Information Complexity and Biology

Franco Bagnoli[1], Franco A. Bignone[2], Fabio Cecconi[3], and Antonio Politi[4]

[1] Dipartimento di Energetica "S. Stecco", Università Firenze, Via S. Marta, 3 Firenze, Italy, 50139, franco.bagnoli@unifi.it
[2] Istituto Nazionale per la Ricerca sul Cancro, IST Lr.go Rosanna Benzi 10, Genova, Italy 16132, abignone@unige.it
[3] Università degli Studi di Roma "La Sapienza", INFM Center for Statistical Mechanics and Complexity, P.le Aldo Moro 2, Rome, Italy, 00185, Fabio.Cecconi@roma1.infn.it
[4] Istituto Nazionale di Ottica Applicata, INOA, Lr.go Enrico Fermi 6, Firenze, Italy, 50125, politi@ino.it

Abstract. Kolmogorov contributed directly to Biology in essentially three problems: the analysis of population dynamics (Lotka-Volterra equations), the reaction-diffusion formulation of gene spreading (FKPP equation), and some discussions about Mendel's laws. However, the widely recognized importance of his contribution arises from his work on algorithmic complexity. In fact, the limited direct intervention in Biology reflects the generally slow growth of interest of mathematicians towards biological issues. From the early work of Vito Volterra on species competition, to the slow growth of dynamical systems theory, contributions to the study of matter and the physiology of the nervous system, the first 50–60 years have witnessed important contributions, but as scattered pieces apparently uncorrelated, and in branches often far away from Biology. Up to the 40' it is hard to see the initial loose build up of a convergence, for those theories that will become mainstream research by the end of the century, and connected by the study of biological systems per-se.

The initial intuitions of L. Pauling and E. Schrödinger on life and matter date from this period, and will gave the first initial full fledged results only ten years later, with the discovery of the structure of DNA by J. Watson and F. Crick, and the initial applications of molecular structures to the study of human diseases few years earlier by Pauling. Thus, as a result of scientific developments in Biology that took place after the 50', the work of Kolmogorov on Information Theory is much more fundamental than his direct contributions would suggest. For scientist working in Molecular Biology and Genetics, Information Theory has increasingly become, during the last fifty years, one of the mayor tools in dissecting and understanding basic Biological problems.

After an introductory presentation on algorithmic complexity and information theory, in relation to biological evolution and control, we discuss those aspects relevant for a rational approach to problems arising on different scales. The processes of transcription and replication of DNA which are at the basis of life, can be recasted into an Information theory problem. Proteins and enzymes with their biological functionality contribute to the cellular life and activity. The cell offers an extraordinary example of a highly complex system that is able to regulate its own activity through metabolic network. Then we present an example on the formation of complex structures through cellular division and differentiation in a model organism (C. elegans). Finally we discuss the essential principles that are thought to rule evolution through natural selection (theory of fitness landscapes).

F. Bagnoli, F.A. Bignone, F. Cecconi, and A. Politi, Information Complexity and Biology, Lect. Notes Phys. **636**, 123–146 (2003)
http://www.springerlink.com/

If one were to judge Kolmogorov's contribution to Biology only on the basis of his papers explicitly devoted to the topic, one might conclude that it is of moderate interest, at least in comparison with his direct intervention in turbulence or dynamical systems. However, one should not forget that, in the past, the limited availability of quantitative data in Biology made this subject quite unattractive for mathematicians. It is therefore remarkable that Kolmogorov nevertheless contributed to three different problems: the analysis of population dynamics (Lotka-Volterra equations), the reaction-diffusion formulation of gene spreading (Fisher Kolmogorov Petrovsky Piskunov – FKPP– equation), and some discussions about Mendel's laws. It is however widely recognized that the relevance of Kolmogorov's contribution is connected to his work on algorithmic information. In fact, after the initial intuitions of L. Pauling and E. Schrödinger on life and matter in the '40s and, especially after the discovery, ten years later, of the structure of DNA by J. Watson and F. Crick, it has become increasingly clear that information plays a major role in many biological processes. The penetration of these concepts in Biology has led to the formulation of the central dogma of genetics: the discovery of a one-directional flow of information from DNA and genes to proteins, and from there to morphogenesis, cellular organization and finally to individuals and communities. In this perspective, life is now viewed as the execution of a computer program codified in the DNA sequence. However, in spite of the elegance of this doctrine, one cannot forget the various difficulties that make the final goal of decoding the program much more difficult than one could expect. First of all, in order to understand what the life-code does without running it, it is necessary to know the logic of the underlying "hardware", or "wetware" as it is sometimes called. Wetware is certainly structured in quite a different way from the hardware of ordinary digital computers. Indeed, living system are highly parallel devices, where the parallelism does not only enhance the "computer" performance, but is also there to guarantee the required redundancy for a robust functioning both in the presence of a noisy environment or of significant damages even. As a result, in living systems, information-theoretic aspects are profoundly interlaced with the physico-chemical mechanisms responsible for their functioning and it could not be otherwise, considering that they are the result of a self-assembling process that has evolved over billions of years.

The roadmap of this contribution is as follows. The first section is devoted to an historical account of the connections between biology and the other major scientific disciplines. This allows us to place Kolmogorov's direct contributions (exposed in the next section) in the right context. Next, we give a brief presentation of information and algorithmic-information theories in relation to biological systems. Finally, we discuss the problem of protein folding as an example of how information, physics and dynamics concur to biological processes at an almost microscopic level of description.

1 Historical Notes

For a long time, biology has mainly dealt with definitions, classification and description of objects connected with the evolution of life forms such as bacteria, plants, animals, biochemical substances, proteins, sugars or genes. The great variety of life forms, present and past, has required a large amount of ground work before general theories could be formulated. The Evolution theory of Charles Darwin is a good case in point. Moreover, the dissection of environments, organisms and cells has been practiced by many in the hope that a precise identification of the objects of study is a necessary and perhaps sufficient condition to understand their role: a sort of reductionism, like in physics, with the difference that no simple general laws have ever been identified.

In this continuous search for an appropriate methodology, biologists have been led to emphasize different aspects (ranging from mechanical, to chemical, thermodynamical, and molecular) depending on the current development of science. As an example, in Fig. 1 we report an apparatus invented by an

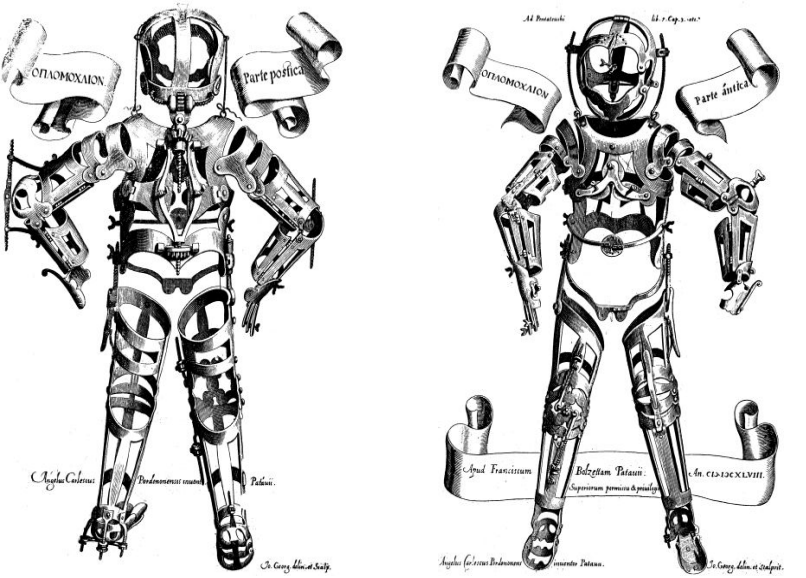

Fig. 1. Girolamus Fabricius ab Acquapendente (1533–1619), machina, Museo Villasneri from [1]. This example of *machina*, a system to bring body parts into correct place–proportions, is due to Fabricius ab Acquapendente, anatomist at the University of Padova. Acquapendente was Professor of Anatomy in Padova during the same period in which Galileo Galilei, who was trained in his youth as a physician, was teaching there (1592–1610). Acquapendente and Galileo were tutors of William Harvey (1578–1657). Harvey, by application of the experimental method, discovered blood circulation, while working at St. Bartholomew's hospital in London [2]

anatomist, Girolamus Fabricius ab Acquapendente at the turn between the 16^{th} and 17^{th} centuries; the machine was meant to be used in order to bring parts of the human body into correct proportions.

The ingenious, naive, primitive, and somehow sadistic mean of fixing body parts through the use of a reference frame is remarkable. The general idea *to fix it* is already there with the means of the period, and has evolved with human knowledge. The present level of technology allows the discussion of atomic-molecular adjustments, but, in many cases, the conceptual aim has not changed much since 1600, aside from a deeper understanding of the inner details involved. The reported picture, nevertheless, warns that the desire to find solutions to biological problems, drawing principles from other sciences, has always been present from the very beginning. Sometimes it has been stretched to the limit.

The limited success of this stratgegy has led in the last century to a strong debate about two alternative approaches, the *holistic* and the *reductionistic* view. The former one assumes that biological systems, in order to be understandable, must be considered and described in their wholeness; the latter one understands that full operational knowledge can be reached only after characterizing all of the single components. While the supporters of the holistic view have, with a few exceptions, fostered more qualitative approaches, the reductionistic attitude has been much more oriented towards quantitative work (for a detailed historical account of this long story, the interested readers are invited to consult the detailed book of B.O. Kuppers, [3], or the historical account of E. Mayr, [4]).

However, in recent years the scenario has started to change and there exists now a chance that the above two points of view can be reconciled within a suitable information-theoretic approach. Indeed, *information processing* represents a truly unifying concept that allows the investigation of seemingly different issues such as the functioning of the brain, the cell cycle, the immune system, or the "simple" task of recognizing food, moving towards it, and so on. Furthermore, one should not forget the problem that has represented a puzzle for centuries: the transmission from one generation to the next of the "plan" for constructing new individuals.

The first results were obtained by G. Mendel (1865) and concerned statistical correlations between phenotypic characters of species. However, the idea that characters are due to elementary units spread about 35 years later, through the work of H. de Vries, C. Correns, and E. Tschermak. Later T. Boveri, W. Sutton, and especially T.H. Morgan and collaborators established the chromosomal theory, the link between characters, genes and the physical existence of chromosomes. Still, all research in genetics up to the beginning of the '40s was done as an inductive reconstruction, through the study of crossing and mutants, with no knowledge of the mechanisms and molecules carrying this information.

All in all, spreading of information-theoretic concepts in Biology occurred by following several convoluted paths. This is true also for the brain: the

system that, more than any other, can be seen as an information-processing unit. It is indeed quite illuminating to cite W.S. McCulloch about the way he came, together with W. Pitts, to the development of his famous simplified model of a brain, in which neurons were treated as boolean interconnected switches [5],

> I came, from a major interest in philosophy and mathematics, into psychology with the problem of how a thing like mathematics could ever arise – what sort of thing it was. For that reason, I gradually shifted into psychology and thence, for the reason that I again and again failed to find significant variables, I was forced into neurophysiology. The attempt to construct a theory in a field like this, so that it can be put to any verification, is tough. Humorously enough, I started entirely at the wrong angle, about 1919, trying to construct a logic for transitive verbs. That turned out to be as mean a problem as modal logic, and it was not until I saw Turing's paper that I began to get going the right way around, and with Pitt's help formulated the required logical calculus.

One of the major obstacles in the development of a theoretical Biology is its nonstationary character: strictly speaking, there are no stationary states – one can at most imagine that, on some scales, quasi-equilibrium is maintained. It is, indeed, since the beginning of the last century that Evolution has been recognized as playing a crucial role and the first models on the time variation of species have been introduced. Initially, the spread of the ideas of C. Darwin strongly depended on the country; in some cases they were partially accepted, allowing evolution but not natural selection – such as, e.g., in France. In others, these ideas were accepted much faster, as in the case of Russia [4]. In the beginning of the century a dichotomy between naturalists and geneticists took place on the way to proceed in order to understand evolution. The former looked more at *final causes*, while the latter, more oriented towards physical and mathematical methods, pursued a strict experimental approach. The major achievement of Genetics in this period was the rejection of the theory of acquired characters – i.e. the pangenesis hypothesis of C. Darwin, or the theories of those biologists who followed J.B. Lamarck [4] –. Without any knowledge of the physical base for the transmission of characters, the demonstration was done by means of statistical methods, and by showing the combinatorial character of traits due to more than one gene (the final experimental demonstration came with work done by Salvatore Luria and Max Delbrück in the '40s).

Attributing a key role to information processing amounts to assuming that the mechanisms through which, e.g., a face or an antigene is recognized, can be understood without the need to characterize in full detail the underlying physical processes and chemical reactions. This is indeed a fruitful hypothesis, formulated already by von Neumann in the book on "The computer and

the brain", that has given rise to the formulation of several more-or-less abstract models introduced in the last 20 years, in the hope of identifying possibly universal mechanisms. One of the most successful models is the Hopfield model [6] that exploited a possible analogy between an associative memory and a spin glass. It shows how information can be robustly stored and retrieved in a context where many connections can be accidentally destroyed, as it is the case of our brains.

Although this is a route that will be useful to pursue in the future as well, one cannot neglect biochemical processes, at least to understand how biological systems can self-assemble. In fact, another discipline that is having an increasing impact on Biology is the theory of dynamical systems. In the last century it has been progressively recognized that most, if not all, processes that are responsible for the functioning of a living system involve nonlinear mechanisms which, in turn, are responsible for the onset of nontrival time dynamics and the onset of spatial patterns. Initial rudimentary attempts to figure out a physico-chemical explanation for the origin of life can already be found in [7], although this remained an isolated attempt, still very qualitative and rooted into J.B. Lamarck ideas. The modelling of oscillations, thanks to the work of A.J. Lotka, where one of the well studied models was introduced, can also be attributed to this path.

Later contributions came thanks to the advances of B.P. Belousov, with his discovery of the chemical reaction that bears his name, the Belousov-Zhabotinsky reaction. Belousov discovered this reaction while attempting to model the Krebs cycle. The Krebs cycle, i.e. the tricarboxylic acid cycle, is the name given to the set of reactions that transforms sugars or lipids into energy. Degradation produces acetyl-CoA, a molecule with two atoms of carbon, which are transformed through the cycle in two molecules of CO_2 while producing energy in the process. He showed that the oxidation of citric acid in acidic bromate, in the presence of Cerium catalysis – [Ce(IV)]/[Ce(III)] –, produces oscillations in the reaction visible through changes in color of the solution. The discovery was made in 1951 but the paper was rejected because the nonequilibrium nature of the thermodynamic process was not understood. He finally published his result in 1958 in the proceedings of a conference. Along this same path, a theoretical paper on the spontaneous formation of patterns in chemical systems was published by Alan Turing in 1952 [8]. However, while Belousov's contribution had an experimental basis, it was not until the beginning of the '90s that Turing's hypothesis was demonstrated experimentally.

The relevance of dynamical system theory in Biology has definitely emerged in the beginning of the '60s in connection with the problem of gene regulation. For instance, in a series of papers, by Jacques Monod, Francois Jacob, and André Lwoff give some hints about the *logic* of living systems and show that regulation of the β-Galactosidase system in bacteria can be seen as a switch. The classical scheme of the Lactose Operon works as a sensor of the presence-absence of Lactose. As a first approximation, a protein

produced in the bacterial cell, the lactose repressor, binds to the operator, a 22 base pairs (bp) long stretch of DNA in front of the genes. This blocks RNA Polymerase that should start transcription of DNA in order to make the mRNAs of the three genes which are part of the Operon (β-Galactosidase, transacetylase and lactose permease). If the lactose is present, it binds to the repressor, unlocking the operator and transcription begins. If the lactose is absent, transcription is blocked. Aside from further complexities, this system, in its baseline, can be considered similar to a boolean switch. These ideas have been pointed out by Jacques Monod and Francois Jacob both in technical papers and in popular articles. Particularly Monod stressed the fact that the logic of living forms follows Boolean algebra, with a series of more or less complex logic circuits at work.

The initial formalization of genes as switches, in a way similar to the modelling of McCulloch and Pitts, is due to M. Sugita in 1961, soon followed by S. Kauffman [9,10]. A similar approach to the study of the dynamics of gene expression was pursued by B. Goodwin [11].

However, recognition that high-level cellular functions are regulated by a plethora of proteins interacting in cells had to wait until the end of the '80s, beginning of the '90s. Since then, it has become increasingly clear that the lower dimensional levels in Biological systems, those of molecules, organelles and cells, are as difficult as the higher dimensional one to solve. Several apparently *simple* functions have revealed a degree of sophistication previously unforeseen. As a result, the currently emerging picture of multicellular systems is much more similar to a *highly regulated society*, than to the simple gene-protein scheme accepted for many years, [12–15].

2 Kolmogorov's Direct Contributions

Before discussing the contributions of Kolmogorov to Biology, it is useful to recall the situation in the Soviet Union. While acceptance of natural selection in the USA and Europe had to overcome political and philosophical barriers, this was not the case in Russia. An important figure in the development of Evolution theories was Sergej S. Chetverikov (1880-1959). In 1906 he published an important study on fluctuations in populations. He was able to demonstrate that not all mutations have a negative impact on fitness: some are almost neutral and, as shown later by Dobzansky, some can even increase the fitness. Moreover, because of heterozygosity – the presence of two copies of each gene – most mutants remain silent within the population, as shown also by R.A. Fisher, and only homozigous individuals will be exposed to selection. Chetverikov demonstrated these facts through back crossing experiments with wild type *Drosophila melanogaster*. His most important result was that the previous idea of the structure of organisms made of independent genes had to be abandoned. No gene has a constant fitness because his expression will depend on the global genetic background. However, his work

was not well-known outside Russia and, after 1926, he had to leave his post for political reasons [4,16].

Independently of the work done within the group of Chetverikov, around 1935, Kolmogorov became interested in some problems of mathematical genetics, probably stimulated by the ongoing debate about Lamarckism occurring in Europe and especially in the Soviet Union.

The first relevant contribution on the subject deals with the propagation of "advantageous genes" (see chapter 9 in this book). The variable of interest is the concentration $0 \leq c(x,t) \leq 1$ of individuals expressing a given gene, at position x and time t. In the absence of spatial dependence, the concentration is assumed to follow a purely logistic growth, $\dot{c} \equiv F(c) = Kc(1-c)$: this dynamics is characterized by two stationary solutions, $c = 0$ and $c = 1$. If $K > 0$, the former one is linearly unstable; any small fraction of individuals carrying the "advantageous" gene tends to grow, but the limited amount of resources put a limit on the growth which converges to the stable solution $c = 1$. In the presence of spatial directions, it is natural to include the possibility of a random movement of the single individuals. As a result, the whole process is described by Fisher's equation [17], proposed in 1937, ten years after the work of Chetverikov,

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} + Kc(1-c). \tag{1}$$

$c = 0$ and $c = 1$ are still meaningful stationary solutions, but now the dynamics of the spatially extended system is determined not only by the evolution of small perturbations, but also by the propagation of one phase into the other. This is a general observation whose relevance goes beyond the original context, since it applies to all pattern-forming dynamical systems. This specific model is important, since it is one of the very few nonlinear reaction-diffusion equations that can be treated analytically. The relevant solutions are front-like ones connecting the two different fixed points (e.g, $c(x,t) \to 1$ for $x \to -\infty$ and $c(x,t) \to 0$ for $x \to \infty$). The relative stability of the two phases is thereby quantified by the front velocity that can be estimated by assuming that the front travels without changing shape, i.e. $c(x,t) \equiv f(x-vt) \equiv f(z)$. By assuming that $f(z)$ decays exponentially to 0, $f(z) \simeq \exp(-\gamma z)$ for large z, one can easily investigate the front propagation by replacing this ansatz into the linearized (1). As a result, one finds that

$$v(\gamma) = \begin{cases} K/\gamma + \gamma D \text{ if } \gamma > \gamma^* \\ \\ 2\sqrt{KD} \qquad \text{if } \gamma \leq \gamma^* \end{cases}$$

where $\gamma^* = \sqrt{K/D}$. If the parameter γ defining the initial profile is smaller than γ^*, then the front propagates with the minimal velocity $v_{min} = v(\gamma^*) = 2\sqrt{KD}$. This is the well-known velocity selection mechanism (see also chapter 9 of this book).

In the same year as Fisher's paper, Kolmogorov, Petrovskii and Piskunov [18] extended the solution of the problem to a fairly general class of local growth-functions $F(c)$, rigorously proving the following expression for the propagation velocity

$$v_{min} = 2\sqrt{F'(0)D} \tag{2}$$

where the prime denotes the derivative w.r.t. the argument (Fisher's result is recovered by noticing that in the logistic case $F'(0) = K$).

There are two other studies by Kolmogorov in genetics. The first one [19] concerns the problem of statistical fluctuations of Mendel's laws. The interest in this subject is mainly historical, since it aimed at refuting a claim by T.D. Lysenko that Mendel's "3:1 ratio"-law is only a statistical regularity, rather than a true biological law. More important is the second contribution which extends the notion of Hardy-Weinberg (HW) equilibrium in population genetics. HW equilibrium refers to the simplest setup, where allele statistics can be studied. It assumes: i) random crossing between individuals; ii) absence of mutations; iii) neutrality, i.e. absence of mechanisms favouring a given allele; iv) closed population in order to avoid exchanges of alleles with the environment; v) an infinite population. Mathematically, the problem can be formulated as follows: Given the probability p ($q = 1 - p$) to observe the allele A (a), the free crossing of genes A and a produces AA, Aa, and aa with probabilities p^2, $2pq$, and q^2, respectively and the frequency of individuals follows a Bernoulli distribution.

The route towards more realistic models requires progressively relaxing the above restrictions. Kolmogorov first investigated the effect of a diffusive coupling in a system of otherwise closed populations and then studied the consequences of selection simulated by a mechanism suppressing the occurrence of the recessive allele a. Here, we briefly summarize the first generalization. Kolmogorov considered a large ensemble of N individuals divided into s populations, each containing the same number n of individuals, ($N = sn$). Like in the HW scheme, random mating (free crossing) is assumed within each population. Moreover, a number of k individuals are allowed to "migrate" towards different populations and thus to contribute to the next generation. As a result of the mutual coupling, a population characterized by a concentration p of the allele A experiences an average drift $F(p)$ towards the equilibrium value p^* (corresponding to the concentration in the total population) with variance σ^2,

$$F(p) = \frac{k}{n}(p^* - p) \qquad \sigma^2(p) = \frac{p(1-p)}{2n}.$$

Altogether, the distribution $\rho(p,t)$ of populations with concentration p satisfies the Fokker-Planck equation,

$$\frac{\partial \rho}{\partial t} = -\frac{\partial F\rho}{\partial p} + \frac{1}{2}\frac{\partial^2 \sigma^2 \rho}{\partial p^2} \quad , \tag{3}$$

whose stationary solution is

$$\rho(p) = \frac{C}{\sigma^2(p)} \exp\left\{2 \int dp \frac{F(p)}{\sigma^2(p)}\right\} = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)},$$

where $\alpha = 4kp^*$, $\beta = 4kq^* = 4k(1-p^*)$ and the Euler beta-function $B(\alpha,\beta)$ accounts for the proper normalization. The frequency of individuals carrying AA, Aa, and aa and, hence deviations from a pure HW-equilibrium can then be estimated by simply averaging p^2, $p(1-p)$, and $(1-p)^2$, respectively over the distribution $\rho(p)$.

Finally, Kolmogorov made some contributions in the modeling of population dynamics, by generalizing the Lotka-Volterra equations. Such a model, on the Volterra side, followed an experimental observation by the Italian biologist Umberto D'Ancona, who discovered a puzzling fact. During the first World War, the Adriatic sea was a dangerous place, so that large-scale fishing effectively stopped. Upon studying the statistics of the fish markets, D'Ancona noticed that the proportion of predators was higher during the war than in the years before and after. V. Volterra, stimulated by his son in law D'Ancona, formulated the problem in terms two coupled differential equations,[1]

$$\frac{\mathrm{d}N_1}{\mathrm{d}t} = (\varepsilon_1 - \gamma_1 N_2)N_1, \qquad \frac{\mathrm{d}N_2}{\mathrm{d}t} = (-\varepsilon_2 + \gamma_2 N_1)N_2,$$

N_1 and N_2 being the abundance of preys and predators, respectively. They exhibit periodic behaviour whose amplitude depends on the initial conditions. This feature crucially depends on the form of the proposed equations, because the dynamics admits a conserved quantity E (analogous to the energy in conservative systems)

$$E = \gamma_2 N_1 + \gamma_1 N_2 - \varepsilon_2 \log(N_1) - \varepsilon_1 \log(N_2),$$

while the periodic orbits are level lines of E. However, E has no direct biological meaning. Kolmogorov argued that the term $\gamma_2 N_1$ is too naive, because it implies that the growth rate of predators can increase indefinitely with prey abundance, while it should saturate at the maximum reproductive rate of predators. Accordingly, he suggested the modified model [22]

$$\frac{\mathrm{d}N_1}{\mathrm{d}t} = K_1(N_1)N_1 - L(N_1)N_2, \qquad \frac{\mathrm{d}N_2}{\mathrm{d}t} = K_2(N_1)N_2,$$

where $K_1(N_1)$, $K_2(N_1)$ and $L(N_1)$ are suitable functions of the prey abundance and predators are naturally "slaved" to preys. With reasonable assumptions on the form of $K_1(N_1)$, $K_2(N_1)$ and $L(N_1)$, Kolmogorov obtained

[1] The same equations were derived also by A.J. Lotka some years before [20,21] as a possible model for oscillating chemical reactions

a complete phase diagram, showing that a two-species predator-prey competition may lead to either extinction of predators, stable coexistence of prey and predator, or, finally, oscillating cycles. He also generalized the differential equation to more than two species[2], introducing most of the phenomenology nowadays known in population dynamics.

Moreover, Kolmogorov pointed to the strong character of the assumptions behind an approach based on differential equations. In particular, populations are composed of individuals and statistical fluctuations may not be negligible, especially for small populations. In practice, there exists a fourth scenario: at the minimum of a large oscillation, fluctuations can extinguish the prey population, thereby causing the extinction of predators too. It is remarkable to notice how Evolution has developed mechanisms to reduce "accidental" extintions. In most species, the birth of individuals takes place during a very short time interval. In some cases, such as for example for herbivores like the gnus – *Connochaetes taurinus* –, living in herds, the birth of puppies is limited to a time span as short as one-two weeks. This mechanism helps in preserving the species since the number of newborns highly exceeds the possibility of killing by predators.

3 Information and Biology

Information is one of those technical words that can also be encountered within natural languages. C.E. Shannon, who was mainly interested in signal transmission, succeeded in formalizing the concept of information by deliberately discarding semantic aspects. He states in the beginning of [24]

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently, the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem.

In fact, before discussing the meaning of single messages, it is necessary to distinguish among them. In this sense, the information becomes the number of independent specifications one needs in order to identify a single message x in an ensemble of N possible choices. Given, for instance, an ensemble of N equiprobable messages x_k, the unambiguous identification of a specific message x requires taking $\log_2 N$ binary decisions. One can indeed split the initial ensemble into two subsets and identify the one containing x. This operation involves the minimal amount of information, a "bit", and it must be recursively repeated until the remaining set contains no more than one element. Accordingly, the amount of information is $I = I(x) = \log_2 N$ bits.

[2] See for instance [23]; the generalized version is sometimes referred as the *Kolmogorov model.*

If messages do not have the same probability, it is natural to define the information of a single message as

$$I_k = -\log_2 p(x_k) \quad , \tag{4}$$

and accordingly introduce the average information

$$H = \sum_k p(x_k) I_k = -\sum_k p(x_k) \log_2 p(x_k) \quad . \tag{5}$$

The quantity H was defined as an entropy by Shannon, since it can also be interpreted as an uncertainty about the actual message.

In many contexts, the object of investigation is an ideally infinite sequence $s_1 s_2 \dots s_n \dots$ of symbols (s_i belonging to an alphabet with 2^b letters) that can be viewed as a collection of strings S_i of length n with probability $p(S_i, n)$. In this case, the information is written as $H(n) = \sum_{i=1}^{2^{bn}} p(S_i, n) \log_2 p(S_i, n)$ and the sum extends to the set of all possible strings. The difference $h_n = H(n+1) - H(n)$ is the information needed to specify the $(n+1)$st symbol given the previous n, while $h = \lim_{n\to\infty} h_n$ is the Kolmogorov-Sinai entropy of the signal. The maximum value, $h = b$ is attained for random sequences of equally probable symbols, while $h = 0$ is a distinctive feature of regular messages.

Another useful indicator is mutual information

$$M(k) = \sum p(s_j, s_{j+k}) \log_2 \frac{p(s_j, s_{j+k})}{p(s_j) p(s_{j+k})}, \tag{6}$$

measuring the statistical dependence between two variables; $M(k) = 0$ if and only if the two variables are mutually independent.

While the concept of information was being formalized, crucial progress was made in Biology that led to the discovery of the DNA double-helix structure by J. Watson and F. Crick [25] in 1953. This was made possible by the development of methods for the analysis of chemical structures based on X-ray scattering, mostly by William and Lawrence Bragg together with the intuitions of L. Pauling for the application of the method to the study of protein and DNA structure [26][3]. One should not however forget also the impact of the book of E. Schrödinger on atoms and life [27], where he argued about the existence of a disordered solid as the medium hiding the secrets of life.

The crucial role of information within genetics has become increasingly clear with the discovery that DNA and proteins are essentially string-like objects, composed by a sequence of different units (bases and amino acids,

[3] The story goes that the interest in protein structure was aroused in Pauling by Warren Weaver, head of the Natural Sciences division of the Rockefeller Foundation, who convinced him to work on the problem, financed by the Rockefeller's funds.

Table 1. Genetic code, translating the codon triplets into amino acids, e.g. UUU and UUC both corresponds to amino acid Phenylalanine (Phe), while Leucine (Leu) is encoded by six possibilities UUA, UUG, CUU, CUC, CUA, CUG. Notice that the symbol "T" is replaced by "U" since the translation codons → aminacids actually involves RNA and not directly DNA. It is evident that most of the redundancy in the code is due to the third base of each codon. Triplets UAA, UGA and UAG are the stop-codons; they do not encode any amino acid but locate the end of the protein

First Position	Second Position				Third Position
	U	C	A	C	
U	Phe	Ser	Tyr	Cys	U
U	Phe	Ser	Tyr	Cys	C
U	Leu	Ser	**Stop**	**Stop**	A
U	Leu	Ser	**Stop**	Trp	G
C	Leu	Pro	His	Arg	U
C	Leu	Pro	His	Arg	C
C	Leu	Pro	Gln	Arg	A
C	Leu (Met)	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
A	Ile	Thr	Asn	Ser	C
A	Ile	Thr	Lys	Arg	A
A	Met (Start)	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
G	Val	Ala	Asp	Gly	C
G	Val	Ala	Glu	Gly	A
G	Val (Met)	Ala	Glu	Gly	G

respectively) linked together by covalent chemical bonds which ensure a chain structure. The properties of DNA, RNA and proteins are briefly recalled below in a specific box; for further details, we recommend that the reader consult any modern textbook on molecular biology [28,29].

The information contained in the DNA is first transferred to RNA and eventually to proteins. In the latter step there is a loss of information because the $4^3 = 64$ possible different triplets of nucleotides – codons – are mapped onto only 20 amino acids (see Table 1). This is therefore an irreversible process and there is no way to go back from the proteins to DNA since different nucleotide sequences can code for the same protein.

After the discovery of DNA's structure and function, a shift of focus took place in genetics and biology: all relevant properties became traced back to the information stored at the molecular level. The DNA sequence is viewed as "the fundamental issue" and the pursuit of the *sequencing projects* for several organisms has been the main direct consequence of this view. The centrality of DNA is so undisputed that *human-like* behavioral characteristics are occasionally attributed to the chemical properties of this molecule.[4] The

[4] The use of the catchy metaphor of the *selfish gene* by R. Dawkins is a good example.

reasons for this are several, the main one being the appealing interpretation of living organisms as complex computing machines. The DNA then represents the code that the program actually runs. However, the relative ease with which DNA molecules can be studied has contributed to this view, since the DNA is relatively constant for a certain organism. Technological advances during the last three decades have made the process sequencing routine, through the use of automatic machines.

Given its naturally symbolic structure, DNA can be directly investigated by means of information-theoretic tools. The zero-*th* order question is whether DNA can be treated as a stationary process, since the very computation of probabilities requires this. Several studies have revealed that a slow drift in the composition may be present and must be carefully accounted for. Once this is made clear, one can proceed by computing the probability $p(S_i, n)$ of each sequence of length n and thus determine the information $H(n)$. If the DNA was a purely random sequence of equally probable bases (four), $H(n)$ would attain its maximum value $H(n) = 2n$. This is almost true for $n \leq 4 \div 5$: for instance $H(1) = 1.95$ in the human chromosome 22 [30], the 0.05 difference from 2 being an indication of the slightly uneven distribution of the four nucleotides. However, upon increasing n, h_n decreases and for $n = 10$ it has already decreased down to 1.7. Going to larger n-values is basically impossible, since one would need such large samples to reliably determine the exponentially small probabilities, that even $10^7 - 10^8$ bp are no longer sufficient. One partial way to go around the problem is by looking for low-order correlations. A straightforward solution consists in studying the standard correlation $C(k) = \langle s_j s_{j+k} \rangle - \langle s_j \rangle^2$. Starting with [31], several studies performed on different living organisms have revealed a slow correlation-decay, $C(k) \simeq k^{-\gamma}$. Such observations have been confirmed by studying also the mutual information $M(k)$ (see (6)) which only requires computing probabilities of pairs of symbols, k bp apart from each other. For instance, in [30], a decay with $\gamma \approx 1/4$ was found up to $k = 10^5$. Many factors seem to be responsible for such a slow decay on different scales, but none of them prevails. For instance, it is known that many almost-equal repeats are interspersed within DNA and some are even as long as 300 bp, but they are responsible for correlations only up to $k \approx 10^2$,

3.1 Algorithmic Information

As soon it was realized that DNA is simply a long message possibly containing the instructions for the development of a living being, algorithmic issues immediately became relevant. In the decade across '60s and '70s, R. Solomonoff, A. Kolmogorov, and G. Chaitin, [32–37], independently set the basis of what is now known as *algorithmic information theory*. Consider the sequences

$$S_a = ATGCATGCATGCATGCATGCATGCATGCATGCATGC$$
$$S_b = AATAGATACAAACATGTCGACTTGACACATTTCCTA,$$

it is clear that S_a is somehow simpler than S_b. Suppose indeed that we have to describe them; while the former string is fully characterized by the statement `8 times ATGC`, the latter cannot be better described than enumerating the individual symbols. However, in order to make more quantitative our considerations about "simplicity", it is necessary to formalize the concept of description of a given string. The Turing machine is a tool to answer this question: it is a general purpose computer which, upon reading a series of instructions and input data (altogether representing the program), produces the required string S. It is therefore natural to consider the program length as a measure of the "complexity" of S. As proven by Solomonoff, Kolmogorov and Chaitin this is an objective definition, provided that the shortest code is first identified. In fact, on the one hand, there exist the so-called universal Turing machines (UTMs) that are able to emulate any other machine. On the other hand, there is no need to refer to a specific UTM, since the unavoidable differences among the lengths of minimal codes corresponding to different UTMs are independent of the sequence length N. More precisely, the Kolmogorov-Chaitin algorithmic complexity $K(S)$, i.e. the minimal code length, is known within a machine-dependent constant and $\kappa(S) = \lim_{N\to\infty} K(S)/N$ is an objective quantity. Unfortunately one consequence of the undecidability theorem, proved by Kurt Gödel, is that there is no general algorithm to determine $\kappa(S)$ which thereby turns out to be an uncomputable quantity.

While information deals with ensembles of strings, algorithmic information aims at measuring properties of single sequences. In spite of this striking difference, there is a close analogy to the extent that we now often speak of "algorithmic information". In fact, one may want to determine the probability $P(S)$ that a given UTM generates a string S when fed with a sequence of independent, equally probable bits. Since Chaitin proved that $K(S) = -\log_2 P$, one can interpret $K(S)$ as the logarithm of the probability that the minimal code is randomly assembled. This observation is particularly suited to discuss the role of chance within biological evolution. Indeed, if the DNA sequence is a randomly selected program, even imagining the Earth as a gigantic parallel processor performing independent tests every cubic millimiter each nanosecond,[5] the probability $P(DNA)$ should be larger than 10^{-50} and, accordingly, $K(DNA) < 200$. In other words, it should be possible to compress the DNA

[5] This is reminescent of an episode of *the hitchiker's guide to the galaxy* by Douglas Noel Adams (whose acronym is DNA)
`http://www.bbc.co.uk/cult/hitchhikers/`: Some time ago a group of hyperintelligent pan dimensional beings decided to finally answer the great question of Life, The Universe and Everything. To this end they built an incredibly powerful computer, Deep Thought. After the great computer programme had run seven and a half million years, the answer was "42". The great computer kindly

sequence down to less than 200 bits (or, equivalently, 100 bp). We cannot exclude that this is the case, but it is hard to believe that all instructions for the development, of e.g. humans, can be compressed within such a short length!

An observation that helps to close the gap is that only part of the genome is transcribed and then translated: according to the most recent results, less than 2% of human DNA is transformed into proteins! A small fraction of the remaining 98% contributes to the regulation of metabolic processes, but the vast majority seems to be only accidentally there. This is so true that onion-DNA contains 3 times more bp than human-DNA! Whatever the algorithmic content of this so-called "junk" DNA, we are clearly left with the crucial problem of discovering the language used to store information in DNA. Several researchers have investigated the DNA structure in the hope of identifying the relevant building blocks. In natural and artificial languages, words represent the minimal blocks; they can be easily identified because words are separated by the special "blank" character. But how to proceed in partitioning an unknown language, if all blanks have been removed? Siggia and collaborators [38] have proposed to construct a dictionary recursively. Given, e.g., the sequence $S_n = s_1 \dots s_n$, it is extended to $S_{n+1} = s_1 \dots s_n s'$, if the probability of S_{n+1} turns out to be larger than the probability of S_n multiplied that of the symbol s'. In the opposite case, the process is stopped and S_n is identified as a new word. The rationale behind this approach is that when a word is completed, a sudden uncertainty arises due to the ignorance about the newly starting one. The above approach has been successfully implemented, allowing the recognition of several regulatory motifs.

Extracting information from the sole knowledge of the DNA sequence seems however to be an exceedingly hard problem, since the products of DNA translation interact with each other and with the DNA itself. In order to gain some insight about living matter, it is therefore useful, if not necessary, to look directly at the structure of the "final product" in the hope of identifying the relevant ingredients. In this philosophy, many researchers have pointed out that living matter is characterized by non-trivial relationships among its constituents. Chaitin [39], in particular, suggested that the algorithmic equivalent of mutual information represents the right setup for quantifying the degree of organization of a sequence S. More precisely, he introduced the d-diameter complexity $K_d(S)$ as the minimum number of bits needed to describe S as the composition of separate parts S_i, each of diameter not greater than d,

$$K_d(S) = min \left[K(\alpha) + \sum_i K(S_i) \right], \tag{7}$$

pointed out that what the problem really was that no-one knew the question. Accordingly, the computer designed its successor, the Earth, to find the question to the ultimate answer.

where $K(S_i)$ is the algorithmic information of the single ith piece and $K(\alpha)$ accounts for the reassembling processes needed to combine the various pieces. If $d > N$, $K_d(S) = K(S)$ and $K_d(S)$ increases as d decreases. The faster the difference $\delta K(S) = K_d(S) - K(S)$ increases, the more structured and organized S is. The beauty of this approach is in the fact that no definition of the constituents is required: they are automatically identified by determining the partition that minimizes the d-diameter complexity.

In the case of a perfect crystal (i.e. a periodic self-repeating sequence), $K(S)$ is very low and $K_d(S)$ remains low, even when S is broken into various pieces, since there is no connection between the different cells. The same is true in the opposite limit of a gas-like (purely random) sequence. In this case, $K(S)$ is maximal and remains large when S is partitioned in whatever way, as all bits are, by definition, uncorrelated with each other.

3.2 DNA → RNA → Proteins

DNA (deoxyribonucleic acid) is a double-stranded polymer made of four elementary components called nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). Nucleotides are small molecules consisting of a phosphate group linked to a pentose (a sugar with 5 carbon atoms) which is in turn bound to one of the bases. The two strands interact via hydrogen bonds linking the pairs A-T and C-G. In its native state, the two DNA-strands spiral around each other and assume the well-known double helix conformation, as proposed by Watson and Crick in 1953 [25]. DNA is the carrier of the genetic information required to build a living organism. Such information is organized in units named genes which, from a molecular point of view, are sequences of DNA nucleotides capable of synthesizing a functional polypeptide. Roughly speaking a gene is a portion of DNA which encodes a protein.

RNA (ribonucleic acid) is generally a single strand polymer made of the same nucleotides as DNA, except for the replacement of Thymine with Uracil (U). RNA is the outcome of DNA transcription, and is copied by using a given portion of DNA as a template. RNAs which carry information to be translated into proteins are called messenger RNAs, mRNA. Other RNAs, such as rRNA and tRNA are involved in the translation of mRNA into proteins.

Amino acids are the proteins' building blocks; even if their number is potentially limitless, only twenty types of amino acid are involved in natural proteins. Amino acids share a common structure: each of them is made by at least one amino group -NH_2 and a carboxyl group -COOH, both linked to a central carbon atom C_α (α-carbon) which is in turn bound to a side chain (functional group or residue). It is the chemical nature of the side chains that differentiate amino acids from one another, conferring to them a structure with chemical and physical specificity. Amino acids are connected together to form the protein chain through peptide bonds, which are established by the chemical reaction between the -NH_2 group of one amino acid and the -COOH

group of another. The sequence of amino acids determines the properties and the function of a protein.

4 Proteins: A Paradigmatic Example of Complexity

In addition to being a technical term introduced in the theory of computation, *complexity* is a word widely invoked in many different contexts ranging from turbulence, to networks of any kind, spin glasses, chaotic dynamics and so on. In spite of the great diffusion of this term, no clear definition of complexity has yet been given. This will presumably remain so in the near future, since it is unlikely that so many different problems share some well-defined properties. Nevertheless, if there exists a scientific discipline where complexity is to be used, it is Biology, both for the diversity of structures existing over a wide range of scales and for the combination of several mutual interactions among the very many constituents.

Proteins have been selected for their mixed digital and analog nature and since they represent the point where the initial set of genetic instructions is transformed into a working device capable of processing information. A protein is uniquely defined by a sequence of amino acids, which in turn follows from the translation of a string of DNA. However, after a protein is linearly assembled by the ribosomes of the cell, it begins to twist and bend until it attains a three-dimensional compact structure – the native configuration – that is specific of each protein and is crucial for its biological function. Because of thermal fluctuations, the final shape is, however, not exactly determined so that a protein can be seen as a digitally assembled analog device. Once assembled, proteins represent the "working molecules", supporting and controlling the life of an organism. Structural proteins, for instance, are the basic constituents of cells and tissues; other proteins store and transport electrons, ions, molecules, and other chemical compounds. Moreover, some proteins perform a catalytic function (enzymes), while others control and regulate cell activity. Most of these processes involve many proteins of the same type at once, so that it is tempting to draw an analogy with statistical mechanics, with a microscopic level (that of the single molecules) and a macroscopic one, characterized by a few effective variables (e.g., concentrations and currents). Accordingly, biological problems are akin to non-equilibrium statistical mechanics and the relevant questions concern how the definition of specific microscopic rules translates into a given macroscopic behaviour.

The lack of theoretical tools for dealing with such systems prevents us of finding general solutions, but analogies with known problems can sometimes be invoked and many help in making substantial progress (examples are the statistical mechanics of disordered systems and reaction-diffusion equations). Moreover, modern microscopic techniques allow the visualization and manipulation of single molecules, so that it is now possible to study proteins experimentally and clarify their mutual interactions.

The first problem one is faced with is to understand how each protein finds its native configuration. Indeed, consider a protein molecule with N amino acids, and assume that there are q preferred orientations of each monomer with respect to the previous one along the chain. Then, there exist q^N local minima of the energy that can a priori be meaningful "native" states. Moreover, one can imagine that the true native configuration is that corresponding to the most stable minimum. If this picture were correct, it is hard to imagine a polymeric chain exploring the whole set of minima in a few milliseconds (the folding time can indeed be so short) to identify the absolute minimum: for a chain with $N = 100$ amino acids and $q = 3$, no more than $10^{-50}s$ should be dedicated to each minimum! This is basically the famous Levinthal paradox [40], which strongly indicates that protein folding can neither be the result of an exhaustive search nor of a random exploration of the phase space.

How can proteins find their native state within such a huge number of possible configurations? A significant step towards the answer was made by Anfinsen and coworkers. They discovered, in an *in vitro* experiment – i.e. outside the cell environment – that the enzyme ribonuclease, previously denaturated, was able to spontaneously refold into its native state when the physiological conditions for the folding were restored. This work [41], that won Anfinsen the Nobel Prize, demonstrated that the folding of protein molecules is a self-assembly process determined, at a first approximation, only by the amino acids sequence. It is not assisted by the complex cell machinery nor by enzymatic activity[6]. This great conceptual simplification in the folding problem gave great stimulus to its study. The activity was not restricted to the fields of Biology and Biochemistry, but was tackled with the methods of Physics, specifically of statistical mechanics. After Anfinsen's work, attention soon shifted towards the so-called "folding code", i.e. the basic rules through which the information stored in a one dimensional structure, the amino acid sequence (also called primary structure), encode the three-dimensional protein structure (tertiary structure). From an information-theoretic point of view, if one is to specify the native configuration out of the above mentioned ensemble of possibilities, the required information is on the order of $N \log_2 q$. This is compatible with the information contained in the DNA sequence, equal to $6N$ bits. However, no algorithm has been found to predict the tertiary structure, given the primary one: this seems to belong to the class of hard computational problems. If a shortcut exists, a solution of the *folding problem* will be more likely found by studying the underlying physics and understanding its dynamics.

The first insight into the folding mechanisms came from the observation of protein shapes. Experimental data on protein structures, collected through

[6] Actually, futher studies revealed that large-protein folding can be assisted by molecular chaperons (chaperonins) and other helper enzymes to prevent protein self-aggregation and possibly dangerous misfolding. However the role of the chaperonins that are themselves proteins is still non fully elucidated.

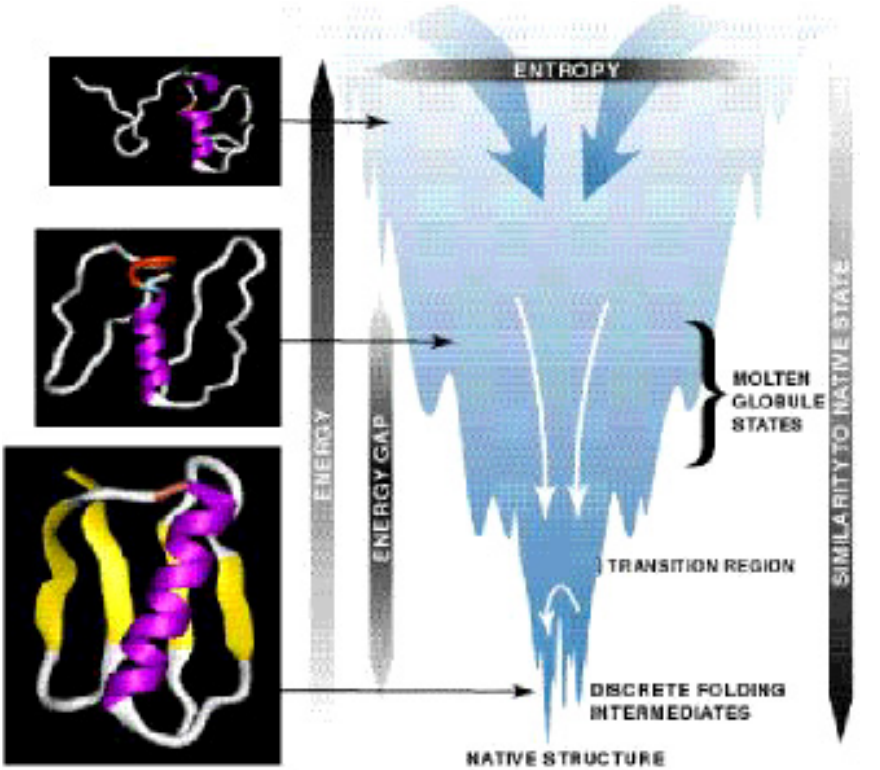

Fig. 2. A possible folding funnel scenario with the corresponding interpretation of folding stages. In the horizontal axis, protein conformations are parametrized by conformational entropy, while the energy is on vertical axis. On the side, typical protein conformations corresponding to states in the funnel.

X-ray spectroscopy and nuclear magnetic resonance (NMR), show that folded proteins are not random arrangements of atoms, but present recurrent motifs. Such motifs, forming the *secondary structure* of proteins, consist of α-helices (L. Pauling), β-sheets and loops (see Fig. 2). The secondary structure formation plays a crucial role in the folding process, since it introduces severe steric and topological contraints that strongly influence the way the native state can be reached.

Another hint about the rules that govern the folding comes from the analysis of the amino acid properties. The twenty natural amino acids can be grouped into two classes: hydrophobic and polar. While polar amino acids are preferentially exposed to water molecules, hydrophobic ones avoid contact with water; this is possible by grouping them together. As a result, most of the hydrophobic residues are buried inside the native structure, while the polar ones are located near the surface. In 1959, Kauzmann [42] realized that the hydrophobic effect is the principal driving force of the folding. However,

even if the hydrophobic collapse has a prominent role in directing the folding, it is not a sufficiently precise criterion to predict the protein structure from the knowledge of the amino acid sequence.

Earlier theoretical efforts to understand protein folding were directly aimed at bypassing Levinthal's paradox. For instance, it was proposed that a protein, during folding, follows a precise sequence of steps (pathway) to the native state without exploring the whole configurational space. This ensures a fast and large decrease of conformational entropy and justifies the relatively short folding times. However, even though it must be true that only a subset of the phase space is explored, several works on the folding kinetics revealed that folding of a given protein does not always follows the same route. The pathway scenario implies also the concept of intermediate states, i.e. states with partially folded domains that favour the correct aggregation of the rest of the protein. However the determination of intermediates is critical because they are metastable states with a relatively short lifetime.

A general theory of the protein folding requires the combination of polymer theory and the statistical mechanics of disordered systems. In fact, several features of the folding process can be understood from the properties of random heteropolymers and spin-glasses. However, there is a great difference between random heteropolymers and proteins: proteins have an (almost) unique ground state, while random heteropolymers have, in general, many degenerate ground states. In other words, proteins correspond to specific amino acid sequences that have been carefully selected by evolution in such a way that they can always fold in the same "native" configuration.

Many of these features have been investigated in what is perhaps the simplest model of a protein, the HP model [43]. It amounts to schematizing a protein as a chain of two kinds of amino acids, hydrophobic (H) and polar (P) ones, lying on a three-dimensional cubic lattice. Accordingly, the primary structure reduces to a binary sequence such as, e.g., $HPPHHHPHH\ldots$. Moreover, pairwise interactions are assigned so as to energetically favour neighbouring of H monomers in real space. Investigation of HP-type models and of more realistic generalizations has led to the "folding funnel" theory [44] which provides the currently unifying picture for the folding process.

This picture, generally referred to in the literature as the "new view", is based on the concept of free-energy landscape. This landscape neither refers to the real space nor to the phase-space, but to the space identified by the order parameter(s). In order to construct such a picture, it is first necessary to identify the proper parameters; the study of spin glasses has shown that this may not be an easy task in disordered systems. In the case of protein models, it was suggested that the proper coordinate is the fraction of correct contacts, i.e. the number of monomer pairs that are nearest neighbours both in the given and the "native" configuration. Theoretical considerations [44] based on simplified models, such as the HP model, suggest that the landscape of proteins is funnel-shaped with some degree of ruggedness (see Fig. 3). The local energy oscillations are a manifestation of frustration, a typical

Fig. 3. Ribbon representation of the protein chemo-trypsin-inhibitor (CI2), showing the characteristic secondary motifs alpha-helix and β-sheets

property of many disordered systems, here induced by the conflicting polar and hydrophobic interactions.

The funnel structure is the essential property ensuring an efficient collapse, because it naturally drives the system towards the minimum of free energy. Moreover, the protein can be temporarily trapped into the deepest relative minima, which correspond to the intermediates observed in kinetics experiments. Accordingly, the funnel scenario is able to reconcile the thermodynamic and kinetic features of the folding process.

References

1. http://www.unipd.it/musei/vallisneri/uomo/13.html
2. W. Harvey, *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus.* Frankfurt am Main, Wilhelm Fitzer, 1628
3. B.O. Kuppers, *Information and the origin of life.* The MIT Press, Cambridge, Mass., 1990
4. E. Mayr, *The growth of biological thought. Diversity, evolution and inheritance.* The Belknap Press of Harvard University Press, Cambridge, Mass, 1982
5. W.S. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115 (1943)
6. J.J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci. USA* **79**, 2554 (1982)
7. S. Leduc, *Théorie physico-chimique de la vie, et générations spontanées.* Poinant Éditeur, Paris, 1910

8. A.M. Turing, The chemical basis of morphogenesis. *Philos. Trans. R. Soc. London B*, **237**, 37 (1952)
9. M. Sugita, Functional analysis of chemical systems *in vivo* using a logical circuit equivalent. *J. Theoret. Biol.* **1**, 415 (1961)
10. S.A. Kauffman, Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437 (1969)
11. B.C. Goodwin, *Temporal organization in cells.* Academic Press, 1963
12. M. Eigen, Self-organization of matter and the evolution of biological macromolecules, *Naturwissenschaften* **58**, 465 (1971)
13. M. Eigen, J. McCaskill, and P. Schuster, The molecular quasi-species. *Adv. Chem. Phys.* **75**, (1989)
14. D. Bray, Protein molecules as computational elements in living cells, *Nature*, **376**, 307 (1995)
15. T.S. Shimizu, N. Le Novere, M.D. Levin, A.J. Beavil, B.J. Sutton, and D. Bray, Molecular model of a lattice of signalling proteins involved in bacterial chemotaxis. *Nat Cell Biol* **2**, 792 (2000)
16. S.S. Chetverikov, *On certain aspects of the evolutionary process from the standpoint of modern genetics. Zhurnal Ekspeimental'noi Biologii*, **A2**, 3 (1926). Translated by M. Barker, I.M. Lerner Editor, *Proc. Am. Phil. Soc.* **105**, 167 (1961)
17. R.A. Fisher, The wave of advance of advantageous genes, *Ann. Eugenetics* **7**, 353 (1937)
18. A.N. Kolmogorov, I.G. Petrovskii, and N.S. Piskunov, Étude de l'équation de la diffusion avec crissance de la quantité de matière e son application à un problème de biologie. *Moskow Univ. Bull. Math.* **1**, 1 (1937)
19. A.N. Kolmogorov, On a new confirmation of Mendel's Laws. *Dokl. Akad. Nauk. SSSR* **27**, 38 (1940)
20. A.J. Lotka, Undamped oscillations derived from the law of mass action. *J. Phys. Chem.*, **14**, 271 (1920)
21. A.J. Lotka, *Elements of physical biology.* Williams and Wilkins, Baltimore, 1925
22. A.N. Kolmogorov, Sulla teoria di Volterra della lotta per l'esistenza. *Giorn. Ist. Ital. Attuar.* **7**, 74 (1936)
23. J.D. Murray, *Mathematical Biology.* Springer, Berlin, 1993
24. C.E. Shannon, A mathematical teory of communication. *The Bell System Technical Journal* **27**, 379 and 623 (1948)
25. J.D. Watson and F.H. Crick, Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171**, 737 (1953). Classical article, *Ann. N.Y. Acad. Sci.* **758**, 13 (1995)
26. L. Pauling and R.B. Corey, Two hydrogen-bonded spiral configurations of the polypeptide chain. *J. Am. Chem. Soc.* **72**, 5349 (1950)
27. E. Schrödinger, *What is Life.* Cambridge University Press, Cambridge, 1992. original appeared in 1944
28. T.E. Creighton, *Proteins: Structure and Molecular Properties.* W.H. Freeman and Company, New York, 2000
29. H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology.* W.H. Freeman and Company, New York (2000)
30. D. Holste, I. Grosse, and H. Herzel, Statistical analysis of the DNA sequence oh human chromosome 22. *Phys. Rev. E* **64**, 041917 (2001)

31. C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley, Long-range correlations in nucleotide sequences, *Nature* **356**, 168 (1992)
32. R.J. Solomonoff, A formal theory of inductive inference, part i. *Inform. Contr.* **7**,1 (1964)
33. R.J. Solomonoff, A formal theory of inductive inference, part ii. *Inform. Contr.* **7**, 224 (1964)
34. A.N. Kolmogorov, Three approaches to the quantitative definition of information. *Problemy Peredachi Informatsii* **1**, 3 (1965). Original in russian, translated in: Three approaches to the quantitative definition of information. *Probl. Inform. Trans.* **1**, 1 (1965)
35. A.N. Kolmogorov, Logical basis for information theory and probability theory. *IEEE Trans. Inform. Theory*, **14**, 663 (1968)
36. G.J. Chaitin, Information-theoretic computational complexity. *IEEE Trans. Inform. Theory* **20**, 10 (1974)
37. G.J. Chaitin, Randomness and mathematical proof. *Scientific American* **232**, 47 (1975)
38. H.J. Bussemaker, Hao Li, and E.D. Siggia, Building a dictionary for genomes: identification for presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA* **97**, 10096 (2000)
39. G.J. Chaitin, *Toward a mathematical definition of "Life"* in R.D. Levine and M. Tribus, The Maximum Entropy Formalism, MIT Press, 1979, 477
40. C. Levinthal, Are there Pathways for Protein Folding? *J. Chem. Phys.* **65**, 44 (1968)
41. C.B. Anfinsen, Principles that govern the folding of protein chains. *Science* **161**, 223 (1973)
42. W. Kauzmann, Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**, 1 (1959)
43. H.S. Chan and K.A. Dill, Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. *Proteins* **30**, 2 (1998)
44. P.G. Wolynes, J.N. Onuchic, and D. Thirumalay, Navigating the folding routes. *Science* **267**, 1619 (1995)