of any analysis of biological sequences. However, one must rely on more "deterministic" approaches to reveal hidden regularities in the real data. In fact, by looking at real sequence data one may encounter surprises and discover peculiar features that cannot be seen in statistical studies alone. We show a few examples that have been unearthed from real bacterial genomes without the need of much biological knowledge as a prerequisite.

The first example we look at is the species-specific "avoidance signature" in bacteria genomes. Take for example, the genome of the harmless K12 strain of *E. coli*. This DNA loop is made of 4 639 675 letters of a, c, g and t. If one collects short strings of length 8 along the loop, shifting one letter at a time, one would collect 4 639 675 strings. However,

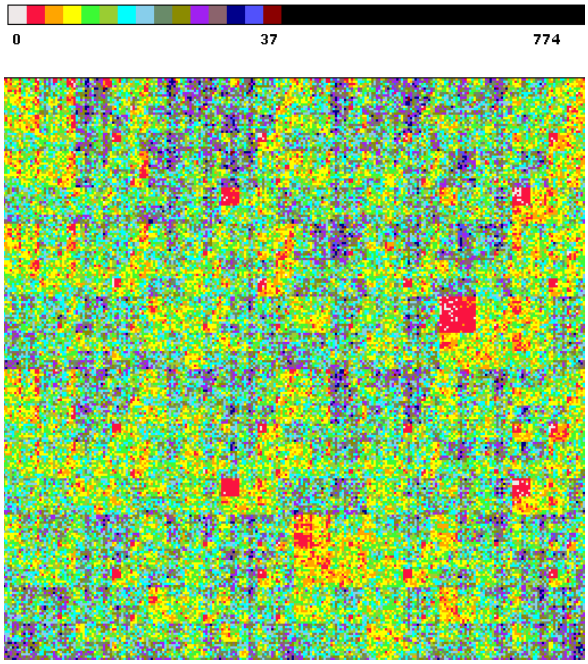


FIG. 2: A two-dimensional histogram of 8-strings in *E. coli*. Each element of this 256×256 square matrix represents the number of appearance of a string type with *gggggggg* in the upper-left corner, *cccccccc* in the upper-right corner and *aaaaaaaa* in the lower-left corner, etc.

these strings can only belong to $4^8 = 65\ 536$ different types. If the *E. coli* genome is a random sequence, each string type would appear $4639675/65536 \approx 71$ times. What happens in reality? Figure 2 is a two-dimensional histogram showing the number of appearances of all string types using a crude color code. In this square there are $256 \times 256 = 65\ 536$ cells each representing a counter for a string type. In order to highlight the missing and under-represented strings, white and bright colors are allocated to zero and small counts. Counts greater than a certain number are put in black. This is also a kind of coarse-graining that leads to the fairly regular patterns seen in the figure. It turns out that *E. coli* does not like strings containing *ctag* as a substring. If one looks at similar "portraits" of other closely related species, e.g., *Samonella* or *Shigella*, they all share this common feature. This might be a reminiscence of the environment when their common ancestors tried hard to avoid the *ctag*-containing recognition site of a "restriction enzyme" produced by their enemy. Different bacteria have different "avoidance signature", but some do not.

If one collects all string counts and draws a one-dimensional histogram by putting the counts along the abscissa and the number of string types in a small bin of counts along the ordinate, one would get an almost continuous distribution biased towards small counts (there are 176 missing 8-string types) with a long tail up to count 777. If one "randomizes" the original genome and does the drawing again, one would get a bell-shaped curve centered around the average value 71. At length 7 there is only one missing string, namely, *gcctagg*. Among the 176 missing strings 8 must be the consequence of the missing *gcctagg*. Therefore, only 168 are "true" missing ones. This observation raises a seemingly simple question: suppose at length $K + 0$ one string is missing, how many strings would it take away at length $K + i$? Simple induction would lead to an answer $4^i(i + 1)$. However, this is only an approximate result, since it ignores the overlap of the leading and ending letters in the string, as was the case with *gcctagg*. The exact answer to this biology-inspired problem leads to a neat piece of mathematics which makes use of combinatorics or language theory.

The second example is the fine structure in the number distribution of K-strings of some *randomized* bacterial genomes. The aforementioned shape of the one-dimensional histograms for the original and randomized *E. coli* genome seems to be natural. Now, let us take another bacterium, say, the *M. tuberculosis* genome – the one-dimensional histogram resembles *E. coli*, that is a continuous distribution biased towards small counts. However, if we randomize this genome, keeping the number of a, c, g and t unchanged, a strange thing happens. See Figure 3.

A fine structure appears in the one-dimensional histogram: instead of a continuous curve we see several peaks. This seems anti-intuitive at first glance as we have got used to the fact that structures usually disappear after randomization. However, a little reflection would tell us that this was caused by the uneven distribution of the nucleotides in this genome: the letters g and c make up almost $2/3$ of the genome. When forming various 8-strings, one simply does not have enough letters for some string types. For a fixed K there are $K + 1$ peaks, each being described by a Poisson distribution. The location and the parameter of the distribution may be calculated precisely.

The avoidance signature in a two-dimensional histogram and the fine structure in randomized one-dimensional histogram reflect what is absent in a genome. The following example deals with what is present in a genome.

The third example comes from species-specific short repeats in bacterial genomes. When

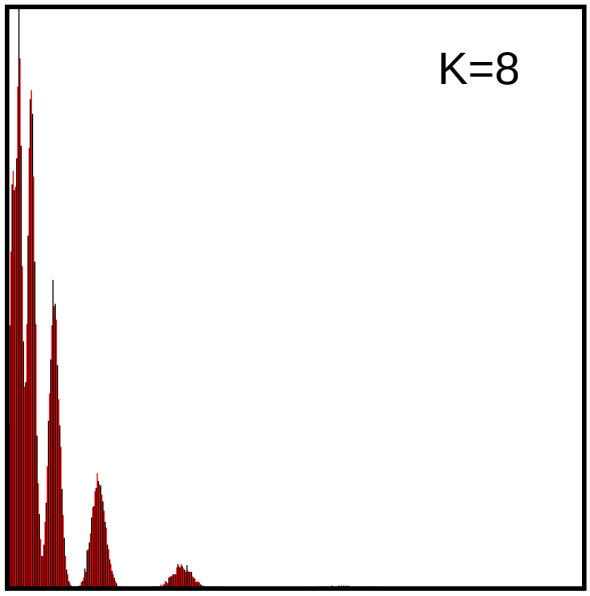


FIG. 3: A 1D histogram of 8-strings in a randomized *M. tuberculosis* genome. The counts are given along the abscissa, the ordinate being the number of string types in a bin. Since we draw attention to the appearance of peaks, no numbers were put along the axes.

studying a genomic sequence by looking at its K-string composition, we see a transition from randomness to some "determinism" with K increasing. The distribution of single letters ($K = 1$) or dinucleotides ($K = 2$) are almost random, while at the longer K more and more species-specific features appear. In particular, there are species-specific repeated segments in some bacterial genomes. For example, the 25-string

aaatcagaccaaaatgggattgaaa

has 107 copies in the *A. fulgidus* genome and 171 copies in the *M. thermoautotrophicus* genome. It is unique in these two genomes. No single copy of such string may be found in all other DNA sequences known so far. Their function remains unknown and existing annotations are questionable.

Another 18-string *gttccaataagactaaaa* exists as repeats in the genomes of three known species from one and the same genus *Pyrococcus* with no appearance in other bacterial genomes. It may serve as a genus marker for *Pyrococcus*. There are many other taxon-specific repeats in bacteria genomes. We emphasize repeats because they are less affected by individual differences or sequencing errors. Plenty of them exist if we confine ourselves to single copy specific segments. However, they may not be robust enough to deserve special

scrutiny.

So far, we have looked at DNA sequences. Our last example deals with protein sequence. Let us look at the following piece of the winter flounder antifreeze protein:

$$MALSLFTVGQ\ LIFLFWTMRI\ TEASPDPAAK\ AAPAAAAAPA$$
$$AAAPDTASDA\ AAAAALTAAN\ AKAAAELTAA\ NAAAAAAATA$$
$$RG$$

It is an alanine(A)-rich protein made of 82 amino acids. Let us decompose this sequence into a collection of overlapping 5-strings (penta-peptides): $MALSL$, $ALSLF$, etc. One gets 78 such strings some of which appear several times. Now we ask the converse: given the collection of these 5-strings, if we reconstruct a sequence by using each penta-peptide once and only once, how unique would the construction be? The inverse problem is solvable because at least one can get the original protein. Obviously, when K is big enough, the reconstruction is unique. This problem has a natural relation to a well-known problem in graph theory, namely, the number of so-called Eulerian loops in a graph. Indeed, by treating 5-string $MALSL$ as a transition from $\boxed{MALS}$ to $\boxed{ALSL}$, etc., and simplifying what we obtained by reducing elements that do not affect the number of Eulerian loops, we get the graph shown in Figure 4.

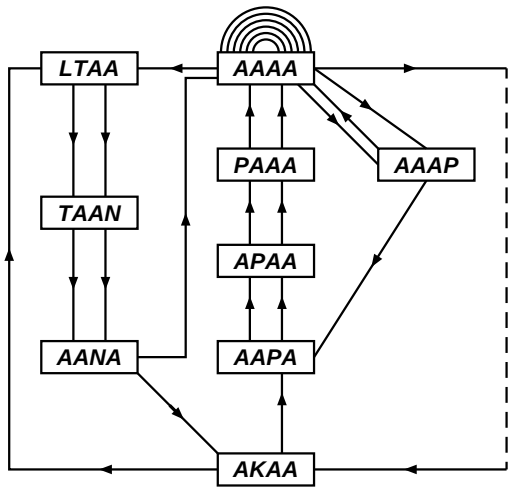


FIG. 4: Euler graph determined by the winter flounder antifreeze protein.

This protein sequence has 1512 different reconstructions at $K = 5$, 60 at $K = 6$, 2 at $K = 7$ and an unique reconstruction at $K = 8$. Most of naturally occurring proteins have

unique construction at $K = 5$ or 6. It has been proved recently that there exist finite-state automata which can recognize whether a given sequence has an unique reconstruction at a given K. The uniqueness problem arises from an attempt to justify a recently proposed method to infer bacterial phylogenetic relationships from the complete genomes that is different from the traditional sequence-alignment methods. It is interesting to note that equipped with the automata just mentioned, one could discover a few proteins with a huge number of reconstructions at moderate K from a big protein database without possessing any further biological knowledge. These proteins deserve further study.

At present the amount of known DNA and protein sequences grows quickly. These sequences are "atypical" according to Shannon's theorem. On one hand, they should be studied almost individually to reveal any common regularities. This is what bioinformatics does. On the other hand, many questions may be asked by inspecting real data and interesting mathematical problems may be posed and solved. In the above examples, we have made connection to combinatorics, graph theory, language and automaton theory, as well as Poisson distributions. The possible width and depth of these biology-inspired problems remind us once more that complexity goes along with specificity. Common features of complex phenomena can only be inferred from exploring the specific richness of real data.

Acknowledgement. The author thanks the Santa Fe Institute for invitation and support. It was a great pleasure to write this essay in the inspiring atmosphere of SFI.

********** ********** ********** ********** **********

Postscript

Complying with the style of *Crossroads* no references were given in the manuscript. In order to help possible readers of *SFI Working Papers* we add the following reference list. Most of the PDFs may be downloaded from Hao's webpage:
`http://www.itp.ac.cn/~hao/`

1. Bailin Hao, Hoong-Chien Lee, and Shuyu Zhang, Fractals related to long DNA sequences and complete genomes, *Chaos, Solitons and Fractals* **11** (2000) 825 – 836.
2. Bailin Hao, Fractals from genomes — exact solutions of a biology-inspired problem, *Physica* **A282** (2000) 225 – 246.

3. Bailin Hao, Huimin Xie, Zuguo Yu, and Guoyi Chen, "Factorisable language: from dynamics to complete genomes", *Physica* **A288** (2000) 10 – 20.

4. Bailin Hao, Huimin Xie, and Shuyu Zhang, Compositional representation of protein sequences and the number of Eulerian loops, arXiv: physics/0103028 (10 March 2001).

5. Huimin Xie and Bailin Hao, Visualization of K-tuple distribution in prokaryote complete genomes and their randomized counterparts, in *Proceedings of CSB2002 Bioinformatics Conference*, IEEE Computer Soceity, Los Alamitos, CA, 31 – 32.

6. Junjie Shen, Shuyu Zhang, Hoong-Chien Lee, and Bailin Hao, SeeDNA: a visualization tool for K-string content of long DNA sequences and their randomized counterparts, *Genomics, Proteomics & Bioinformatics* **2**(3) (2004) 192 – 196.
   The source code `SeeDNA.tar.gz` is downloadable from Hao's website.

7. Bailin Hao, Ji Qi, Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance, *J. Bioinformatics & Computational Biology*, **2** (2004), 1 – 19.

8. Qiang Li and Huimin Xie, Finite automata for testing uniqueness of Eulerian trails, arXiv: cs.CC/0507052 (20 July 2005).

9. Li Xia and Chan Zhou, Phase transition in sequence unique reconstruction, accepted for publication by *J. Systems Sci. & Comlexity*, 2006.

10. Bailin Hao, Huimin Xie, Factorizable language revisited: from dynamics to biology, a brief review submitted to *Int. J. Modern Phys.* **B**, 2006.

11. Xiaoli Shi, Huimin Xie, Shuyu Zhang, and Bailin Hao, Decomposition and reconstruction of protein sequences: the problem of uniqueness and factorizable language, submitted to *J. Korean Physical Soceity*, 2006.