The problem of Form in molecular biology

Laura Nuño de la Rosa & Fernando M. Pérez Herranz

In *Philosophical Essays on Physics and Biology*, ed. by J. L. González Recio (Hildesheim, Zürich, New York: Georg Olms Verlag, 2009), pp. 128-150.

0. Introduction

Since the Ancient Greeks, the category of Form has been a fundamental explanatory tool in the understanding of living beings. However, we are usually told that the great scientific revolution in the history of biology consisted on the triumph of efficient causality upon both formal and final causality in the understanding of natural phenomena. Just like Newton was able to explain movement by means of an external force, Darwin managed to explain evolutionary change in virtue of an also external and efficient cause: Natural Selection (Depew & Weber 1995). Nevertheless, the problem of morphogenesis, i.e. the generation of organismal form throughout development, was still explained in terms of formal and final causality. Yet, the special status of the theoretical framework of embryology did not remain for too long. The discovery of the chromosomal determination of morphological traits and the rediscovery of Mendel's laws inaugurated a reductionist program that took all causal power away from Form. Finally, the finding of the genetic code reduced Form to an algorithmic sequence, defined by the linear disposition of amino acids in DNA and interpreted as a Turing machine tape (Adleman, 1994). It seemed eventually possible to explain the realm of organic forms without paying any attention to Geometry.

The geometrical approach to biology persevered in some isolated spirits such as D’Arcy Thompson, Richard Goldschmidt or Nicolas Rashevksky, but it was not until the 1970s that Form was integrated into the explanatory framework of biology by a more or less homogeneous group of theoreticians. Moving along different points of view, authors like Conrad H. Waddington, René Thom, Stuart Kauffmann, Brian Goodwin or Stephen Jay Gould proposed an internalist and morphological explanation of living forms. The so-called *structuralist school* rejected the attempts to codify forms in the unidimensional nucleotide chain, stressing the epigenetic character of development and the resulting discreteness of 'morphospace', the space of possible biological forms. The category of Form was, for these authors, an irreducible explanatory resource.

Due to the hegemony of Genetics, the vindication of Form was mainly posed against the belief in the molecular determination of morphology. Consequently, the debates about the role of Form are usually found in the context of the confrontation between antireductionist organismal approaches and reductionist molecular approaches to biology. However, it is not obvious at all that molecular biology implies by nature an 'anti-morphological' vision of the organic realm. On the contrary, we claim that the problem of Form is also present in the very field of molecular biology in many unresolved manners. Moreover, this presence allows us to vindicate an irreducible explanatory role for the category of Form in molecular biology.

Throughout this paper we shall not be concerned with the relationship between molecular and organismal parts. It is rather the discussions around the nature of the very biological macromolecules, and the way in which the category of Form becomes an integral part of the understanding of their nature, that we shall focus on. *Structural biology* is the branch of molecular biology concerned with the study of the structure and shape of biopolymers (DNA, RNA, proteins). It is within this field that we will explore how the debate on the role of Form, which started at the macroscopic scale with the question on the organismal and taxonomical forms, is also more or less explicitly present at the microscopic level[1].

Furthermore, we will try to show how molecular biology turns out as a privileged field to clarify the classical philosophical problem of Form and its relationship with Matter and Function. All these terms have a contemporary biological translation into the also problematic triad conformed by genotype, phenotype and function. In the molecular realm, DNA and RNA are conceived as strings of nucleotides (what is often referred as *genotype*) directly mapping into sequences of amino acids in proteins, but they only acquire a *function* when folded into a three-dimensional structure (*phenotype*). Since all these aspects (genotypic, phenotypic and functional) converge into the same entity (the molecule), biological macromolecules are exemplar objects for theoretical biology (Stadler et al. 2001) and for analysing the explanatory role of Form in biology.

Our philosophical exploration will be stringed together by the analysis of the experimental techniques and theoretical models used for the determination and explanation of molecular forms. Our goal is to unravel the ontological consequences of structural biology research programs in order to confront them with the orthodox reductionism of molecular biology. Throughout the chapter, we will see how the theoretical questions of the morphological tradition and the objections it posed to reductionist approaches reappear when we go deep into the molecular forest.

On the first section, we analyse questions about the *definition* of molecular form arising from different structural analysis techniques. On section two, we explore the different attempts to *explain* the generation and maintenance of molecular forms in comparative and computational biology. Finally, we move into the many ways in which Form relates to Function in the molecular realm.

1. The definition of biomolecular form

What is Form? and how can we know it? These are classical philosophical questions that have survived in Natural History and modern biology with special resistance. The definition of Form and the characterization of morphological diversity have been the main goals of Comparative Anatomy and Theoretical Morphology. But with the advent of Modern Synthesis, both disciplines were neglected as sources of biological explanation and their theoretical scope was drastically reduced. For instance, the *morphological* definition of species was replaced by Mayr's biological definition in terms of *reproductive isolation*, and the concept of *type* replaced with the concept of *common ancestor*. In a nutshell, morphological approaches to biology were branded as typological and 'population thinking' was considered the only legitimate access to biological problems. This claim was challenged by the structuralist school, which built a new morphological agenda for biological research, committing biology to address new morphological questions and old-neglected ones, such as the stability of body plans, the

[1] In this paper we are dealing with a synchronic approach to biomolecular forms. We leave for further occasion the analysis of the evolutionary consequences of the structuralist approach to biology (developmental constraints, phenotypic stability, homology, punctuationism, novelty), their conceptual parallels in molecular biology and their philosophical implications.

problems of homology and modularity or the properties of morphospace. In this section, we will see how this new set of problems has also arisen in molecular biology.

1.1.The forms of Form: shape and structure.

From the organismal and taxonomical point of view, Form has been traditionally understood in two ways: as shape and as structure. *Structure* refers to the topological properties of organic systems and the spatial arrangement of organs, whereas *shape* refers to the 'contour' that encloses both a living being and its macroscopic parts. Both approaches to organismal form conformed the two great morphological programs of the history of biology. On the one hand, the topological approach to *structure* was the great theoretical discovery of Geoffroy Saint-Hilaire (1772-1844), who tried to reduce the vast morphological diversity to a unique abstract type that varied in shape and function but not in the topological arrangement of its constituent parts. On the other hand, the study of *shape* was one of the main goals of D'Arcy Thompson, who, for instance, was able to reproduce many fishes' shapes by means of mathematical transformations (D'Arcy Thompson 1917).

Biological macromolecules have also shape and structure, two aspects revealed by different structural analysis techniques: Scanning Probe Microscopy for the case of shapes and X Ray crystallography and Nuclear Magnetic Resonance Spectroscopy for the case of structures. We will focus on protein form in order to illustrate the operation of both techniques and the derived ontological consequences.

The vision of proteins as surfaces or contours has recently become possible thanks to Scanning Probe Microscopy (SPM), which scans the samples' surfaces with a mechanical probe. This visualization of biomolecules has been essential for the understanding of their function, for it represents proteins as they are 'seen' by another protein. As we will see later, proteins have a hydrophobic inner core that is not accessible to other proteins. Functional interactions are thus produced between complementary surfaces whose structural scaffolding is negligible. And this is precisely what SPM allows us to observe.

As we can see in **Figure 1**, protein structure has many aspects: amino acid sequence (*primary structure*), regularly repeating local structures such as α-helices and β-sheets (*secondary structure*); and the overall *tertiary structure* (also called *native conformation*) consisting of the spatial relationship among secondary structures (the final shape that the full protein adopts). Until the first diffraction pattern of a protein was obtained in 1936, it was thought that proteins were unstructured random coils. Thus, except for the primary sequence, none of the afore-mentioned structural aspects (secondary and tertiary) of proteins was known. The main advance was thus provided by the use of radiation sources with a wavelength similar to atomic dimensions. This is the case of X Rays, electrons and neutrons, all of them discovered between the end of the 19th and the beginning of the 20th century and applied to macromolecular structures decades after their discovery[2]. Since the determination of the three-dimensional structure of myoglobine by Max Peruz, X Ray crystallography became the main structural analysis technique. In a nutshell, the method lies in determining the arrangement of atoms in a protein crystal from the way in which X Rays are dispersed by the crystal's electrons. In a lesser degree, structures are also determined by Nuclear Magnetic Resonance Spectroscopy. NMR is based on the fact that certain atomic nuclei subjected to an external magnetic field absorb electromagnetic radiation in RF region. Since the exact frequency of this absorption depends on the surrounding of these nuclei, it can be used to determine the structure of the molecule.

[2] For a comprehensive review of the history of structural biology see Campbell 2002.

Both techniques offer complementary information about macromolecules and are usually applied in concert. X Ray crystallography is the most powerful structural analysis tool, but NMR is essential to determine mobile regions that do not easily crystallize. However, as we will see in the next section, these two approaches to protein structure have also given rise to different ontological commitments regarding the nature of molecular form and the role played by movement in its comprehension.

1.2.Static and Dynamic Forms.

The relationship between organic form and movement has been a classical controversial topic in philosophy of nature and theoretical biology. Embryology oriented researchers have traditionally outlined the dynamic character of organic forms, whereas natural historians devoted to the organization of morphological diversity have usually offered a static geometrical or topological view of them. In the 19th century *Naturphilosophen* revolted against the statical forms graved in Natural History treatises and vindicated a dynamic form irreducible to the stages it traverses throughout its development. Current embryologists stressing the need to take into account the organisms' whole life-cycle of an individual from fertilization to death in order to understand morphologies make a similar point (Hall 1999).

This very same controversy between the explanatory preference of static or dynamic forms also permeates biology at the molecular scale. Protein form has traditionally been considered a single state with a well-defined tertiary structure, as determined by X Ray crystallography. But as in the case of organismal biology, where the anatomical and the embryological approaches to organic forms determined the preference for structures or processes, the inclination to the static or the dynamic vision of molecular form heavily depends on the structural analysis technique. Since X Ray crystallography studies molecules in solid phase, it offers a rigid static image of molecular forms. On the other hand, NMR works in liquid phase, providing images of proteins in very different conformations, and thus permitting to conceive proteins as dynamical systems (see **Figure 2** for a comparison of the two kinds of representation of a protein structure: a static representation on the left (**2a**) and a dynamic one on the right (**2b**), pictured as a superposition of the multiple conformations of the same macromolecule).

Despite the fact that X Ray crystallography and NMR were almost simultaneously discovered, the early success of crystallography on delivering reliable results slowed down the use of NMR and, consequently, the dynamic conception of Form. The structures obtained by crystallography gave rise to a *geometric-based* definition of protein conformation as a set of atomic coordinates. The statical structure of these 'frozen' structures became a model for the understanding of the nature and function of molecular forms. However, since the late 1990s, biologists working with NMR spectroscopy began to insist on the insufficiency of the information given by the statical three-dimensional structures, as shown by X Ray crystallography (Dobson & Hore 1996; Lewis 1998). NMR images revealed a high structural flexibility, ranging from side chain rotation to complete rearrangement of secondary structure elements.

The recognition of the role of dynamics has lead to two great ontological inferences. On the one hand, Denton and co-workers consider structural flexibility as a demonstration of their Platonic conception of natural forms. On their view, the fact that proteins can keep on folded despite permanent conformational perturbations demonstrates the special nature of organic forms, whose robustness reveals an ideal essence unknown in artificial objects (Denton et al. 2002). On the other hand, more experimentally oriented biologists have proposed to replace the geometric-based definition of molecular structure by a thermodynamics-based one. Thus the three-dimensional structure is considered

> “not as a specific microstate, but as a macrostate, which can be envisioned as a collection of microstates separated from each other by low-energy barriers. According to this view, a protein conformation is a continuous subset of the conformational space (i.e., a continuum of well-defined configurations) that is accessible to a protein confined to a certain local minimum.” (Kaltashov & Eyles 2005)

We have just seen how the two great approaches to organic form have met again at the microscopic level of molecular biology. But what about the parts conforming these macromolecular wholes? Can we say that biomolecules have parts as the organisms have organs? And in this case, how do they relate to each other and to the whole they conform? Can forms be reduced to their component parts?

1.3.The morphological whole and its morphological parts

The paradoxical relationship between organic wholes and their morphological units is another classical topic in the history of Morphology. Organic forms are decomposable into other morphological parts, but, at the same time, it appears that they cannot exist outside the whole they integrate. In contemporary terms, “phenotype is neither atomistic nor holistic, but modular.” (Griffiths 2002)

The fact that phenotypic wholes can be decomposed into parts dates back to Aristotle, and has gained renewed attention under the current debate on *modularity* (see Callebaut & Rasskin-Gutman 2005). Organisms are organized into 'modules' that can be defined as “cohesive units of organismal integration.” (Eble 2005) In this view, modules arise from stronger interactions within than among other modules, whereas organismal integration reflects differential interactions among them.

Again, the distinction between organic wholes and modules is not absolute but relative to a certain scale of organization. Although traditionally defined at the morphological scale, the concept of modularity is now assumed to occur at different levels of the biological hierarchy, ranging from systems and organs to genetic networks and molecules (Abouheif 1997). Proteins are made up in a 60 per cent by a very reduced number of local structural motifs stabilized by hydrogen bonds (α-helices, β-sheets, and turns) and domains, which conform the secondary structure. The tertiary structure is achieved through the three-dimensional arrangement of these elements via different coordinate kinds of interactions, mainly hydrophobic forces, but also salt bridges, hydrogen and disulfide bonds, and post-translational modifications.

At first it was thought that a great amount of structural motifs was to be discovered and that their combination would give rise to an infinite number of protein structures. But it has been found that many proteins adopt similar common structural motifs resulting from combinations of a limited set of secondary structure elements[3], as we can schematically see in **Figure 3**. The metaphor of Nature as a tinkerer and not as an inventor (Jacob, 1977), vindicated by the structuralist school at the organismal scale, reappears again in molecular biology.

However, despite the ontological autonomy attributed to organic parts or modules, the holistic nature of organic forms has been traditionally claimed to be an elementary difference between artificial and natural beings. This quasi-independence of parts lead Aristotle and Kant to outline the need to take into account the organismal context in order to explain the parts which make it up. Similar claims have been made in the

[3] Steric restrictions limit the conformational volume accessible to proteins, what is usually represented using conformational maps or *Ramachandran plots* (Ramachandran et al. 1963)

context of the reductionism-holism debate in current biology (see, e.g. Laubichler & Wagner 2000)

We have just seen how proteins are conformed by a very reduced number of motifs that result on higher order structures. But, as in the case of the organs within a whole organism, molecular modules depend on their being part of the protein's native conformation, outside of which they have no independent existence. This has lead to recover a holistic conception for these macromolecular entities:

> “Proteins, like sentences, are intensely holistic entities. All the current evidence suggests that the various parts of the fold [...] exert what appears to be a mutual and reciprocal formative influence on each other and on the whole, which itself in its turn exerts a reciprocal formative influence on all its constituent parts.” (Denton et al. 2002)

We will see this issue reappearing in the debates about the generation of molecular forms. But for the time being, we just want to outline how the ghost of holism has not been exorcized from molecular entities. There cannot be parts outside of biomolecular wholes, and this requires a return to the category of Form as the holistic morphological context where parts are con-formed and 'naturalized'.

2. The explanation of molecular form

Up to now, we have dealt with macromolecular entities from a strictly morphological point of view, asking questions about the very nature of the structures and shapes of biopolymers and their 'morphological' parts, without taking into account their ‘material’ substrate, i.e. the nucleotides and amino acids as component molecules[4]. This brings us to the great challenge of molecular biology, namely its alleged ability to reduce biological forms to the sequence that 'codifies' them. This view of macromolecular wholes as the ‘sum’ of their amino acid parts has been called the 'Linear Sequence Hypothesis' (LSH) and instantiates the more general issue of reductionism, understood as a question about the relationship of parts to wholes (Bechtel & Richardson 1993).

But the LSH is, actually, a two-fold statement (Hüttemann & Love, Unpublished). On the first place, the 'information construal' of the LSH is a question of whether native protein conformation (whole) can be inferred from the linear sequence of amino acids (parts). It is a question about the *identity* for the folded protein, which is consider to be predictable solely from its sequence. On the second place, the 'process construal' of the LSH is a question of whether there is a *causal* part-whole reduction of how the three-dimensional structure is generated, defending that sequence contains the necessary and sufficient information to specify the generative process. The first question construes the problem of reduction in molecular biology in terms of *prediction,* whereas the second one glosses it as a question of *explanation*. We dedicate the two sections of this chapter to each one of these problems.

2.1.Matter and Form, Sequence and Structure

Despite the celebrated image of the double helix, DNA's geometry had no influence on the re-conceptualization of heredity that determined the research program of molecular biology. Indeed, as soon as the fundamental DNA structure was found, molecular biology focused on the research of the sequence[5]. Even the double helix suggested “a possible copying mechanism for the genetic material” (Watson & Crick 1953), the copy

[4] We take the notion of ‘matter’ on the Aristotelian sense: not as rought stuff or piece of mass, but as the set of specific compositional elements whose spatial organization gives rise to morphological wholes.

mechanism (and the morphological aspects involved in the process) had no influence in the definition of the information to be copied. The discovery that "all genes have roughly the same three-dimensional form told [biologists] that the differences between two genes reside in the order and number of their four nucleotide building blocks along the complementary strands" (Alberts 2003: 97). In 1958, Crick defined biological 'information' as the sequence of the bases in the nucleic acids and of the amino acids in proteins, and established the celebrated 'central dogma of molecular biology': "[sequential] information cannot be transferred back from protein to either protein or nucleic acid" (Crick 1958). In this way, the research program of molecular biology became highly constrained, to the extent that proteins were consigned to be a simple product of the expression of genetic code. Therefore it might be argued that the discovery of DNA's three-dimensional structure became a great discovery to be immediately forgotten and included as a *ceteris paribus* clause that made sequences the variational elements by means of which biological phenomena should be explained (and themselves in need of explanation by Natural Selection).

However, from the 1970s onwards the sequence-based definition of biological information has been strongly put into question. At the macroscopic level, whether organismic morphologies (organs, limbs, segments, etc.) could be reduced-to or directly deduced-from the molecular level (particularly from DNA sequences) has become a topic of much debate (see, e.g. Alberch 1982; Goodwin & Saunders 1989). What has received fewer attention is the prior assumption that sequence (of either nucleotides or amino acids) can, by itself, specify morphologies (either DNA or protein conformation) at the very molecular level. In other words, the question about the possibility of a DNA-to-organism mapping has obscured the more fundamental question of the very possibility of a direct sequence-structure mapping.

The sequencing of genomes and proteins, on the one hand, and the determination of biomolecular structures by X Ray crystallography and NMR, on the other hand, constitute the two great research programs of molecular biology. Both sequencing and structural analysis techniques began to develop in the 1960s and 1970s. However, as of February 2008 there were 82,853,685 sequence records at the GenBank database (GenBank 2008), and the number continues to grow exponentially. Yet, as for the 3th June 2008, there were only 51,155 entries at the Protein Data Bank (Protein Data Bank 2008). To be sure, this asymmetry is partly due to technical reasons: DNA is easily sequentiable, whereas the techniques for structural analysis are much more expensive and time-consuming. But it also reveals the theoretical bias of molecular biology. Nevertheless, the determined structures are numerous enough so as to compare them with the corresponding sequences in order to check if the results of this mapping correspond with the theoretical assumptions of the reductionist program.

The relationship between sequence and structure, or between genotype and phenotype, has been modelled as a function that ties the set of sequences to the set of structures. Depending on how this function is formulated, it can give rise to two kinds of spaces: metric and non-metric. *Metric spaces* involve a complete symmetry between both sets, i.e. each sequence codifies for a single structure. This is the kind of space underlying the strong research program in molecular biology. *Non metric spaces*, on the contrary, have the fundamental property vindicated by the structuralist school at the organismal and taxonomical scales, namely that the relationship between sequence (genotype) and structure (phenotype) is non symmetric: for each structure, there are one or more

[5] Actually, DNA double helix was not taken seriously until DNA replication was connected to protein synthesis. Up to then, there were two contrasting theories under discussion on protein synthesis: the peptide theory, where proteins were thought to be made by coupling of many peptide units; and the template theory, involving synthesis on genetic templates (Olby 2003).

sequences leading to it (Stadler et al. 2001).

In a series of papers, Schuster, Fontana and co-workers have explored the RNA's morphospace properties, studying the relationship between RNA sequences and RNA secondary structures (see, e.g. Fontana & Schuster 1998; Schuster 2001). The systematic exploration of this sequence-structure map has proved that the number of sequences is much bigger than the number of secondary structures. This result may be easily generalizable to other biological macromolecules, such as proteins, where very similar tertiary structures can be adopted by quite dissimilar primary sequences (Zhang & DeLisi 2001). “Therefore, the fold universe appears to be dominated by a relatively small number of giant attractors, each accommodating a large number of unrelated sequences” (Kaltashov & Eyles 2005; see Hou et al. 2003 for a global view of the 'protein structure universe').

The theoretical consequences of the asymmetry of the sequence-structure map are immense. If we consider the problem of Form at the unidimensional space of the nucleotide and amino acid chains, then there are almost infinite combinatorial possibilities, and the questions around the unity and diversity of organic forms should be posed in the field of comparative genetics. But if we treat the problem at the three-dimensional scale of protein structures (rather than sequences), there appears to be a constrained number of possible morphologies. In this view, Morphology cannot be reduced to information as measured on a nucleotide-bit way. In other words: there seems to be a 'morphological information' that is not reducible to sequential information. Only a full understanding of the mechanisms governing the transformations that lead from sequence to structure will be capable of delivering an appropriate understanding of the 'morphological code', as claimed by epigenetic developmental biology (see, e.g. Waddington 1962).

2.2.The generation of Form: the protein folding problem.

The recognition and characterization of the asymmetry of the sequence-structure map, leads immediately to investigate the causes of this asymmetry. Modern Synthesis explained it in terms of convergence: Natural Selection has conserved the same adaptive structures codified by different genes. As we mention before, evolutionary developmental biology criticized this externalist approach and defended that the discreteness of morphospace is *the result* of the internal properties of the developmental system which generate the morphospace (Alberch 1980). In other words, if developmental systems constraint the possible morphologies, then the causes of the structural identities should be looked for in the very developmental processes. In fact, the conviction that the generation of Form is fundamental to understand the resulting morphologies roots in the 19th century. This was the main leitmotiv of many transcendental morphologists such as Lorenz Oken, Johann Friedrich Meckel or Étienne Serres, and of the school of Evolutionary Morphology lead by Ernst Haeckel and Francis Balfour. However, after the triumph of Genetics, the link between development and morphology disappeared from the field of mainstream biology (Hamburger 1980): if sequences contained all the construction rules necessary for the building of morphologies, development could be ignored as an epiphenomenon (Gould 1977). However, the 'morphogenetic school', integrated by embryologists such as Joseph Needham or C. H. Waddington, survived as a minority group which kept on investigating the epigenetics of development. Current attempts to build the developmental bridge between genes and forms can be explored in Forgacs and Newman's *Biophysics of the Developing Embryo* (2006).

Despite the views of development as a molecular 'computing' of the embryo (Rosenberg

1997), the big question to be solved in developmental biology is still finding out "the function that maps molecular input into embryological output." (Laubichler & Wagner 2004). The same problem appears in molecular biology when trying to understand the three-dimensional folding of nucleotide and amino acid chains. This time the question is not to be solved in the field of experimental molecular biology, but mainly in the field of computational molecular biology. One of the most challenging goals of current bioinformatics is precisely "to devise a computer algorithm that takes, as input, an amino acid sequence and gives, as output, the tertiary structure of a protein." (Dill et al. 2007). Indeed, the Protein Folding Problem (PFP) is not just the great defy of molecular biology, but it is considered to be one of the biggest unsolved problems in science (*Science* 2005). From a philosophical point of view, the PFP is also a fascinating topic, since it resumes in a concrete and limited way the philosophical problem of reductionism (Sarkar 1998).

Despite acknowledging that neither the mechanisms of structural stability nor the protein folding processes are well understood, molecular biology textbooks state that "*[a]ll of the information necessary for folding the peptide chain into its 'native' structure is contained in the amino acid sequence of the peptide.*" (Garret & Grisham 1999: 161. Italics in the original). This version of the Linear Sequence Hypothesis states that correct folding is solely a function of the linear order of amino acid components, incorporating the time dimension to the problem of reduction: the properties of the parts (amino acids) at *t* cause the whole (the folded protein) to have some properties at a later time *t+1* (Hüttemann & Love, Unpublished). The justification of this statement is usually traced back to Anfinsen's experiments at the early 1960s, which demonstrated that most proteins can, in vitro, fold back to their original conformation without being aided by any cellular machinery[6]. It is interpreted that "[i]n such experiments, the only road map for the protein, that is, the only 'instructions' it has, are those directed by its primary structure" (Garret & Grisham 1999: 161).

Hüttemann and Love have dealt with the PFP in the context of the debate about the causal role of intrinsic and extrinsic factors, putting into question the 'spontaneity' of the folding process and thus the possibility of a reduction explanation. The LSH states that the native conformation is determined by the intrinsic properties of the amino acid sequence in a given environment. However, folding has been proved to be critically dependent on the 'normal physiological medium', which includes not just physico-chemical components of this medium, but also other macromolecular structures, specially the chaperone proteins in charge of guiding protein folding (Hüttemann and Love, Unpublished).

We do agree with this objection, but for our goals, the comprehension of the emergence of Form in the folding process, does not just rely on the necessity of extrinsic components but on the appearance of new irreducible topological relations in the three-dimensional biological space. In order to test whether these factors are taken into account in current biological practice, we will explore current computational methods of protein folding prediction. As we shall see, and despite of programmatic statements, current structure prediction methods do not solely consider the components participating in folding, but also the 'structural landscape' emerging from the very process of folding.

The three-dimensional structure of a native protein in its physiological milieu represent the free energy minima among all possible states. Thus, protein structure prediction

[6] Anfinsen added denaturants (such as urea) to ribonucleases, which caused them to loose tertiary structure and revert to a random coiled state. After removal of the denaturants, proteins spontaneously folded back into the native conformation.

methods consist of finding a *search strategy* to explore the space of possible structures, and an *energy function* to identify the optimal structure (recent books reviewing protein structure prediction methods are Webster 2000, Tramontano 2006, Zaki & Bystroff 2007). There are two great strategies in protein structure prediction: Homology Modelling and *De novo* Modelling. Homology Modelling or comparative structure prediction exploits the similarity between the protein whose folding is to be predicted and an homologous protein of known structure (see Nayeem et al. 2006 for a review). On the contrary, *De novo* Modelling does not make any assumption about the final state and thus involve a much harder task; that of searching through the space of a large number of possible structures, requiring a huge computational cost. Although Homology Modelling is the most successful tool, it is the *De novo* method that remains of philosophical relevance for us, since we are interested in how a protein comes to reach an unknown state.

The narrowing of conformational possibilities demanded by the so-called 'Levinthal paradox' is one of the main goals of *de novo* computational biology. In 1968 Cyrus Levinthal pointed out that if a protein had to reach its most stable conformational state by sampling all the possible conformations, the time required for a real-time exploration of all the possibilities would be longer than the age of the universe (Levinthal 1968). The fact that proteins attain their native states so quickly lead to conclude that they cannot do so by a random search through all possible pathways. It is at this point where the type of questions raised by scientists devoted to the protein folding problem become very similar to those posed by embryologists:

> “Do proteins take shape gradually or in fits and starts? Is there only one folding sequence for each protein? How sensitive is folding to cellular conditions? What comes first - an "outline" of the shape or its details?” (Jayant's web page)

In fact, the two great strategies essayed to solve Levinthal paradox have ‘recapitulated’ the main approaches to the explanation of development we find throughout the history of embryology. Both development and protein folding have been seen either as (i) a series of developmental stages or (ii) as the exploration of a landscape of possible conformations leading to the final three-dimensional form.

At first, Levinthal paradox led to hypothesize that proteins fold by specific 'folding pathways' implying intermediate, partially folded conformational states. It was thought that protein segments independently adopt local secondary structures (the α-helices and β-sheets), which in turn depend on its amino acid composition. In this approach, secondary structures form first and then interact to build tertiary structures. This 'hierarchical view' of protein folding fitted perfectly in the LSH: sequence was thought to determine structure in a linear, local and unidirectional way and secondary components were thought to aggregate in a lego-like manner.

However, the 'hierarchical view' was built upon a methodological limitation: traditional experiments worked only with average quantities and so they were unable to detect individual folding processes. Latter, statistical mechanical modelling permitted to recognize that folding does not involve a single folding pathway, but a potential funnel-shaped *energy landscape* (Baldwin 1995) in the conformational space (Kaltashov & Eyles 2005). This ‘new view’ of protein folding can be summarized in two principles: First, folding may proceed through multiple pathways, rather than a single route. Second, regardless of the starting point, the conformational space is progressively funnelled.

Both facts regarding the generation of Form were first recognised in developmental biology. On the first place, the fact that proteins can achieve their native structure following different folding pathways, is parallel to the embryological phenomenon captured by the notion of 'morphogenetic field', developed to explain the fact that a same organ can be formed throughout different developmental pathways (Spemann 1938; Goodwin 1982; Gilbert et al. 1996). On the second place, energy landscapes are analogous to the *epigenetic landscape* by means of which Waddington explained development. As cells at the early development, protein chains can adopt multiple forms at the beginning of the folding process. But throughout the process itself, the conformational space of both cells and proteins is progressively reduced until the final form is achieved.

We see that in this view the generation of form and thus of the very nature of molecular form is not reducible to the understanding of the sequence properties. The explanatory resources must include not just the components of the physiological medium, but the emergent tertiary interactions that stabilize protein native states and, more importantly, the understanding of how the conformational space is generated and transformed throughout the very process of folding.

3. Form and Function

After our examination of the nature and generation of molecular form, we are ready to examine one of the most intriguing topics of the history of philosophy and biology: the relation between Form and Function. What is the relation between Form and Function in the biomolecular realm? Are they reducible? An if so, in which direction? This is probably the main topic stringing together the full history of biology (see Russell 1916). Since the triumph of Darwinism, morphologies were reduced to be the result of the gradual action of Natural Selection, and 'externalist functionalism' became the main approach to the explanation of Form. From the 1970s onwards, the hegemony of this paradigm was challenged from the point of view of both Function and Form. On the one hand, functional morphology (Arnold 1983) explored an internalist conception of function, understood in a biomechanist manner: How do morphologies give rise to their functions? On the other hand, the importance of Form (and the finding of the possible or available developmental forms) in constraining the action of Natural Selection became a focus of attention: in which sense can morphological factors constrain the attainment of certain functions? Both questions find equivalent counterparts in molecular structural biology.

Focusing on proteins, the chief actors within the cell, it is universally accepted that only when they are in their native structure they are able to work efficiently: catalysing chemical reactions as *enzymes,* participating in *cell signalling* and *signal transduction* or conferring rigidity to the cell as *structural proteins*. However, the recognition that only folded proteins are functional, comes always with the reminder that structure depends on the sequence. Thus, once again, it is inferred that function is determined by amino acid sequence and, at a last resort, by the nucleotide sequence which codified the polypeptide chain. So the principle/dogma heading all textbooks of molecular biology is that “Function derives from three-dimensional structure and the three-dimensional structure is established by the amino acid sequence.” (Lodish 2005: 60).

Nevertheless, we have seen the many ways in which structure is not reducible to sequence. Similar arguments can be found at the level of functions. For instance, mutation experiments have given rise to many 'neutral' proteins, where the functional properties were not affected by amino acid substitutions (Watson et al. 2004). And the other way round: comparative genomics has shown many cases where homologous proteins have different functions.

All these evidences strongly suggest that protein function must be understood in terms of its spatial organization and not just of its sequential composition. This is the goal of this section. Firstly, we analyse how Form can have a meaning in itself and thus be significant for other forms with which it interacts. Secondly, we investigate how Form can impose topological constraints for the fulfilment of the function of other forms.

3.1.Structure, dynamics and function

It has often been argued that forms are functional in and of themselves. Regardless of how they are achieved, forms have 'semantic' properties which make them functional. For instance, many functional properties of the cell have to do with its spherical shape, which permits to maximize the volume-surface ratio. A similar relation between Form and Function can be predicated of many biomolecular structures. One of the best examples is the double helix structure of DNA and its replicative function. As it is apparent in **Figure 5**, if the two phosphate backbones are pulled apart, each strand can then be used as a template for a new base-pair complementary strand.

But the most interesting examples for analysing functional forms are those related to the coupling of forms in functional interactions[7]. The study of the functional and mechanical relationships among structures and of the link between morphology, performance and fitness (Arnold 1983) has become an active area of research (see Kingsolver & Huey 2003 for a review). In the field of molecular biology, the morphological reading of molecular interactions is fundamental, as it is demonstrated by the omnipresent metaphor of the lock-and-key fitting in the explanation of enzymatic reactions (e. g. Watson et al. 2004: 49).

It is said that the dynamical view of protein structure has challenged the 'function-structure' paradigm according to which "the enzyme was a rather rigid negative of the substrate and that the substrate had to fit into this negative to react" (Kaltashov & Eyles 2005). According to the alternative 'induced fit theory', the reaction between the enzyme and substrate occurs after a conformational change induced by the ligand. Protein function is thus related to conformational changes: "Like the Greek sea god Proteus, who could assume different forms, proteins act through changes in conformation" (Garret & Grisham 1999: 168). This dynamic view, in which the ligand induces conformational changes that result in reactive properties, might be thought of as challenging the claim that structure or Form determines Function. However, we claim that it is rather an enrichment of this view what is required. It is true that conformations need to change in order to perform their functions and so dynamics must be incorporated into the definition of molecular function: "A deep insight into the function of proteins will only be obtained through a combined study of both structural and motional properties of these inherently dynamic molecules" (Lewis 1998). But surface coupling is still the necessary condition for protein activity and so the role of Form is still fundamental in the comprehension of Function.

We want to remark that the fact that Form is fundamental for the understanding of Function does not imply that Function is reducible to Form (*as a whole*). This is specially evident if we consider the role played by domains in protein function. Functional domains are those fragments of a protein that bind to other macromolecules in biochemical reactions. Their structure and biochemical properties determine whether and how these reactions take place, whereas other parts of a protein may be substantially less important in relation to function. For this reason, two variants of a

[7] A peculiar but poorly followed morphological approach to functional relationships was the application of Catastrophe theory to ethology (Thom 1972; see Pérez Herranz 1994 for a philosophical examination)

protein (even within the same individual) may have identical functional domains, but can vary in the rest of the structure. Thus, a protein function cannot always be derived from its overall form, but, in any case, from the form of its parts (rather independently of the rest of the molecular form). This is, for instance, the case of topoisomerases, all of them in charge of the same functions, but made up of different sequences and having different overall structures :

> "[D]espite little or no sequence homology, both type IA and type IIA topoisomerases from prokaryotes and the type IIA enzymes from eukaryotes share structural folds that appear to reflect functional motifs within critical regions of the enzymes. [...] The structural themes common to all topoisomerases include hinged clamps that open and close to bind DNA, the presence of DNA binding cavities for temporary storage of DNA segments, and the coupling of protein conformational changes to DNA rotation or DNA movement." (Champoux 2001)

3.2.Topological constraints

The existence of structural constraints to Function has become a major topic in evolutionary biology debates above the role of Natural Selection (see Maynard-Smith et al. 1985 for a review; Alberch 1983). The 'constraints school' outlined that variation is not isotropic and thus free to be moulded by Natural Selection, but appears constrained or "shaped" by the developmentally available morphologies. The discipline of evolutionary developmental biology ('evodevo') studies the internal variational properties of developmental systems and the ways in which this defines possibilities for evolutionary change. The same point can be made in the molecular realm. A very illustrative case comes from the topological constraints given by the circular form of prokaryote's DNA chains. Linear DNA molecules can freely rotate to accommodate changes in the number of twists. But if (as illustrated in **Figure 6**) the two ends are linked to form a covalently closed, circular DNA (cccDNA), the absolute number of times the chains can twist about each other does not change. Such a cccDNA is said to be *topologically constrained*, in the sense that its form severely limits its functional properties. As a consequence, "understanding the topology of DNA and how the cell both accommodates and exploits topological constraints during DNA replication, transcription, and other chromosomal transactions is of fundamental importance in molecular biology." (Watson et al. 2004: 112, 139)

A very interesting consequence of topological constraints lies in the role they may play in the prediction of other molecular entities which are able to overcome these constraints in order to perform certain tasks. Let us keep on with the example of cccDNA in order to illustrate this issue.

The two strands of the double helix must rapidly separate in order for DNA to be duplicated. But, as they are twisted around each other, this cannot occur without permanently breaking a covalent bond in the sugar phosphate backbones. The easiest topological (and less energetically expensive) way of separating the two circular strands without breaking any bond, consists of cutting one of them and passing it through the other repeatedly. This is the function of the afore-mentioned topoisomerases, which are precisely "dedicated to solve the topological problems associated with DNA replication, transcription, recombination, and chromatin remodelling by introducing temporary single- or double-strand breaks in the DNA." (Champoux 2001). So if we assume as a general rule in biology that minimal energy is used to fulfil a function and this corresponds to the easiest topological way of doing so, we can say that Form plays a predictive role in the postulation of molecular entities in charge of fulfilling certain functions.

4. Conclusions

Throughout this paper we have seen how despite molecular biology is claimed to be reductionist by nature, it faces the same questions posed by naturalists interested in macroscopic organic forms. Thus, our analysis of biomolecular forms allows us to restate traditional ontological questions about Form:

1. *Living forms are dynamical entities*, a fundamental property for both the fulfilling of functions and the maintenance of form.
2. The relationship between protein wholes and their constituent parts, as studied on current attempts to solve the PFP, suggests that *biological macromolecules are holistic entities* not reducible to their 'parts'.
3. Both comparative analysis and folding studies demonstrate that *Form is not reducible to sequence.*
4. *The* comprehension of *Form cannot be dissociated from the understanding of the process governing its generation.*
5. Folding is not explainable by just appealing to the nature and position of amino acids in the polypeptide chain, or the inter-atomic forces acting on the sequence. *The generation of molecular form depends on the structural landscape that is generated during the very process of folding.*
6. *Form is a fundamental category in the understanding of Function,* because of both its intrinsic functional properties and the constrains it imposes upon the fulfilling of certain biological tasks.

All these facts demand:

1. A geometrical-topological treatment of molecular substances in contrast with the informational approach that has dominated mainstream philosophy of biology.
2. A more pluralist conception of causality that roots into Aristotelian philosophy and its conception of causality as the set of principles that are necessary and sufficient to explain a phenomenon. In this sense Matter, Form and Function appear as irreducible explanatory dimensions to understand biological phenomena.
3. Within the wider context of the organism, and against the consideration of genes as the elementary units (and of proteins as mere accidents of genetic essences) proteins (including their morphological and dynamical properties) should be favoured as the minimal compositional units of living organization.

However, we want to make clear that the fact that many traditional problems of biology regarding the nature and the role of Form reappear at the molecular level does not mean that they are going to be solved at this level. Rather it means that molecular reductionism will have to look, on the first place, at the many problems that Form poses in the very field of molecular biology.

5. Acknowledgements

Laura Nuño de la Rosa wants to thank Lino Camprubí and very specially Xabier Barandiaran for their very helpful comments and suggestions on earlier drafts of this paper.

6. References

Adleman L.M. 1994. Molecular Computation of Solutions to Combinatorial Problems. *Science* 266: 1021-1024.

Abouheif, E. 1997. Developmental genetics and homology: A hierarchical approach. *Trends in Ecology and Evolution* 12:405-408.

Alberch, P. 1980. Ontogenesis and morphological diversification. *Amer. Zoologist* 20: 653-667.

Alberch, P. 1982. Developmental constraints in evolutionary processes. In: J. T. Bonner (ed.) *Evolution and Development*. Dahlen Konferenzen. New York: Springer-Verlag.

Arnold, S. J. 1983. Morphology, performance, fitness. *Amer. Zoologist* 23: 347-361.

Baldwin R. L. 1995. The nature of protein folding pathways: the classical versus the new view. *J. Biomol.* NMR 5: 103–109.

Bechtel, W. and Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research.* Princeton: Princeton University Press.

Budgaonkar, J.B.'s web page: http://www.ncbs.res.in/jayant/jayant.htm

Callebaut, W. & Rasskin-Gutman, D. (eds) 2005. *Modularity: Understanding the Development and Evolution of Natural Complex Systems*. Cambridge, MA: MIT Press.

Campbell, I. D. 2002. The march of structural biology. *Nature Reviews. Molecular Cell Biology 3.*

Crick, F. H. C. 1958. On protein synthesis. *Symp. Soc. Exp. Biol.* 12.

Champoux, J. J. 2001. DNA topoisomerases: Structure, Function, and Mechanism. *Annu. Rev. Biochem.* 70.

Denton, M. J., Marshall C. J. & Leggew M. 2002. The Protein Folds as Platonic Forms: New Support for the Pre-Darwinian Conception of Evolution by Natural Law. *J. theor. Biol.* 219.

Dill, K. A., Ozkan, B. Weik T.R., Chodera, J. D & Voelz, V.A. 2007. The protein folding problem: when will it be solved? *Current Opinion in Structural Biology* 17.

Depew, D.J. & Weber, B.H. 1995. *Darwinism Evolving: Systems and the Genealogy of Natural Selection*: MIT Press.

Dobson & Hore. 1996. Kinetic studies of protein folding using NMR spectroscopy. *Nature Structural Biology* 173.

Eble, G.J. 2005. Morphological Modularity and Macroevolution: Conceptual and Empirical Aspects. In Callebaut, W. & Rasskin-Gutman. 2005.

Editorial: So much more to know. *Science* 2005, 309: 78-102.

Fontana, W. & Schuster, P. 1998. Shaping space: The possible and the attainable in RNA genotype-phenotype mapping. *J. Theor. Biol.* 194: 491-515.

Garret, R. H. & Grisham, C. M. . 1999. *Biochemistry* 2nd ed: Saunders College Publishing: Harcourt Brace College.

Gilbert, S. F., Opitz, J. M. & Raff, R. Review. Resynthesizing Evolutionary and Developmental Biology. *Developmental Biology* 173: 357–372

Gen Bank: http://www.ncbi.nlm.nih.gov/Genbank/

Gould, S. J. 1977. *Ontogeny and Phylogeny*. Cambridge MA: Harvard Univ. Press.

Goodwin, B. 1982. Development and Evolution. *J. Theor. Biol.* 97: 43-55

Goodwin, B. & Saunders, P. 1989. *Theoretical Biology: Epigenetic and Evolutionary Order for Complex Systems*: Edinburgh University Press.

Griffiths, P. E. 2002. The philosophy of molecular and developmental biology. In *Blackwell Guide to Philosophy of Science*: Blackwell Publishers.

Hall, B.K. 1999. *Evolutionary Developmental Biology*. 2nd edition. Chapman and Hall, London/Kluwer Academic Publishers, Netherlands.

Hamburger, V. 1980. Embryology and the modern synthesis in evolutionary theory. Evolutionary theory in Germany, A Comment. In Mayr, E. and Provine, W.B. *The Evolutionary Synthesis. Perspectives on the Unification of Biology* (de.) Harvard Univ. Press.

Hou J., Sims G. E., Zhang C., Kim S. H. (2003). A global representation of the protein fold space. *Proc. Natl. Acad. Sci.* U.S.A. 100.

Hüttemann & Love, A. Unpublished. Forms of Reductive Explanation in Biological Science: Intrinsicality, Temporality, and Part-Whole Relations.

Ishima, R. & Torchia, D.A. 2000. Protein dynamics from NMR. *Nature Structural Biology* 7(9)

Jacob, F. 1977. Evolution and tinkering. *Science* 196: 1161-1166.

Kaltashov, I.A. & Eyles, S. J. 2005. General overview of basics concepts in molecular biophysics. *Mass Spectrometry in Biophysics: Conformation and Dynamics of Biomolecules*, John Wiley & Sons, Inc.

Kay, Lewis E. 1998. Protein dynamics from NMR. *Nature structural biology. NMR supplement.*

Laubichler, M. D. & Wagner, G. P. (2000) Organism and Character Decomposition: Steps towards an Integrative Theory of Biology. *Philosophy of Science*, Vol. 67, Supplement. Proceedings of the 1998 Biennial Meetings of the Philosophy of Science Association. Part II: Symposia Papers.

Levinthal C. 1968. Are there pathways for protein folding? *J. Chim. Phys. 65*

Lodish et al. 2004. *Molecular Cell Biology*, 5th, W.H. Freeman and Company: New York.

Maynard Smith, J., Burian, R., Kauffman, S., Alberch, P., Campbell, J., Goodwin, B., Lande, R., Raup, D. & Wolpert, L. 1985. Developmental constraints and evolution. *Quarterly Rev. Biol.* 60.

Mayr, E. 1942. *Systematics and the Origin of Species*. Columbia University Press. New York.

Nayeem A., Sitkoff D., Krystek S Jr. 2006. A comparative study of available software for high-accuracy homology modeling: From sequence alignments to structural models. *Protein Sci* **15**: 808-824.

Olby, R. 2003. Quiet debut for the double helix, *Nature* 421 (23)

Pérez Herranz, F.M. 1994. "La Teoría de las Catástrofes de René Thom, nuevo contexto determinante para las ciencias morfológicas", El Basilisco, nº 16, págs. 22-42.

Protein Data Bank: http://www.rcsb.org/pdb/

Ramachandran, G.N., Ramakrishnan, C. & Sasisekharan, V. 1963. *Stereochemistry of polypeptide chain configurations.* In: *J. Mol. Biol.* vol. 7, p. 95-99.

Spemann, H. 1938. Embryonic Development and Induction.Yale Univ. Press: New Haven.

Rosenberg, A. 1997. Reductionism redux: Computing the embryo. *Biology and Philosophy* **12**: 445—470

Schuster, P. 2001. Evolution in silico and in vitro: The RNA model. *Biol. Chem.* 382.

Stadler, B., Stadler, P., Wagner, G., Fontana, W. (2001) The Topology of the Possible: Formal Spaces Underlying Patterns of Evolutionary Change. *J. theor. Biol.* 213.

Thom, R. 1972. *Stabilité structurelle et morphogénèse*, Intereditions, Paris.

Wand. 2001. Dynamic activation of protein function: A view emerging from NMR spectroscopy. *Nature structural biology.*

Waddington, C. H. 1962. *New patterns in genetics and development*. Columbia University Press, New York.

Waddington, C.H. et al. 1968. *Towards a Theoretical Biology*, International Union of Biological Sciences & Edinburgh University Press.

Watson, J. D. Baker, T.A., Stephen P. Bell, Alexander Gann, Michael Levine & Losick, R. 2004. *Molecular Biology of the Gene*, 5th Ed.: Benjamin Cummings.

Watson, J. D. 1968. The Double Helix: A Personal Account of the Discovery of the Structure of DNA: Atheneum.

Zhang C. & DeLisi C. (2001). Protein folds: molecular systematics in three dimensions. Cell. Mol. Life Sci. 58: 72–79.

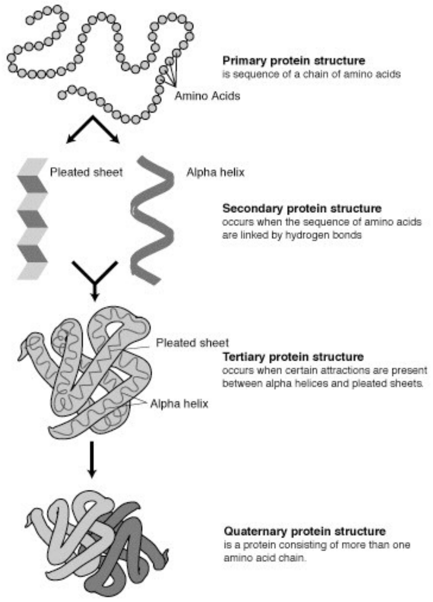
Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one amino acid chain.

Figure 1: Protein structure.
Courtesy: National Human Genome Research Institute.

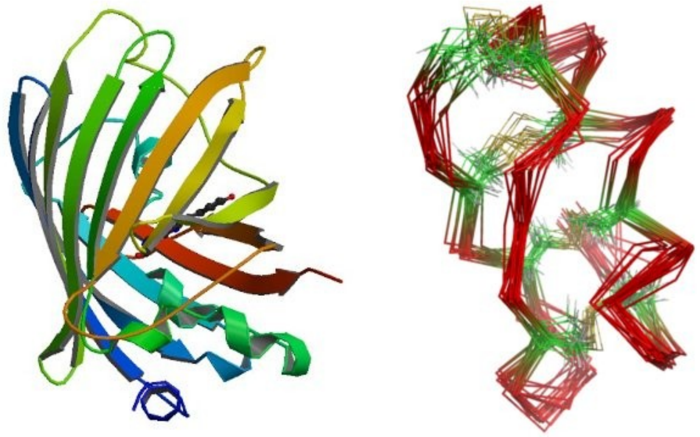

Figure 2: a. Crystal structure of the Green Fluorescent Protein variant YFP-H148Q. **b.** Protein NMR structure of the four-disulfide-bridge scorpion toxin HsTx1. Multiple structures are shown to reflect the natural fluctuations in native-state protein structure in solution. (*Wikimedia commons.* GNU Free Documentation License.)

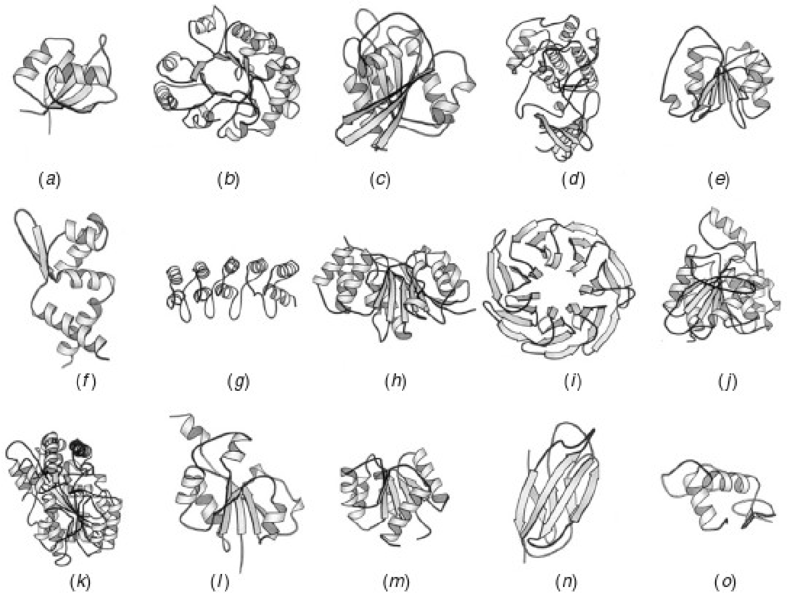

Figure 3: Protein motifs (Kaltashov & Eyles 2005)

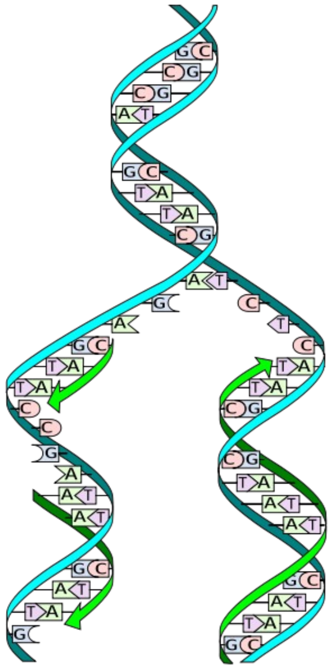

Figure 4: DNA replication. The double helix is unwound and each strand acts as a template. Bases are matched to synthesize the new partner strands. (*Wikimedia commons.* GNU Free Documentation License.)

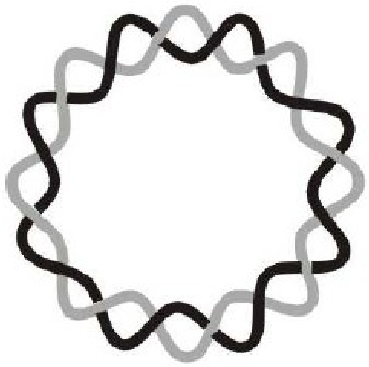

Figure 5: Schematic representation of a cccDNA. The linking number represents the number of times that each curve winds around the other, and it is an invariant topological property of cccDNA, regardless the shape of the DNA molecule.