



# Functional labels and syntactic entropy on DNA strings and proteins

A. Carbone<sup>a,b,\*</sup>, M. Gromov<sup>a,c</sup>

<sup>a</sup>*Institut des Hautes Études Scientifiques, 35, route de Chartres, 91440 Bures-sur-Yvette, France*

<sup>b</sup>*Laboratoire d'Algorithmique, Complexité et Logique, Université de Paris 12, Paris, France*

<sup>c</sup>*Courant Institute of Mathematical Science, New York University, New York, USA*

---

## Abstract

The DNA of a cell is an object which admits a simple mathematical description and a convenient representation in a computer (it is given by an easily manipulatable list, a finite sequence in four letters typically of length between one million and 10 billions). In contrast to this, there is no simple way of describing the cell neither statically and even less temporally (dynamically). We shall indicate here a possible formalism of combinatorial and numerical (entropic) structures on spaces of sequences which reflect, up to some degree, the organization and functions of DNA and proteins.

© 2002 Elsevier Science B.V. All rights reserved.

*Keywords:* DNA sequences; Genes; Proteins; Macromolecules; Protein networks; Gene networks; Bioplexi; Genoplexi; Hypergraphs; Open label structure; Evolution; Entropy; Coentropy

---

## 1. A view on DNA out of the cell

Let  $\mathcal{S}_i$  be the space of sequences in the letters  $A, T, C, G$  of length  $i$  and let  $\mathcal{S} = \cup \mathcal{S}_i$  be the disjoint union of the spaces of sequences of length  $1, 2, 3, \dots$ . Denote by  $E$  the *environmental parameter space* (of the cell), represented as  $E = [0, 1]^d$  where  $d$  is of order  $10^2 - 10^4$  for unicellular organisms.

**Remark.** The individual parameters in  $E$  represent the temperature, pH content, the concentrations of particular chemical compounds, etc. We exclude the interaction of

---

\* Corresponding author. Institut des Hautes Études Scientifiques, 35, route de Chartres, F-91440 Bures-sur-Yvette, France.

*E-mail addresses:* [carbone@ihes.fr](mailto:carbone@ihes.fr) (A. Carbone), [gromov@ihes.fr](mailto:gromov@ihes.fr) (M. Gromov).

cells with macromolecules, such as proteins, which is unavoidable in true biological situations, especially for a cell being a member of a multicellular organism. In fact, if we allow proteins in the environment, the potential number of different species of molecules becomes exponentially large, of order  $20^{300}$ .<sup>1</sup> The drastic difference between the information content of the environment for a free living cell and a member of a multicellular organism is consistent with the physiology. A significant part of the genome of a multicellular organism, and to a lesser extent in microbes, is occupied by genes involved in *signalling pathways* responsible for cell interaction with intercellular proteins (these are proteins that are secreted by one cell and influence another cell). The formalism which we sketch below is essentially motivated by the structure of imaginary bacteria living in a free environment, not in another organism and not interacting with other cells and viruses.

*Fitness* is a function  $F: \mathcal{S} \times E \rightarrow \mathbb{R}$ , where the value  $F(s, e)$  represents the viability of a cell with a genome  $s$  in the environment  $e$ . Besides sheer viability, one is concerned with how a cell interacts with the environment. For example, how a bacterium metabolizes a particular compound. The definition of fitness can be adjusted to this, and in what follows it can be understood in either of the two ways. Also notice that, despite our notation, the sets  $S$  and  $E$  are distinct in nature and the behavior of  $F$  with respect to  $S$  and  $E$  is quite different.

**Main conjecture of computational genetics:**  $F$  can be approximated by a function of reasonable complexity. More pragmatically, one can design experiments that could be implemented within several decades, say 50 years, such that the results can be encoded in a database  $\mathcal{D}$  of order  $10^{10}$ – $10^{13}$  bytes, and such that, using  $\mathcal{D}$  one can design a feasible<sup>2</sup> algorithm for computing  $F$ .

### 1.1. Remarks and explanations

0. The objects we introduced,  $F, E, \mathcal{D}$ , have several faces: they may be quantities found by (potential) experiments, defined by their formal mathematical properties, or computed by some algorithm. The choice among the three is usually clear from the context.

1. A biologist is usually concerned with DNA sequences of length between  $10^6$  and  $10^{10}$ ,<sup>3</sup> thus the relevant  $\mathcal{S}$  is large but still a finite space. On the other hand,  $E$  being a continuum, appears infinite. However, the relevant ranges of values of parameters are rather small, in practice, of order of 10 for suitable choices of parameterization. Thus,  $E$  may have cardinality  $10^{100}$  which is much smaller than the cardinality  $4^{10^6}$  of the space of smallest genomes.

<sup>1</sup> A protein is encoded by a sequence of 20 amino-acids with an average length  $\approx 300$ .

<sup>2</sup> The word “feasible” refers to a realistic time measured in minutes, hours or weeks at most.

<sup>3</sup> The smallest viruses have genomes of 1200 base pairs (bps), the smallest bacteria such as *Mycoplasma genitalium* have less than half a million bps, the most studied bacterium, *Escherichia coli* has  $4.6 \times 10^6$  bps, the second most studied mammal, *Homo sapiens*, has DNA of about  $3 \times 10^9$  bps, and some genomes, e.g. of the lungfish, have more than  $10^{11}$  bps.

2. Customary  $F$  represents the reproduction rate of a cell. A rougher  $F$  would be a 0, 1-valued function expressing the idea of existence of a viable (alive) cell in a given environment. Thus,  $F$  is represented by the subset  $\mathcal{C}_* = \mathcal{C}_F \subset \mathcal{S} \times E$  of functional cells, where  $F(s, e) = 1$ . The subset  $\mathcal{C}_*$  is the union of the subsets  $\mathcal{C}_e \subset \mathcal{S}$  corresponding to genomes viable in the environment  $e$ .

3. The algorithm in the conjecture is supposed to be neither deterministic nor sharp. It may occasionally fail and even when it works, it might give only an approximate value of  $F$ .

4. Designing experiments and the software for constructing genomes and predicting the properties of the corresponding organisms makes a large part of the ongoing project in genomics. The conjecture is an abstraction of this program.

5. Denote by  $\mathcal{G} \subset \mathcal{S}$  the set of genomes found in living organisms. This set is much smaller than  $\mathcal{S}$  as it contains, by current estimates, at most  $10^8$  (essentially) different sequences, corresponding to different species of organisms, including bacteria that make this bound so large. In formulating the conjecture, one should limit  $\mathcal{S}$  to a certain “neighborhood”, or to an extension  $\mathcal{G}_* \supset \mathcal{G}$  in  $\mathcal{S}$  consisting of sequences reachable from  $\mathcal{G}$  by feasible chains of mutations where all intermediate genomes are viable.

6. This conjecture is not Popperian: even if stated in the most refined form, it is not falsifiable unless one admits an unfeasible search for all possible programs of a certain length. The positive solution, if a solution exists, will be found in many intermediate steps guided by sub-conjectures: our purpose is to contribute to the development of tools for the formal analysis of sequences of such steps.

7. We believe that in order to settle the conjecture, one needs a mathematical interface between experimental biology and computational biology (bio-informatics). Identifying formal mathematical structures encoded in our  $F$  and  $\mathcal{C}_*$  may facilitate the design of the algorithm and also may help biologists in the design of relevant experiments.

The purpose of this article is to give an outline of possible mathematical structures that could be used for designing an interface between bench and computer experiments. In order to be concise, we are faithful neither to the biological reality nor to the mathematical rigor. In future articles we shall give a more rigorous presentation better connected to biophysical features of the cellular chemo-architecture, such as folding and self-assembly of macromolecules, chemical kinetics and enzymatic activities. Also, we shall try to formalize more elaborate aspects of the chemo-architecture such as the organization of membranes and compartments of the cell that can be (partially) captured by the combinatorics in the sequence space rather than by a straightforward geometrical and physical description.

## 2. Combinatorial and open label structures in genomes

*Motivations:* Let a genome of an organism be represented by a long sequence (string) of letters  $A, C, T, G$ . There are biologically significant substrings  $\sigma$  within this long sequence such as protein encoding genes and regulatory genes. We think of such a

feature as a member of an abstract set (possibly equipped with some mathematical structure) and we regard a statement such as “ $\sigma$  is a gene” as a function assigning the label “gene” to the segment  $\sigma$ . The distinguished segments, such as genes, can be related to each other in a variety of ways. For example, several exons may be contained in the same gene or several genes may be encoding proteins involved in a common function in the cell. Thus, we distinguish specific subsets of segments and assign to such a subset a label(s) indicating the nature of the mutual relationship between the segments within the subset.

*Definition of a genoplex:* A *genoplex* (a network of genetic links)  $\Sigma$  is given by the following data:

- (1) a finite collection, also called  $\Sigma$ , of *strings*,  $\sigma_1, \dots, \sigma_k$ , where each  $\sigma_i$  is a finite sequence of letters  $A, T, C, G$ ;
- (2) a set  $\mathcal{L}$  of objects, called *labels*, divided in several classes called *types of labels*; the number of types  $\tau$  will be few and in what follows we shall distinguish two types of labels: *positional* labels and *functional* labels. The set of labels within each type is not a priori specified and this is especially important for functional labels as they correspond to certain functions performed in the cell, some of which may yet be unknown.
- (3) for each type  $\tau$ , there is a subset  $A_\tau \subset 2^\Sigma$  and a function assigning to each  $L \in A_\tau$  a label  $\mathbf{L} \in \mathcal{L}$  of type  $\tau$ . The set  $A_\tau$  is called *domain* of the labelling and the function  $A_\tau \rightarrow \mathcal{L}$  is called it  $\tau$ -*labelling*.

Recall that  $2^\Sigma$  denotes the set of all subsets in  $\Sigma$  and it has exponential cardinality, that is  $\text{card}(2^\Sigma) = 2^{\text{card}(\Sigma)}$ . On the other hand, in a biological context,  $A_\tau$  is a relatively small subset of at most polynomial size in  $2^\Sigma$ . This represents the commonly accepted idea that the number of different biologically significant functions in the cell is relatively small (at most polynomial in  $\text{card}(\Sigma)$ ), otherwise the biological information could not be represented by a database of a feasible size without introducing radically new concepts.

*Hypergraph structures and open label structures in a genoplex:* The “combinatorial skeleton” of a genoplex is represented by several hypergraph structures on the set of strings, that is by a  $A_\tau$  for all types  $\tau$ . Recall that a *hypergraph structure* on a set  $\Sigma$  of *vertices*  $\sigma \in \Sigma$  is given by distinguishing certain subsets in  $A \subset 2^\Sigma$ , where the subsets  $L \in A$  are called *hyperedges*, or *links* in our context.<sup>4</sup> This reduces to a graph structure if the cardinality of the subsets equals 2. Observe that every hypergraph can be represented by a (bipartite) graph where the vertices are colored by two colors, say black and white, and where the black vertices correspond to the vertices of the hypergraph and the white ones represent the hyperedges. A black and a white vertices are joined by an edge if the black vertex in the hypergraph is in the hyperedge represented by the white vertex. This may be useful for visual representation of an abstract hypergraph but we do not use it in the paper.

The second combinatorial component of a genoplex, distinguishing it from a plane hypergraph (or several hypergraph structures on a given set), is the representation of

<sup>4</sup> In the present context, every vertex  $\sigma \in \Sigma$  is regarded as a link.

the vertices of the hypergraphs by strings of symbols in a given alphabet, that is  $A, C, T, G$  in the present case.

Labels are not so formally defined and each of them represents a specific information attached to every link. The set of labels makes a universe on its own independent of a particular genoplex. Eventually one wants to find a mathematical structure in the set of biological labels, as complete as possible, and to bring genoplexi associated to the cells maximally close to formal mathematical objects. In the present stage, neither the full set of labels nor the relations between them (compare examples in c below) are specified. One is open to introducing new labels and relations between them. This motivates the “open” terminology.

*Apology:* It seems impossible to follow the mathematical tradition of introducing all fundamental concepts at the beginning, as this depends on the current state of knowledge derived from the experiments. Our logic is rather similar to the logic behind the development of a certain software, where we allow the introduction of new objects, i.e. labels, and new logical connections in the course of arrival of new information. The combinatorics of the resulting logical graph is not necessarily constrained by common mathematical requirements such as connectivity, consistency, etc. The overall structure contains a well-defined mathematical core as well as less sharply defined chemical, genetic and biological data, with a continuous flow of information between the three empirical components coordinated by the core.

*Examples of genoplexi:* (a) Let us give a more detailed description of the genoplex associated to a given genome:

- (1) *segmentation* of the genome: we distinguish certain segments, the strings of the genoplex; in the simplest case, it consists of dividing the full genome sequence into non-overlapping segments of biological significance such as genes,<sup>5</sup> regulatory regions, possibly exons and at times introns, etc.;
- (2) *functional labelling:* the links *connect* (correspond to groups of) segments involved in common functions in the cell, with the labels describing these functions;
- (3) *positional labelling:* this specifies the mutual positions of the segments in the genome. A label may say, for example, that “ $\sigma_1$  is contained in  $\sigma_2$ ”, or that “ $\sigma_1$  lies a certain number of base pairs upstream of  $\sigma_2$ ”. The choice of a particular positional information depends on the problem at hand. In what follows, we suppress this labelling for the simplicity of the exposition, but whenever needed the information encoded by this structure can be brought back in an obvious way.
- (4) *other types of labels:* as an example of these, we indicate “ $\sigma_1, \sigma_2, \dots$  are homologous” and “DNA segments corresponding to  $\sigma_1, \sigma_2$  are spatially close in a chromosome”.

(b) *A virus–bacterium system,* can be regarded as a genoplex where the relevant set of strings consists of the protein coding genes of the virus and of the part of the bacterial genome encoding proteins relevant for the virus life cycle.

(c) *An organism which is genetically modified by introducing several genes.* The pertinent genoplex consists of these genes and the original genes of the organism which directly interact with the newly introduced genes.

<sup>5</sup> Coding sequences may overlap. This is rather common in viruses.

*Examples of labels on strings*

- “ $\sigma$  is a gene”,
- “ $\sigma$  is a protein coding gene”,
- “ $\sigma$  is a gene encoding a ribosomal protein”.

The above indicates possible structures in the space of labels; for instance, the three labels are linearly ordered in the obvious way.

*Examples of labels on links*

- “ $\sigma_1$  is a regulatory region for the gene  $\sigma_2$ ”, for the 2-link  $\tau = \{\sigma_1, \sigma_2\}$ ;
- “the proteins encoded by  $\sigma_1$  and  $\sigma_2$  make a dimer”;
- “ $\sigma_1, \sigma_2, \dots, \sigma_k$  represent binding sites involved in the regulation of the same gene”;
- “the proteins associated to genes  $\sigma_1, \sigma_2, \dots, \sigma_k$  are involved in a specified metabolic process”.

**Remark.** A customary way in biology is to speak about *graphs* or *networks* relating genes and/or proteins produced by these genes. Among them, one distinguishes the *protein-protein binding graphs*, and *gene-gene regulatory graphs*. A priori, graphs have more compact representation than hypergraphs since a graph on  $k$  strings is given by at most  $k^2$  non-zero entries in the incidence matrix while the full hypergraph may require up to  $2^k$  entries. However, only relatively few links will enter the hypergraph with non-empty labels, and moreover the sets of links and labels carry additional structures that eventually allow a compression of the representation of biological genoplexi. A simple example is a simplicial complex of such a structure, that is a hypergraph where every subset of a link is again a link. Here, one only needs the simplices of the maximal dimension as their faces are automatically in the hypergraph. Whenever such a situation arises in biology, namely if we are only concerned with the maximal set of genes/proteins involved in a given function, we do not spend time and space in enumerating all subsets of this set unless there is a special reason for that.

There is a simple dictionary translating from “graphs” to “hypergraphs”; choosing the particular language depends on the suitability for the problem at hand. The links of our hypergraph typically represent clusters of genes and proteins involved in the same function. Such a hypergraph may be, sometimes but not always, formally derived from the underlying graph structure, e.g. where the clusters appear as connected components of the graph.

*What genoplexi might be good for:* The combinatorial structure of a genoplex (encoded in a hypergraph structure and/or in the combinatorial structure of the space of labels) mediates between the syntactic structure of the strings and the biological information carried by the labels. Ideally we want the “category of genoplexi”<sup>6</sup> to adequately approximate the “category of biological systems”; also, ideally, the combinatorics of genoplexi should be *uniquely* determined by the syntactic content of the constituent strings. Granted this, one could explicitly relate the genotypes and phenotypes of organisms.

<sup>6</sup> Here, we refer to categories as understood in abstract algebra or to structures in a similar spirit.

A more modest goal is to express the bulk of biological constraints imposed by functional requirements of an organism in terms of the genomic sequences. We think of biological functions as “equations” imposed on genomic sequences and suggest a *quantitative* approach to these equations in Section 5.

### 3. Collective label structures in $\mathcal{S}$

Unlike an individual sequence (string), the space  $\mathcal{S}$  of all sequences carries a variety of intrinsically defined combinatorial structures. Moreover, these structures are enhanced by biological labels attached to them. They are called *collective* as they involve interrelations between different genome sequences.

*Point substitution: deletion and insertion.* The space  $\mathcal{S}$  naturally serves as the vertex set of the graph where edges correspond to (all possible) *point mutations*; a point mutation is a deletion, insertion or substitution of a single letter into a sequence. We extend our usage of labels to the present context and consider the following kinds of labels associated to the edges:

- (1) Type of the mutation: “deletion”, “insertion”, “substitution”.
- (2) Syntactic content of the mutation, e.g. a letter  $A$  is substituted by  $G$ .
- (3) Position of a mutation.
- (4) Probability of mutation; this is a number between 0 and 1 expressing the probability of this mutation per generation.

The above (1)–(4) define a mathematical structure in  $\mathcal{S}$ , the *syntactic label structure*. The following type of labels is of biological nature, and hence not formal:

- (5) Basic physiological effects of the mutation: “neutral mutation” which means that it does not significantly change the function  $F(s, e)$  as for  $s$  mutated to  $s'$ ; “lethal mutation” which makes  $F=0$  for all  $e \in E$ ; “advantageous (disadvantageous) mutation” which increases (decreases) the value of  $F$  (since  $F$  depends on  $E$ , one should specify the range of the parameters where  $F$  increases (decreases)); “unknown mutation” where the information on the effect of the mutation is not available.

The probability of occurrence of two consecutive point mutations is assumed to be the product of the two of them. However, a mutation of a segment, may have much higher probability than the product of the probabilities of mutation of its constituent letters (e.g. due to recombination and horizontal transfer of genes). Because of this, one distinguishes segment mutation.

- (6) *Segment mutation*: in the course of this mutation a given segment may disappear, it may double, it may invert, it may interchange the location with another segment, or it may appear in several copies in a tandem. Each of these characteristics is viewed as a label attached to the edge of the corresponding graph structure on  $\mathcal{S}$ . Also, we assign positional, functional and probabilistic labels as for the point mutations, and we consider insertion of segments coming from other genomes (it may happen naturally, e.g. through viruses, or artificially by means of genetic engineering).

#### 4. Relations between the structures

Let us explain how the segmentation structure can be derived from the labelled graph structure on  $\mathcal{S}$  by identifying extremal (minimal or maximal) “significant” segments in a genome. The word “segment” may have two meanings: (1) *content-segment*, earlier referred to as a string, that is a sequence of letters of relatively short length, (2) *position-segment*, that is a subsegment in a longer sequence where one forgets the letter content of the subsegment and remembers only its position.

Given a content-segment, we consider possible *insertions* of it in all sequences  $s \in \mathcal{S}$  and see what kind of changes this makes in the values of the fitness potentials at  $\mathcal{S}$ . Here are several possibilities:

- (1) All *fitness potentials*, that are  $F_e(s) = F(s, e)$  for  $e$  running over  $E$ , change very little, for all, most or many genomes of viable cells.
- (2) For all, most or many genomes, the change is lethal.
- (3) The values of some fitness potentials change strongly without being lethal, for all, most or many genomes.

The second and third possibility, is referred to as a “significant” change. In what follows, we want to pinpoint significant content-segments which bring non-lethal significant changes. Given a content-segment  $\sigma$ , let  $\Delta F(\sigma) = \Delta F_e(\sigma, G, p)$  denote the variation of the fitness potential(s) of the genome  $G$  when  $\sigma$  is inserted at the position  $p$ . The segment  $\sigma$  is considered to be “significant” if  $\Delta'F(\sigma) = \Delta F(\sigma) - \Delta F(\sigma')$  is large for random perturbations  $\sigma'$  of  $\sigma$ .<sup>7</sup>

There are many unavoidable ambiguities in this definition, such as the choices of  $e, G, p$  and the notion of “random perturbation”. Let us explain possibilities for the latter. It may be a replacement of  $\sigma$  by a totally random  $\sigma'$  having nothing to do with  $\sigma$ ; another possibility is a random modification of a small number of letters in  $\sigma$ ; or it can be a random modification of a certain percentage of the letters. In any case,  $\Delta'F(\sigma)$  appears as a random variable and its largeness should refer to a suitably chosen expectation value of this variable.

One can think that significant segments are those which are recognized by cells as meaningful words (sentences), when inserted in the genome of the cell.<sup>8</sup>

Now, let us see what can happen when we *remove* a segment  $\sigma$  from a given genome, thought of as a window in the genome sequence. Here,  $\Delta'F(\sigma)$  refers to changes of  $\Delta F$  (inflicted by the removal of  $\sigma$ ) which occur when we replace  $\sigma$  by a nearby segment  $\sigma'$  (i.e.  $\sigma'$  is obtained by a small sliding, stretching or shrinking of the window). Another possibility leading to essentially the same picture appears when we make a random modification of  $\sigma$  rather than removing it from the genome. The (position of)  $\sigma$  is called *significant*, if  $\Delta'F(\sigma)$  is large.

<sup>7</sup> The case of lethal segments needs a somewhat different treatment due to the binarity of death versus alive alternatives that should be augmented by some measure of “lethality”.

<sup>8</sup> The classical Shannon information theory does not distinguish “meaningful” words from random words: the latter carry maximum information in Shannon theory and essentially zero information within the cell. Despite the efforts of augmenting Shannon theory with “meaning”, apparently there is no quantitative theory of this kind applicable to the cell that would justify the idea of “reads”, that are sequences which can be read and biologically interpreted by the transcriptional/translational machinery of the cell.

**Remark.** The above is only a sketch of a definition meant to illustrate the idea of how a gene can be defined via the collective structure in  $\mathcal{L}$ , where one should distinguish *minimal* significant segments. It is meant to capture, besides protein encoding genes and regulatory regions, such entities as exons, regions producing functional RNA's and possibly some other segments whose function is still unknown.

*Functional and combinatorial role of labels on strings:* We want to reconstruct the labels on the strings that are significant segments of the genome in the above sense, along the same lines as we distinguished significant segments. Of course, one cannot reconstruct the biological *function* of a particular distinguished segment, e.g. being a protein coding gene, but one can trace such a property in the combinatorial geometry of the genoplex. Eventually, we want to find some similarity function on (pairs of) strings depending on the distribution of these labels in the full combinatorial architecture of the totality of genoplexi corresponding to living cells. For example, one can regard two genes translated into proteins with metabolic functions as similar if significant fragments of the metabolic pathways involving these proteins are isomorphic.

*Links and their labels:* One can try to identify significant links in the same way as we defined significant strings by substituting “letters” by “strings” and “strings” by “links” in the discussion above. In other words, significant links are those whose substitution/removal from a genome, has a distinctly significant effect on the fitness potentials.

**Remark.** The above combinatorial description ties up labels to particular strings, links and genomes. However, it is desirable to define biological significant labels independently of particular genomes as it was indicated for the above examples, such as “protein” label, “protein complex” label, etc. This is necessary both for a mathematical satisfactory formalism and for conceiving databases.

*Viability, consistency and evolutionary feasibility:* A genoplex is called *viable* if it can be implemented by a set of DNA strands<sup>9</sup> of some genome(s) such that the physiological functions (properties) of these strands agree with what is written in the labels.

A genoplex is called *consistent* if its labelling is consistent with the present day biophysical and biochemical data on the functioning of the cell.

A genoplex is called *evolutionary feasible* if there is a feasible<sup>10</sup> chain of mutations leading to building such a genoplex within the allotted evolution time.

<sup>9</sup> A strand refers to a segment of DNA made out of the bases indicated in the corresponding string.

<sup>10</sup> The available biological data suggest paths between present-day organisms but specific constructions of these paths, where each edge represents a single mutation, may need genoplexi which do not describe actual organisms. Some of them may correspond to extinct organisms in the context of natural evolution and some may be implemented by artificial evolution. Feasibility for natural evolution allows from hundred thousands to billions of generations. In artificial evolution one is limited to tenths, hundreds and rarely to thousands of generations.

Despite the fact that these three notions are not precisely defined, they will guide our requirements on genoplaxi: we want them to be viable, consistent and evolutionary feasible.

## 5. Syntactic geometry and entropy in sequence spaces

Let  $\mathcal{S}_n$  denote the space of sequences of length  $n$  in a finite alphabet of  $r$  letters, not necessarily made of  $A, C, T, G$ . Examples we have in mind are spaces of  $A, C, T, G$ -sequences making a gene of length about 1000, and spaces of sequences in 20 letters of length about 300 representing amino-acid contents of proteins. We are interested in subsets  $L \subset \mathcal{S}_n$  corresponding to labels  $\mathbf{L}$  associated to sequences. Such a set  $L$ , consisting of the strings that fulfill the function described by  $\mathbf{L}$ , is called the *syntactic image* of  $\mathbf{L}$ . Typical examples are sets of amino-acid sequences encoding globular proteins that properly fold under specified (sometimes unspecified) conditions, and/or having functional domains with specified binding or enzymatic properties.

The number of subsets  $L \subset \mathcal{S}_n$  is double exponential in  $n$ , and therefore, for large  $n$  ( $n > 100$  is safe), one can assume it faithfully reflects the informational content of any conceivable biological label. On the other hand, due to the size of the set, it is unfeasible to describe it formally and explicitly without appeal to the biological meaning of the label. What is more practical is to identify essential properties of such a set  $L$  and relate them to the biological features of  $\mathbf{L}$ .

The basic characteristics of a set  $L$  is its cardinality  $|L|$  as compared to the cardinality of the space  $\mathcal{S}_n$ . This can be conveniently measured by the *total syntactic entropy* and by the *average syntactic entropy per site*:<sup>11</sup>

$$ent_{\text{synt}}(L, n) = \log_r |L|$$

and

$$ent_{\text{synt}}^{\%}(L) = ent_{\text{synt}}^{\%}(L, n) = \frac{\log_r |L|}{n}.$$

To have a feeling on which kind of subsets  $L$  may appear in this context, let  $r$  be a power of a prime, e.g.  $r = 4$ , and think of  $\mathcal{S}_n$  as the  $n$ -dimensional vector space of the field of  $r$  elements. We want to think of a label  $\mathbf{L}$  to be a system of constraints imposed on sequences representable by a system of linear equations, where  $L$  represents the set of solutions of these equations. If we have  $m$  independent equations, then  $ent_{\text{synt}}(L, n) = n - m$ . In other words, this entropy plays the role of dimension. This suggests the definition of *syntactic coentropy* for an arbitrary label  $\mathbf{L}$

$$coent_{\text{synt}}(\mathbf{L}) = n - ent_{\text{synt}}(L, n).$$

<sup>11</sup> Subsets in the binary sequence spaces are extensively studied in coding theory. We deviate from the standard notation and terminology of coding theory. For example, what we call average syntactic entropy per site appears there under the name of “transmission rate” and/or “dimension”.

Similarly, normalizing by  $n$ , one introduces

$$coent_{\text{synt}}^{\%}(\mathbf{L}) = 1 - ent_{\text{synt}}^{\%}(L).$$

The total coentropy is suitable when the label refers to a number of amino-acids which is small compared to the number of residues that make the active domain. The average coentropy per site is more convenient when the constraint imposed by  $\mathbf{L}$  is global (non-easily localizable), e.g. saying that the whole chain makes an  $\alpha$ -helix or just stating that the protein properly folds, since these properties involve all or most of the residues making the chain.

The basic heuristic principle manipulating these codimensions reads

if  $\mathbf{L} = \mathbf{L}_1 \wedge \mathbf{L}_2$  then

$$coent_{\text{synt}}(\mathbf{L}) = coent_{\text{synt}}(\mathbf{L}_1) + coent_{\text{synt}}(\mathbf{L}_2)$$

provided there is no apparent mutual dependence between the functions encoded by  $\mathbf{L}_1$  and  $\mathbf{L}_2$ , and where the notation  $\mathbf{L}_1 \wedge \mathbf{L}_2$  means that the label  $\mathbf{L}$  consists of both  $\mathbf{L}_1$  and  $\mathbf{L}_2$ .

A typical instance of two labels attached together is where  $\mathbf{L}_1$  is the proper folding label and  $\mathbf{L}_2$  stands for binding specificity. One does not expect these labels to be truly independent since an unfolded protein cannot specifically bind. This leads to the introduction of a more realistic coentropy

$$coent_{\text{synt}}(\mathbf{L}_2|\mathbf{L}_1) = coent_{\text{synt}}(\mathbf{L}_1 \wedge \mathbf{L}_2) - coent_{\text{synt}}(\mathbf{L}_1).$$

Next, suppose that  $\mathbf{L}_3$  is the label for a binding or enzymatic activity in some region of protein, far away from the active side of  $\mathbf{L}_2$ . For example,  $\mathbf{L}_2$  and  $\mathbf{L}_3$  refer to two different folding domains or to two different (idealized) *zinc fingers*, that are proteins binding to DNA by several separated small binding domains. Then,

$$coent_{\text{synt}}(\mathbf{L}_2 \wedge \mathbf{L}_3|\mathbf{L}_1) = coent_{\text{synt}}(\mathbf{L}_2|\mathbf{L}_1) + coent_{\text{synt}}(\mathbf{L}_3|\mathbf{L}_1).$$

These rules for evaluating entropies of composed labels can be justified not only by the intersection rules of linear (and more generally, algebraic) subvarieties but also by the corresponding properties of independent random subsets. On the other hand, the geometric properties of syntactic images of biological significant labels do not appear random with respect to the natural geometry in  $\mathcal{S}$ , e.g. with respect to the Hamming metric.

To grasp the picture, let us evaluate

*Spread of random subsets in  $\mathcal{S}_n$* : Look at the Hamming ball  $Ball(\sigma, i)$  in the space  $\mathcal{S}_n$  of binary sequences of length  $n$ . The cardinality (thought of as volume)  $b(n, i)$  of this ball is  $\sum_{j \leq i} \binom{n}{j}$ . If  $i \ll \sqrt{n}$  then, one can think of  $b(n, i)$  as a polynomial in  $n$  of degree  $i + 1$ :

$$b(n, i) \approx C_i n^{i+1} \quad \text{for } C_i = 1/i!.$$

For large  $i \approx n$ , the function  $b(n, i)$  becomes exponential in  $n$ . In fact, by the Stirling formula  $m! \approx e^{-m} m^m$  one has, for  $i = \alpha \cdot n$ ,

$$\binom{n}{i} \approx (\alpha^\alpha (1-\alpha)^{1-\alpha})^{-n}.$$

Fix a point  $\sigma \in \mathcal{S}_n$ , and take  $2^{\delta \cdot n}$  random points in  $\mathcal{S}_n$ . The probability that none of these points is contained in the  $Ball(\sigma, i)$  equals

$$\left(1 - \frac{b(n, i)}{2^n}\right)^{2^{\delta \cdot n}} = \left(\left(1 - \frac{b(n, i)}{2^n}\right)^{2^n / (b(n, i))}\right)^{2^{(\delta-1)n} b(n, i)}.$$

Since

$$\left(1 - \frac{b(n, i)}{2^n}\right)^{2^n / (b(n, i))} \approx e^{-1},$$

the latter expression is

$$\approx e^{-2^{(\delta-1)n} b(n, i)}.$$

This is close to one if and only if  $b(n, i) \ll 2^{(1-\delta)n}$ . It follows that a typical point in a random subset  $L$  with  $ent_{\text{synt}}^{\delta}(L) = \delta < 1$ , contains no other points of  $L$  within distance  $i$ , unless  $i$  is of the order of  $n$ .

If a biologically significant label  $\mathbf{L}$  as represented by such a set  $L$ , then *every* mutation at  $i$  locations, with  $i \ll n$ , would destroy the implied biological function of  $\mathbf{L}$ . This drastically contradicts to observed rates of mutation for most genes and proteins, and tells us that the syntactic images of biological functions are very far from being random.

As another extreme, look at the least random subsets in the set  $\mathcal{S}_n$  of binary sequences, that are

*Coordinate planes:* A coordinate plane  $L$  of dimension  $m = \delta n$  is defined by specifying the values 0, 1 at given  $n - m$  locations in the sequence, and leaving free the remaining  $m$  locations. If  $\sigma \in L$ , then

$$|Ball(\sigma, i) \cap L| = \sum_{i \leq m} \binom{m}{i} \approx C_i m^{i+1} = C_i \delta^{i+1} n^{i+1},$$

where  $C_i = 1/i!$  and where we assume  $i \ll \sqrt{n}$ . Thus, the local size of  $L$  near  $\sigma$  is much larger than that for random sets  $L$ . This picture is closer to the biologically significant situations and motivates the following definitions.

*Local entropies and coentropies:* The  $\delta$  in the above formula, adapted to a realistic (non-Hamming) metric(s) in  $\mathcal{S}$ , can be experimentally computed since one can analyze the biological functionality of (polynomially many) strings  $\sigma'$  obtained by a few point mutations of  $\sigma$ . Therefore, this  $\delta$  can be used for computing (experimentally unreachable) syntactic coentropies of a label by the study of the data coming from natural and artificial evolution, which deliver the  $\sigma'$ 's.

The  $\delta$ , encoding the cardinalities of the intersections of  $L$  with small balls around points in  $L$ , can be regarded as a *local coentropy* or *functional entropic rigidity* of  $\mathbf{L}$  and used for the computation of the global entropy. (Similar to how the dimension of a manifold can be computed by looking at the tangent space at a generic point. In fact, every genome comes along with a family of “infinitesimal neighborhoods” represented by genomes of the evolutionary related species.)

*Super-rigid bioplexi:* In certain situations the above local entropy is essentially zero as it happens, for example, for Histone 3 and Histone 4 proteins, and possibly for some part of viral genomes. In this cases, one should renormalize the entropy in order to obtain a meaningful number. One can hardly compute it in the case of histones by the present day techniques, but the “renormalized viral entropy” can be hopefully evaluated from the data on the evolution of viruses (due to their fast reproduction rate and small genome sizes).

*Evaluation of entropy by stereo-chemistry: molecular coentropy.* Consider a protein with  $n$  residues and suppose that a label refers to an active site composed by a relatively short peptide subchain of  $q$  amino-acids. The space of (stereo-chemically) possible spatial configurations of this chain makes a domain  $A$  in the Euclidean space of dimension proportional to  $q$ , where the determination of  $A$  is given by stereo-chemical data on such chains, e.g. the Ramachandran plot, that is a certain graphical representation of constraints on the angle within the polypeptide chain. Denote by  $A_L \subset A$  the set of configurations compatible with the label  $\mathbf{L}$ . In the first approximation, the syntactic average coentropy per site of  $\mathbf{L}$  is proportional to  $1 - (1/n) \log_{20} \text{vol } A_L / \text{vol } A$ , where  $\text{vol}$  stands for the Euclidean volume in the configuration space. This formula suggests that if the protein has a single active domain, then  $n$  must be of the order  $\log_{20} \text{vol } A_L / \text{vol } A$  and consequently,  $n$  must be  $\approx c \cdot q$ , where  $c$  is a constant that depends on the type of biological function performed by the domain. If there are  $k$  domains, then accordingly  $n \approx c \cdot k \cdot q$ . This may be compared with the fact that the exposed (2-dimensional) part of a globular (of dimension 3) protein is of the order  $n^{2/3}$ .

In the formula above, one should replace the Euclidean volume by a suitable Gibbs type measure,<sup>12</sup> which can be evaluated on the basis of known chemical/physical data. (One may think of mutated sequences preserving  $\mathbf{L}$  as Monte-Carlo samples of  $A_L$ .) The resulting number can be regarded as the *molecular entropic rigidity* of the label.

What is more difficult to evaluate is the geometry of the set  $A_L$  since it depends on particular functional constraints imposed by  $\mathbf{L}$ . In the case of highest specificity,  $A_L$  consists of a single point, or rather of a ball of radius  $\varepsilon$  around a single point, where  $\varepsilon$  is of the order of a 1 Å or less, and it can be evaluated more precisely depending on the physical/chemical nature of  $\mathbf{L}$ .

Following these lines of reasoning, one can evaluate the length of the whole protein needed to realize a given function  $\mathbf{L}$  by considering the map from the space of pro-

<sup>12</sup>The distribution of states of a physical system often obeys the Gibbs law: the probability of a state is proportional to  $e^{-E/kT}$  where  $E$  is the energy of the state,  $T$  is the absolute temperature and  $k$  is the (normalizing) Boltzmann constant.

tein sequences  $\mathcal{S}_n$  to  $A$  and thus relating the syntactic entropy to a suitable  $\varepsilon$ -entropy of  $A$ .<sup>13</sup>

Finally, the syntactic coentropy can be estimated by evolutionary data on the conservation of a given protein, thus suggesting the notion of an *evolutionary entropic rigidity*. The evaluation of this entropy is straightforward for point-mutations, but the availability of segment substitutions depends on the genetic pool within a given organism and/or within the population. Apparently, there is a difference between these mutations for prokaryotes and eukaryotes. The former use horizontal transfer of genomic segments between organisms where most of the segments are functionally significant. On the other hand, eukaryotic genomes contain large amount of non-functional quasirandom sequences, thus allowing substitution of random segments into the genome. Also, one should distinguish mutation/variation possibilities for diploid and haploid genomes in the context of the evolutionary entropic rigidity.

When everything (molecular composition, function and phylogenetic tree) is taken into account, the three kinds of rigidity must coincide. If the three entropies are far away, one should reassess the coentropies of the labels and/or search for extra functional constraints (labels).

This approach can be extended to more complicated *bioplexi*, including *proteoplexi* and *genoplexi*, where a proteoplex is defined in the same way as a genoplex with nucleotide sequences replaced by amino-acid sequences, and where a bioplex refers to any kind of a labelled hypergraph on a set of strings or on a set of syntactically describable biochemical objects. For example, one may think of a protein as a *peptoplex*, and of a metabolic network as a *metaboplex*. Also, a general notion of a bioplex should incorporate the environment where  $\mathcal{S}$  is replaced by “space of strings”  $\times$  “space of environmental parameters”. Eventually, one wishes to bring together evolutionary, molecular and functional data for achieving the entropic (rigidity) consistency of the bioplexi.

## 6. Omissions

This article represents a fragment of a general formalism (the language of bioplexi) that we believe may be useful for describing biological systems. The missing components are:

- our formalism needs to be linked to the static and temporal (stochastic dynamics) cellular chemo-architecture of the cell. A part of this architecture, e.g. a protein traffic along DNA, can be expressed in the language of bioplexi. Other biological aspects of the chemo-architecture, such as the spatial localization of an enzyme or the schedule of a particular process, can be incorporated into the labels.

- we did not describe the interfaces between different bioplexi. Some of these are rather immediate such as the relations between genoplexi and proteoplexi, due to the

---

<sup>13</sup> The ordinary  $\varepsilon$ -entropy measures the minimal number of  $\varepsilon$ -balls needed to cover  $A$ , but this is too crude for the present purpose.

linear correspondence between genes and proteins. Other relationships are less clear, such as the relation between peptoplexi and metaboplexi.

– in our definition of bioplexi we insisted that the vertices of the hypergraphs supporting the label structures are represented by strings of symbols. In certain situations, e.g. for peptoplexi and metaboplexi, the vertices may be represented by non-linear structures, e.g. (the standard Lewis) diagrams of molecules.

– there are several natural operations that one can perform over bioplexi such as taking inclusion (sub-bioplexi), factorization (e.g. by compressing or forgetting the information contained in certain links), amalgamation of two bioplexi along isomorphic sub-bioplexi, substitution of vertices of a genoplex with other genoplexi (e.g. proteins in a proteoplex may be seen as peptoplexi) etc. Such operations, when they are implemented by evolution, can be (hopefully) seen in suitably labelled phylogenetic trees. Other operations may correspond to artificial genetic modifications.

– the organization of natural bioplexi bears traces of the essential features of the structure of cellular processes. Among them, one distinguishes the *specificity/universality principle*. Many mechanisms in the cell are universal such as the production of RNAs and proteins, functioning of tRNAs, phosphorylation of proteins, methylation of DNA, various pathways of degradation (e.g. the ubiquitin system), etc. Some of these processes are implemented by universal molecular machines such as RNA polymerase, ribosomes, proteosomes, chaperones, ubiquitin, etc. that can be viewed as bioplexi serving as vertices of higher level bioplexi.

Among specific phenomena, one finds the preferential binding of proteins to particular targets, especially in the context of immunoplexi, and enzymatic activities.

One needs to explicitly identify the combinatorial properties of bioplexi reflecting this principle.

– when one evaluates the information content (coentropy) of specific strings and links in a bioplex, one should keep track of mutual relations between these links which reduces coentropy. In particular, the symmetry is systematically employed by the cell to “save” information, with the most pronounced example given by the icosahedral symmetry of viruses.

One may distinguish at least five kinds of symmetries:

- *syntactic symmetry* seen, for example, in the repetitions of genomes and in palindromic patterns,
- *spatial symmetry*, e.g. the symmetry of virus coats and of many polymeric proteins,
- *temporal symmetry* seen in cyclic behavior of biochemical processes,
- *combinatorial symmetry of bioplexi* such as the metaboplexi; the essential characteristic of such symmetry is measured by the degree of repetitiveness of small sub-bioplexi inside a bigger one, while the entropy measures the number of combinatorially distinct sub-bioplexi,
- *functional symmetry*, expressing high degree of similarity of certain functions in the cell; this symmetry is close to *universality*.

The basic problem is to relate these symmetries and to use them for the evaluation of the entropic characteristic of genoplexi. In general, there is no simple link between different symmetries, but there are some exceptions such as the palindromic symmetry of binding sites of homodimeric restriction enzymes.

– besides symmetry/entropy, one seeks for other invariants of bioplexi expressing their overall complexity, where one should differentiate between the overall size and the combinatorial depth of the structure. The latter refers to several layers of organization such as metabolic, regulatory, signal transduction, etc. where the number of layers increases in the course of structurally innovative evolution.

– one believes that, among all organisms, viruses are those whose genoplexi have their maximal possible functional coentropy, and therefore these can be used as reference points for the study of the entropies of bioplexi of other organisms.

– for an effective computer implementation, a bioplex should be reduced to a size not exceeding the informational content of a realistic genome. Moreover, the information contained in a bioplex should be organized in several levels according to their biological and logical significance, such that even the first level carrying incomplete information could be conceptually and practically usable.

## 7. Summary and programme

We proposed a framework that organizes the common language for describing biological systems and fragments of these. This allows the incorporation of evolutionary data as well as physical/chemical characteristics of macromolecules in the cell. We indicated a mathematical formalism for finding correlations between these two kinds of data, some coming from *inside* the cell (e.g. biochemistry) and some from *outside* (e.g. the data on evolution of genomic and protein sequences). We plan to analyze existing data, having in mind practical applications such as the evaluation of time needed for the design of proteins with specific properties by means of artificial evolution/selection; also, this may apply to natural evolution/selection processes including the immune system.

## Acknowledgements

We are greatly thankful to François Képès who read the first draft of the manuscript and suggested a variety of corrections and improvements. We also want to thank the referee whose remarks helped us to clarify many points. Vic Norris brought forth to us a biologist view on many issues discussed in this paper and beyond. We wish our formalism could incorporate his ideas.

## Further reading

- [1] S.A. Benner, M.A. Cohen, D.L. Gerloff, Correct structure prediction? *Nature* 359 (1992) 781. (The authors discuss the method for finding spatial conformation of proteins by combining stereo-chemical constraints with evolutionary data. An essential part of our approach is an attempt to extend this idea (in a relaxed form) to a more general context.)

- [2] A.J. Cann, *Principles of Molecular Virology*, Academic Press, New York, 1999. (The book provides an introduction to the molecular architecture, the genetics and the life-style of viruses.)
- [3] A. Carbone, M. Gromov, *Mathematical Slices of Molecular Biology*, *Gaz. Math. Soc. Math. France* (Numéro Spécial) 88 (2001) 11–80. (We give a brief overview of molecular biology aimed at the mathematical audience and indicate possible models for formalizing various fragments of what is called bioplexi in the present article.)
- [4] P. Clote, R. Backhofen, *Computational Molecular Biology*, Wiley, New York, 2000. (This book describes basic algorithms for alignment, for finding secondary structures and for comparing higher order macromolecular foldings. It includes the necessary biological and mathematical background.)
- [5] E.H. Davidson, *Genomic Regulatory Systems—Development and Evolution*, Academic Press, New York, 2001. (This book describes a fundamental body of work on gene regulation circuitry in embryonic development of bilaterally symmetric organisms. It proposes general principles for gene regulation and the functioning of the associated protein machinery.)
- [6] M.B. Elowitz, S. Leibler, A synthetic oscillatory network of transcriptional regulators, *Nature* 403 (2000) 335–338. Letters to Editor. (The paper describes the theory and the implementation of an artificial gene regulation circuit in a bacterial cell.)
- [7] A.J.F. Griffiths, J.H. Miller, D.T. Suzuki, R.C. Lewontin, W.M. Gelbart, *An Introduction to Genetic Analysis*, Freeman, New York, 2000. (This text book covers the basics of classical and molecular genetics and highlights the essentials of the genetic engineering techniques.)
- [8] O. Lichtarge, H.R. Bourne, F.E. Cohen, The evolutionary trace method defines the binding surfaces common to a protein family, *J. Mol. Biol.* 257 (1996) 342–358. (This approach allows to localize active sites of proteins by looking at the evolutionary neighborhood of polypeptide sequences.)
- [9] J. Maynard-Smith, E. Szathmary, *The Major Transitions in Evolution*, Oxford University Press, Oxford, 1997. (The authors give a structural overview of molecular and phenotypic evolution. One can interpret their analysis as an evaluation of the combinatorial depth of the corresponding bioplexi.)
- [10] L. Patthy, *Protein Evolution*, Blackwell Science, Oxford, 1999. (The book provides a survey on natural evolution in eukariotic and prokaryotic organisms.)
- [11] M. Ptashne, *A Genetic Switch*, 2nd ed., Cell Press and Blackwell Science, Oxford, 1992. (The book gives a non-technical account on the fundamental research on the gene regulation of the  $\lambda$ -phage. It presents the author's ideas on the global gene regulation mechanism in cells in prokaryotic and eukaryotic cells.)
- [12] A. Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, *Mol. Biol. Evol.* 18 (2001) 1283–1292. (The paper provides an analysis of what we would call coentropies of the labels associated to the edges of the protein–protein interaction network, with a special attention payed to the evolution following gene duplication events.)