Title: Network properties of genotype-phenotype mappings

AUTHOR: Rebeca de la Fuente Marañón

Master's Thesis

Master's degree in Physics of Complex Systems

at the
UNIVERSITAT DE LES ILLES BALEARS

Academic year 2015-2016

Date September 5, 2016

UIB Master's Thesis Supervisor Emilio Hernández García
Network properties of genotype-phenotype mappings

Rebeca de la Fuente Marañón

September 25, 2016
1 Introduction

One fundamental question in biology is which are the processes that give rise to a differentiation between individuals, species and groups. Since the discovery of genes as the units of heredity, there has been a tendency to view genes as the determinants of the physical characteristics. It was W. Johannsen [1] who first proposed the distinction between genotype and phenotype by studying the heritable variation in plants. The hereditary dispositions of organisms was differentiated from the ways in which those dispositions have physical manifestation. The genotype-phenotype distinction was a conceptual framework from which studies of development, genetics and evolution achieved a new perspective, being the next step the understanding of how is the mapping between these two differentiate spaces.

The first section of this work contains a literature review of the main concepts that have been developed in time with the study of the genome and its expression and from which the genotype-phenotype mapping is considered nowadays as a crucial framework which picks up the complex mechanism by which individuals are developed [1-14]. It is also shown in the last part of this section actual genotype-phenotype mappings which resulting properties are going to be analyzed and compared in the third section [18,19][21-23]. The second section shows a literature review of the concept of neutral evolution, which explains the invariance of phenotype behind multiple point mutations, and how from this mechanism which is observed in actual organisms along evolution is possible the punctuated equilibria phenomenon. Behind this concept of evolution scientific community is constructing neutral networks, which are defined from the genotype-phenotype mappings, and studying some properties of these networks having a biological interpretation [15-17]. This section also shows different constructions of neutral networks arising from genotype-phenotype mappings defined in the first section. Note that most of the models analyzed are based on RNA folding. The third section shows network properties of genotype-phenotype mappings of the different models which are defined in the lastest sections and are compared to each other. The fourth section explains a new model based on a cellular automata. This model is thought to be appropriate because of its originality and mathematical nature. Some measures that have been used in previous models are applied over the new model in order to compare the resulting biological properties. For finalizing the work, a new measure is developed and applied to the model to obtain interesting biological interpretations.

The main goal of this work is to compare different models of neutral networks to a new one based on a cellular automaton. From this comparison, we can elucidate that from a very simple and abstract model it is possible to obtain similar results to very much complicated models based on RNA folding or protein function. As a second goal, biological properties such as robustness, evolvability and accesibility are studied to get an intuitive idea about how evolution operates behind neutral mutations.
1.1 Genotype and phenotype differentiation

Life is possible because of the survival of genotypes with time, by which the genetic information passes between successive generations. It has the reproductive property which allows the persistence of life during million of years. The purpose of this work is not to study the progressive character of evolution, but rather to study the main characteristics of evolution arising from the genotype-phenotype relation. The way life is generated and its endless effect comes from genotype characteristics but the genetic information is not strictly sufficient for life because there are external factors that are coupled to the process that give structure to life. This is showed at the level of phenotype, where observable traits of life emerge. The outcome of development is terminated at the death of the organism, while the genome is an element which pathway goes from the first stage in the life to the “final individual”; there is no reciprocal effect of the development of an organism passing through generations. So, the genotype is defined as the inherited material transmitted between generations, whereas the observable traits of an individual was referred as the phenotype, which involves physical attributes as morphology, development and behaviour. So it is on the phenotype space where it is possible to apply epigenetic factors. There could be no inheritance of acquired characteristics.

This thought comes from a rediscovery of Mendel’s work (1822-1884) on inheritance, which study shows a clear distinction between the genome and the phenome and its ambiguity relation. He observed that plants that carry one specific member of a gene pair specifying red flowers and another one specific member specifying white flowers are indistinguishable from plants carrying two copies of both. It is common that a large fraction of genetic variation present in a population is hidden at the level of phenotype and we need experimental techniques to explore it. Although Mendel’s laws have some exceptions, its work immediately generated an apparent conflict with the theory of gradual evolution of Darwin. It was the clarification of the genotype and phenotype concepts which let the understanding of heredity and its separation from the developmental process where evolutionary forces as natural selection act. This distinction is essential to the developmental of an evolutionary theory. Following this reasoning, heritable variation is generated at the genotypic level, Mendelian laws specify how this variability is transmitted to the next generation while selection operates at the level of phenotype. From the evolutionary genetics point of view, development is the function that maps the genotype onto the phenotype. At higher levels of interaction, such as morphological traits, the genotype-phenotype relation is more complex and non-linear.

The development of molecular biology began with the definitive identification of DNA as the material basis of genes in 1940. Followed by this identification, there began the study of the chemical and physical structure of DNA, and the molecular mechanism by which genome produces copies of itself and molecules of physiological and developmental functions. All these studies give rise to the confirmation of the causal independence of the hereditary behaviour of the genome from its developmental functions.

DNA is known to be a long polymer composed of only four types of subunits, which resemble one another chemically. DNA is composed of two strands of the
polymer wound into a helix. The observation that DNA is double-stranded is of crucial significance and provide one of the major clues that led to the Watson-Crick structure of DNA. Only when this model was proposed did the DNA’s potential for replication and information encoding become apparent. Each of these chains is known as a DNA chain, or a DNA strand. Hydrogen bonds between the base portions of the nucleotides hold the two chains together. Nucleotides are composed of a five-carbon sugar to which are attached one or more phosphate groups and a nitrogen-containing base. In the case of the nucleotides in DNA, the sugar is deoxyribose attached to a single phosphate group (hence the name deoxyribonucleic acid), and the base may be either adenine (A), cytosine (C), guanine (G), or thymine (T). The nucleotides are covalently linked together in a chain through the sugars and phosphates, which thus form a backbone of alternating sugar-phosphate-sugar-phosphate. Because only the base differs in each of the four types of subunits, each polynucleotide chain in DNA is analogous to a necklace (the backbone) strung with four types of beads (the four bases A, C, G, and T). These same symbols (A, C, G, and T) are also commonly used to denote the four different nucleotides; that is, the bases with their attached sugar and phosphate groups.

![Building blocks of DNA](image1)

The way in which the nucleotide subunits are lined together gives a DNA strand a chemical polarity. If we think of each sugar as a block with a protruding knob (the 5’ phosphate) on one side and a hole (the 3’ hydroxyl) on the other, each completed chain, formed by interlocking knobs with holes, will have all of its subunits lined up in the same orientation. Moreover, the two ends of the chain will be easily distinguishable, as one has a hole (the 3’ hydroxyl) and the other a knob (the 5’ phosphate) at its terminus. This polarity in a DNA chain is indicated by referring to one end as the 3’ end and the other as the 5’ end. The three-dimensional structure of DNA –the double helix– arises from the chemical and structural features of its two polynucleotide chains. Because
these two chains are held together by hydrogen bonding between the bases on the different strands, all the bases are on the inside of the double helix, and the sugar-phosphate backbones are on the outside. In each case, a bulkier two-ring base (a purine) is paired with a single-ring base (a pyrimidine); A always pairs with T, and G with C. This complementary base-pairing enables the base pairs to be packed in the energetically most favorable arrangement in the interior of the double helix. In this arrangement, each base pair is of similar width, thus holding the sugar-phosphate backbones an equal distance apart along the DNA molecule. To maximize the efficiency of base-pair packing, the two sugar-phosphate backbones wind around each other to form a double helix, with one complete turn every ten base pairs.

Genes carry biological information that must be copied accurately for transmission to the next generation each time a cell divides to form two daughter cells. Two central biological questions arise from these requirements: how can the information for specifying an organism be carried in chemical form, and how is it accurately copied. The discovery of the structure of the DNA double helix was a landmark in twentieth-century biology because it immediately suggested answers to both questions, thereby resolving at the molecular level the problem of heredity.

DNA encodes information through the order, or sequence, of the nucleotides along each strand. Each base -A, C, T, or G- can be considered as a letter in a four-letter alphabet that spells out biological messages in the chemical structure of the DNA. Organisms differ from one another because their respective DNA molecules have different nucleotide sequences and, consequently, carry different biological messages: genes contain the instructions for producing proteins. The DNA messages must therefore somehow encode proteins. This relationship immediately makes the problem easier to understand, because of the chemical character of proteins. The properties of a protein, which are responsible for its biological function, are determined by its three-dimensional structure, and its structure is determined in turn by the linear sequence of the amino acids of which it is composed. The linear sequence of nucleotides in a gene must therefore somehow spell out the linear sequence of amino acids in a protein. The exact correspondence between the four-letter nucleotide alphabet of DNA and the twenty-letter amino acid alphabet of proteins -the genetic code- is not obvious from the DNA structure. Gene expression is the mechanism through which a cell translates the nucleotide sequence of a gene into the amino acid sequence of a protein.
The complete set of information in an organism’s DNA is called its genome, and it carries the information for all the proteins the organism will ever synthesize. The amount of information contained in genomes is staggering: for example, a typical human cell contains 2 meters of DNA. At each cell division, the cell must copy its genome to pass it to both daughter cells. The discovery of the structure of DNA also revealed the principle that makes this copying possible: because each strand of DNA contains a sequence of nucleotides that is exactly complementary to the nucleotide sequence of its partner strand, each strand can act as a template, or mold, for the synthesis of a new complementary strand. Each separated strand then serves as a template for the production of a new complementary partner strand that is identical to its former partner. The ability of each strand of a DNA molecule to act as a template for producing a complementary strand enables a cell to copy, or replicate, its genes before passing them on to its progeny.

Transcription and translation are the means by which cells read out, or express, the genetic instructions in their genes. Because many identical RNA copies can be made from the same gene, and each RNA molecule can direct the synthesis of many identical protein molecules, cells can synthesize a large amount of protein rapidly when necessary. But each gene can also be transcribed and translated with a different efficiency, allowing the cell to make vast quantities of some proteins and tiny quantities of others. Moreover, a cell can regulate the expression of each of its genes according to the needs of the moment, most obviously by controlling the production of its RNA. The first step a cell takes in reading out a needed part of its genetic instructions is to copy a particular portion of its DNA nucleotide sequence, a gene, into an RNA nucleotide sequence. The information in RNA, although copied into another chemical form, is still written in essentially the same language as it is in DNA: the language of a nucleotide sequence. RNA differs from DNA chemically in two respects: The nucleotides in RNA are ribonucleotides, that is, they contain the sugar ribose (hence the name ribonucleic acid) rather than deoxyribose and although, like DNA, RNA contains the bases adenine (A), guanine (G), and cytosine (C), it contains the base uracil (U) instead of the thymine (T) in DNA. Since U, like T, can base-pair by hydrogen-bonding with A, the complementary base-pairing properties described for DNA apply also to RNA. So, in RNA, G pairs with C, and A pairs with U. It is not uncommon, however, to find other types of base pairs in RNA: for example, G pairing with U occasionally.
Despite these small chemical differences, DNA and RNA differ quite dramatically in overall structure. Whereas DNA always occurs in cells as a double-stranded helix, RNA is single-stranded. RNA chains therefore fold up into a variety of shapes, just as a polypeptide chain folds up to form the final shape of a protein. The ability to fold into complex three-dimensional shapes allows some RNA molecules to have structural and catalytic functions.

All of the RNA in a cell is made by DNA transcription, a process that has certain similarities to the process of DNA replication. Transcription begins with the opening and unwinding of a small portion of the DNA double helix to expose the bases on each DNA strand. One of the two strands of the DNA double helix then acts as a template for the synthesis of an RNA molecule. As in DNA replication, the nucleotide sequence of the RNA chain is determined by the complementary base-pairing between incoming nucleotides and the DNA template. When a good match is made, the incoming ribonucleotide is covalently linked to the growing RNA chain in an enzymatically catalyzed reaction. The RNA chain produced by transcription is therefore elongated one nucleotide at a time, and it has a nucleotide sequence that is exactly complementary to the strand of DNA used as the template. Most genes in a cell produce mRNA molecules that serve as intermediaries on the pathway to proteins. Once an mRNA has been produced, by transcription and processing the information present in its nucleotide sequence is used to synthesize a protein. Transcription is simple to understand as a means of information transfer: since DNA and RNA are chemically and structurally similar, the DNA can act as a direct template for the synthesis of RNA by complementary base-pairing. In contrast, the conversion of the information in RNA into protein represents a translation of the information into another language that uses quite different symbols. Moreover, since there are only four different nucleotides in mRNA and twenty different types of amino acids in a protein, this translation cannot be accounted for by a direct one-to-one correspondence between a nucleotide in RNA and an amino acid in protein. The nucleotide sequence of a gene, through the medium of mRNA, is translated into the amino acid sequence of a protein by rules that are known as the genetic code. The sequence of nucleotides in the mRNA molecule is read consecutively in groups of three. RNA is a linear polymer of four different nucleotides, so there are 64 possible combinations of three nucleotides: the triplets AAA, AUA, AUG, and so on. However, only 20 different amino acids are commonly found in proteins. Either some nucleotide triplets are never used, or the code is redundant and some amino acids are specified by more than one triplet. Each group of three consecutive nucleotides in RNA is called a codon, and each codon specifies either one amino acid or a stop to the translation process.

Although a few slight differences in the code have been found, these are
chiefly in the DNA of mitochondria. Mitochondria have their own transcription and protein synthesis systems that operate quite independently from those of the rest of the cell, and it is understandable that their small genomes have been able to accommodate minor changes to the code.

1.2 Genotype-Phenotype relationship

The reading of the genotypic information by the cells and the use of this information to produce molecules that underlie development of the characteristics of the phenotype is carried out by the processes of transcription and translation and it follows a different pathway than the process by which heredity is possible. It is the transcription of the DNA into a separate molecule of RNA the critical point in the separation of the heredity and the developmental functions of the genome.

As we already know, a protein consists of a string of amino acids, each one of which is coded by a triplet of nucleic acids in the string of DNA constituting a gene. A replacement of any of one of these amino acids will prevent the physiological activity of the protein. It is almost impossible from observing phenotypic variation be sure about the exact variation in genotype space. In humans there are three million nucleotide differences on the average between any two people taken at random. No two organisms have identical genomes, with the exception of twins or individual clones. Even in the cases of organisms with identical genetic information, there exists a phenotypic variation. If the mechanism of development were such that changes in genotype result in changes in phenotype and different phenotypes was the consequence of a difference in genotype, then the study of the genotype-phenotype mapping would be very simplified. This is not the case: there is many-to-many relation in which any given genotype corresponds to many different phenotypes and there are different genotypes that undergo the a given genotype. This many-to-many relationship between genotype and phenotype comes from the relations:

- The relation between the DNA and the chemical structure of proteins.
- Relations between transcription products and translation.
- The dependence of development on the genotype and the temporal sequence of environments in which the organism develops and functions.
- Stochastic variations of molecular processes within cells.

Nowadays, scientists have no doubt that multiple genes interact with multiple environmental variables such as abiotic factors, culture or symbiots to produce phenotype. The genotype-phenotype relationship is a useful framework in the context of pleiotropy, epistasis, and environmental effects. Pleiotropy occurs when one gene influences two or more apparently unrelated phenotypic traits and the term epistasis describes a certain relationship between genes, where an allele of one gene hides or masks the visible phenotype of another gene.
For point mutations there are about $10^{-8}$ mutations per base and generation (20-25 years) in the core of the human genome. This implies about 64 mutations per generation [26]. Also, each cell express different parts of genome from each other. In asexual unicellular organisms as bacteria that are reproduced by division of parental cell differ in their genomes because mutations of DNA are sufficiently common that at least one of the nucleotides that constitute their DNA will have undergo a spontaneous change during cell division. Mutations isolated from laboratory strains have been instrumental to the understanding of the genotype-phenotype mapping. Under the classical scheme, a mutation is compared to a wild-type reference, and its phenotypic effects are used are used to infer gene function. In some cases a genetic change causing a variation in phenotype, it is convenient to assimilate the corresponding gene as a causal determinant of a trait [6]: It is common to find headlines which express the discovery of the 'longevity' or 'well-being' gene. What this actually means is that a variation at a given gene causes variation in a given phenotype. Actually, a gene alone cannot be associated to a unique phenotype. Genes need a cellular environment, the combined action of multiple other genes and physico-chemical conditions to have an observable effect on organisms. A mutation that causes a phenotypic outcome is required to allow genetic evolution and adaptation by natural selection [7]. This is because a new allele formed in genome would generate a different phenotype outcome each time it ends up in a different organism. Competition occurs between alleles that span the same genetic locus. There is a natural selection acting directly on the genotype variation with is associated with a given phenotypic variation.

The accumulating data on the mutations that give rise to phenotypic variation share a common characteristic: The comparative developmental biology reveals that animals share a common set of key regulatory genes with conserved functions [8]. Differences between plants and animals for which the genetic basis have been partially identified are due to mutations at homologous genes and very few are due to new genes [3] and multiple cases of similar phenotypic changes have been shown to involve mutations of the same homologous genes in independent lineages even across large phylogenetic distances. A phenotype is said to be canalized if mutations do not affect considerably the physical properties of the organism. A canalized phenotype therefore may be result from a large variety of different genotypes and it is difficult to be sure about what is the exact genotype which has generated the specific phenotype. At the other hand, if there is no canalization, small changes in genome have some effect on the phenotypical traits. Here it is show a many-to-one relation, where many phenotypic features show no variation and one could think that there is no genetic variation for the specific trait and also that its development is resistant to environmental disturbances. One example of this occurs with flies where the development of the fly is disturbed such that some flies present two or fewer ocelli [11]. The offspring of those with two or fewer ocelli are abnormal flies than the parental generation. When the process of selective breeding from abnormal flies is continued over many generations a line of flies is produced that consistently have two ocelli, even in the absence of any external disturbance of development. This proves that different genotypes map onto the same phenotype. Developmental effects from genetic variation are prevented by the system of buffering.
Phenotypic plasticity defines the degree at which phenotype is determined by its genotype, because environmental factors can have high influence on the phenotype. One example of high plasticity can be observed in larval newts [9], in which predator-induced plasticity is showed: when two species are exposed to nonlethal larvae, different phenotypes arise. Because environmental factors affect to the developmental, the genotype is always in a dependent cause of the current environmental factors. Because of this, two identical genotypes do not express the same phenotype, even considering the case in which there has not been any mutation in cell division. By this way, the complete genotype does not contain the total information for specify the final product which is the organism with its physical traits. The outcome of the developmental processes depends on the genotype and the temporal sequence of environments in which the organism develops. One experimental proof of this is the experimentation with clones of the plant *Archillea* [10], where individual immature plants were collected from nature and from each plant three clones were produced and each one was let to growth at different earth elevations. The result was that there is no correlation among the plants in their growth for the different environments. Specific elevations were good environment to a specie to grow and not good for another ones. If the phenotype of an organism of a given genotype is plotted against an environmental variable, the function that is produced is called the norm of reaction of the genotype. The common experience is that norms of reaction of different genotypes show irregular patterns across each other. So, it is not possible to predict with certainty the phenotype result of different genotypes exposed to a given set of environmental factors. The outcome of development of any genotype is a unique consequence of the interaction between genome and environment. For example, in many turtle species, a change in temperature during egg evolution is associated with the resulting male/female sex determination. In this case, sex chromosomes and temperature have the same phenotypic effect on turtles. Therefore, not all phenotypic changes can be attributed to genetic changes.

From the information given before, one could thing about one-to-many relation between genotype and phenotype. But there exists also epistasis in the mechanism of the production of functional proteins. A physiological effect usually needs the reading of many different genes being all necessary, so alterations in any one of the genes will disturb the effect. From here, it is not difficult to have an intuition of the complex interaction between genes to produce a phenotype. Then one can elucidate the many-to-many relationship between genotype and phenotype.

Symmetry is an important question with respect to a biology because in average there is fluctuating asymmetry in the pattern of the phenotype. For example, humans do not have the same fingerprints on their left and right hands being the genes on cells of both hands identical, under the assumption that there was not been mutation. There is a random variation rising to an asynchrony of cell division. A single bacterial cell will divide into two cells after an hour. Those two cells will then each divide about one hour later but not simultaneously. The offspring will divide again, but each few minutes earlier or later than others. Then, population of cells is growing in time with no synchronization of division. The same asynchrony of divisions occur at the stages in the division of cells from...
the fertilized egg in embryos. The source of these asymmetries and asynchronies is the very low number of copies of biologically important large molecules in each cell. Also, for a reaction to occur between molecules they must be in proximity and each molecule in the reaction must be in the right vibrational state of interaction. Vibrational state are fluctuating for each molecule ultimately as a consequence of quantum uncertainty [12]. As a consequence of the stochastic variation in number, spatial location and reactivity of each kind of molecule, there is considerable random fluctuation from cell to cell in the timing of cell division and in its outcome. These stochastic effects are important sources of phenotypic variation.

1.3 Actual GP mappings

It was Alberch in 1991 [4] who first introduced the concept of genotype-phenotype mapping. By his work he reminds biologists that genes do not specify development and organismal form, but are instead one of several causal factor to be the phenotype determined. Alberch worked with a pattern-formation model by which he introduced a metaphor: a defined mapping function by a given parameter space which defines the function. Parameters would be developmental in nature and their values would be affected by gene expression [4]. This mathematical model of pattern formation define patterns as the morphology resulting from the set of genetic and developmental interactions. Interactions have properties that emerge from the dynamics of the system and are not taking into account in the genotype space. Morphological diversity is generated by perturbations in parameter values, such as rates or diffusion, cell adhesion, and so on, or by perturbations in initial conditions. The system therefore will generate a discrete subset of phenotypes. In general, the relationship between the morphogenetic parameters and the phenotype can be mathematically stated as

\[
dP/dt = f(P, x_i),
\]

where \( P \) is a phenotype, \( f \) is an unspecified function describing the nature of interactions and \( x_i \) is a finite number of interacting morphogenetic parameters.
A hypothetical parameter space composed by six phenotypes is shown in Figure 1, which are represented by capital letters. These phenotypes are determined by the interaction of two parameters $x_1$ and $x_2$. Alberch derived four general conclusions for the conceptual mapping:

- The map is very complex, so there is a relation of many-to-many between genotype and phenotype. In particular, the same phenotype can be obtained from different combination of genetic informational sources.
- The area in parameter space is associated to the stability of a given phenotype. This stability refers to alterations of developmental parameters: environmental and genetic perturbations. The area of a domain is therefore related to canalization.
- The parameter space has boundary conditions such that small changes in the parameters could cause a transition from one phenotypic state to another one.
- The stability of a phenotype of a given population depends on the area of parameter space it occupies, and if it is sensitive to small perturbations due to the proximity to a transformational boundary. In figure 1 it is shown that species 1 exhibits phenotype D, a very stable phenotype whereas the distribution of parameter values for species 1 is very close to a basin boundary. Also, species 2 have a more stable position.

This work showed a new level of selection, where selection is done among pattern generating systems favoring the ones that exhibit the adequate balance between stability and potentially to generate sufficiently phenotypic variability. This is a framework from which we can view genotype-phenotype mapping as a complex interaction that needs mathematical tools.

As we shall see, once researchers were actually able to tackle real genotype-phenotype maps, intuitive properties of evolution in phenotypic space arising
from the Alberch’s work turn out to be fundamentally correct and now can be expressed on a more precise foundation. There are two main systems of computational and empirical studies: RNA folding and protein function.

1.3.1 RNA folding
As we already know, RNA consists of a string of the four types of nucleotides and forms a unique sequence, representing a genotype. The biochemical function of RNA is given by the three-dimensional structure. Therefore, the genotype-phenotype map of RNA is a map from sequence to structure and a map from structure to function. In particular, for short sequences the tertiary structure can be approximated by the secondary structure fold. Then, most of the genotype-phenotype models are based on RNA secondary structures representing the phenotypes. Since there are many sequences which fold into the same structure, there is a many-to-one relation.

There are sophisticated computer models capable to predicting the three-dimensional folding of a linear sequence of nucleotides based on thermodynamic considerations. For example, the Program RNAfold of the Vienna RNA package is commonly used by scientific community and it is the one used by the simulation models which are going to be studied [19,22]. By this program, it is possible to fold \textit{in silico} all the RNA sequences. Nucleotides in an RNA sequence tend to form pairs to minimize the free energy of the molecule and therefore it is used as structure the minimum free energy secondary structure predicted by routine fold of the RNAfold program. The RNA secondary structure folding is formed by base pairs connections between nucleotides of the same sequence. Parameters are the base pairing considerations and the temperature, which is usually set to 39°C. There are three conditions [24]:

- An individual nucleotide can only participate in one base pair.
- Base pairs between nearest neighbors are not considered.
- There are no pseudoknots

1.3.2 Protein function
Proteins fold in an almost rigid three-dimensional structure. This folding is induced by the interaction between the sequence of aminoacids forming the primary structure. One focus of neutrality happens at this level of the biological organization because most of the replacements of aminoacids in the sequence leaves the protein unchanged. But some aminoacids are crucial because if they are replaced by others the conformation of the protein changes. As the tertiary structure determines the protein function, it turns out that many aminoacid substitutions do not modify the structure, and thus have no effect on the phenotype. Also at this level, if one protein is replaced by another one, there is no major change in the metabolic pathway or regulatory circuit in which they participate. This is an example of robustness property belonging to biological systems. Also, the genotype space is characterize by large areas of neutrality.
that facilitate evolvability. In this mapping is found that many fewer protein molecules than RNA molecules fold, but they fold into many more structures than RNA. Then the genotype space for every phenotype is smaller and consequently protein molecules forming the same phenotype are more similar between them. RNA folding gives more novel structures than protein molecules and consequently it is more likely to the evolution of new structures in RNA folding than in protein molecules [25].

In this work we do not analyzed protein molecules for genotype-phenotype mappings models because we are most interested in RNA folding models. However, it is interesting the different approaches to get a genotype-phenotype mapping and the differences that one can get from different models.
2 Neutral evolution

The XXth century is characterized because of the efforts to elaborate a formal theory of evolution by using the processes of replication, mutation and selection. However, large empirical data show that there is no an optimal genotype which is searched by the evolutionary process.

Kimura surprised the scientific community in 1968 with the statement that most mutations are neutral in mammals [13]. The term neutral comes from this concept: most mutations in genome have no effect on their phenotype. In his study, Kimura shows that there are 16 million of substitutions in the whole genome every 28 million of years, so there is a substitution every 2 years and this mutational rate can only be applied to the actual evolutionary process if the majority of these mutations have no effect on their phenotype; in other words, it is only possible if most mutations are neutral.

Recent studies support this idea of neutrality at least at the molecular level. We already know that genes are involved in a complex regulatory network in which the proteins codified by some genes activate or inhibit the coding of other proteins, so one protein is in direct dependence of the action of other ones. The phenotype is thus the effect of the genome as a whole, rather than a linear combination of traits. We study here the properties such robustness, evolvability and accesibility arising from neutrality; in other words, if we study the effect of point mutations on the phenotype, it is accessible to analyze some properties that are intrinsic in the evolutionary process. There are in general two currents explaining the evolutionary process: Gradualism and Punctuated Equilibrium.

Darwin took from Geology the idea of gradualism, and it is one of the strong arguments of selection in his work *The Origin of species*. Gradualism defines evolutionary changes by slow accumulation of small changes caused by the very infrequent appearance of benefits mutations. This current is supported by selection behind the idea that most mutations are deleterious and consequently dissapear due to molecular responses and selection. Actually, there are many mutations that we are not able to discover because the organism act in such a way that eliminate the trait. Also, selection operates over the resulting set of phenotypes in a biological system (molecular, community, species, and so on), favoring the ones which fitness if higher than others. In this sense, there exists a gradual force of evolution, by which small changes in phenotype is induced by small changes in genotype. In this sense, a few number of mutations would cause a small variation in the phenotype.

But gradualism is one of the most controversial points of evolutionary theory because it conflicts with the fossil record, where species are observed to remain nearly unchanged for long periods of time and in some points of the history be quickly replaced by new species, which process is termed as punctuated equilibrium [15]. Now, the accumulation of neutral mutations seems to lead the existence of many similar individuals of the same species with genotypes that are very far apart from each other. In this situation, a new mutation may induce a big phenotypic change in one of these individuals but not in others because the effect seems to be over the genome as a whole in just one step because all the previous steps were in silence. This effect is a case for punctuated equilibrium.
2.1 Neutral networks

To study the effect of point mutations over the phenotype, it is useful to construct a neutral network, where nodes of a single neutral network are the genotypes having all the same phenotype. Then the number of phenotypes resulting from the mapping is the number of neutral networks we get. Now, two nodes in a given neutral network are connected if both are at a Hamming distance 1; in other words, if their sequences differ in only one digit. Neutral networks, one for each phenotype, could be connected or disconnected. In the later case subnetworks are formed. Therefore, each neutral network has a number of subnetworks and a subnetwork has nodes which are connected if their give rise to the same phenotype and are at a Hamming distance one. During neutral evolution, genes can randomly move through neutral networks and traverse regions of sequence space which may have consequences for robustness and evolvability. To obtain a neutral network connecting two genotypes yielding the same phenotype, there has been needed to define previously a genotype-phenotype mapping. Most of researches in this area consider a biological sequence of length L, and position \( i \) can be formed by one of \( k \) variants. One can think about this set of variants as the alphabet, depending of the type of sequence. It can be formed by \( k=4 \) bases of which DNA or RNA are made, or by \( k=20 \) aminoacids that build up proteins, or \( k=2 \) if we consider a binary string for a conceptual framework. Every realization of a sequence is a genotype and represents a point in genotype space. There are \( k^L \) different genotypes. Mutations can even modify the length, but we consider in this work only point mutations. Then for every neutral network, two nodes are connected if they are at a Hamming distance one, i.e, if they differ in only one nucleotide between the two sequences. Since genotypes are connected by only one-point mutation, mutations move the sequence from a node of the network to one of its neighbours which differ from it in just one position and have the same phenotype. Also, it is important to analyze the dependence of the genotype space on the distribution of phenotypes. Genotypes with the same phenotype are considered as genotypes with the same fitness in literature, so a sequence can move across any connected component of the network without cost of fitness. It is usual to find disconnected network, where subnetworks are sets of the genotype space in which movements by point mutations have no effect on their phenotype. So, regions where two different subnetworks are close, a point mutation may generate a genome that belongs to a different network, and consequently changes in phenotype arise. Here one can find pathways where punctuated equilibrium is possible, explaining the sudden changes in phenotypes observed after long periods of stasis [14]. The movement of a population on the neutral network have effects at the genomic level but does not cause any visible change. Once a population is trapped in a position of the neutral network close to the old phenotype, it will take a while until it diffuses again on a new subnetwork.

There are three basic characteristics that share RNA and protein molecules neutral networks [23]:

- Most neutral networks are not abundant; they contain few genotypes, whereas relatively few contain many genotypes.
• Large neutral networks are adjacent by point mutations to a greater diversity of phenotypes than small neutral networks.
• Large neutral networks span the entire sequence space.

2.2 Evolution

Evolution and adaption is understood as a process of search and fixation of fitter phenotypes, and it is conditioned by the connectivity and relationships between genotypes. There are a number of results that relate the topology of the networks with the equilibrium states of populations and the dynamics of adaptation on the neutral network. It has been shown that the distribution of a population evolving on a neutral network is solely determined by the topological properties of connectivity.

2.2.1 Robustness, Evolvability and Modularity

Phenotypic robustness and evolvability are defined properties of the neutral network. The term robustness is referred to the persistence of high level traits or fitness under variable conditions. It is a concept akin to the older ideas of homeostasis and canalization. In contrast, evolvability refers to the capacity for heritable and selectable phenotypic change. Robustness and evolvability are crucial to the persistence of life and their relationship is vital for understanding evolution. In nature, organisms are presented with a multitude of environments and are occasionally exposed to new and slightly different environments. Under these variable conditions, organisms must on the one hand maintain a range of functionalities in order to survive and reproduce. On the other hand, organisms must also be flexible enough to adapt to new conditions that they have not previously experienced. This dual presence of robustness and adaptiveness to change is observed at different scales in biology and it has been responsible for the persistence of life over billions of years. So, robustness and evolvability are related to canalization and adaptive phenotypic plasticity when organisms are exposed to environmental changes.

Wagner asserts that "understanding the relationship between robustness and evolvability is key to understand how living things can withstand mutations, while producing ample variation that leads to evolutionary innovations" [16]. At first, robustness and evolvability appear to be in conflict because intuition usually gives the following reasoning: maintaining developed functionalities is not compatible with the exploring new ones at the same time. But this conflict dissapear when robustness is conferred in both the genotype and phenotype space. In other words, if the phenotype is robustly maintained in the presence of genetic mutations, then a number of cryptic genetic changes may be possible and their accumulation over time might expose a large range of distinct phenotypes by movement across the neutral network. In this way, robustness of the phenotype might actually enhance access to heritable phenotypic variation and thereby improve long-term evolvability. So, if there is a high number of distinct phenotypes that are accessible through point-mutations between subnetworks then a wide variety of accessible phenotypes can be explored while remaining
close to a viable phenotype. It depends on the place you are placed in the network. Consequently robustness is quantified by number and size of neutral subnetworks that share the same phenotype. Because of this connectivity, some mutations or perturbations will leave the phenotype unchanged. On the other hand, evolvability is achieved over the long-term by movement across a neutral network.

Another important concept is modularity, which is defined as the degree of interconnectedness among components of a gene network. The solution is to be found is that genetic networks with intermediate degrees of connectedness strike a balance between too much robustness and too little evolvability, and vice versa. As natural selection can act on the degree of connectedness of a gene network, this means that it can alter both robustness and evolvability [17].

2.2.2 Network properties

In order to describe evolution in neutral networks we need to study the topological properties that are related to biological properties as robustness and evolvability. A property of the genotype space is its connectivity \( C(N) \), which is the proportion of connected genotypes. A configuration where the population has evolved mutational robustness is located in a region of the neutral network where the connectivity is as large as possible, and thus where mutations affect as less as possible the current phenotype. A high connectivity can be as an indicator of the presence of communities of phenotypes: common structures that arise because of the formed clusters that are densely connected in subsets [21]. If this value decreases and it approaches intermediate degrees of connectedness, then we can find a balance between too much robustness and too little evolvability and vice versa [19].

The degree of a sequence is a measure if its robustness to mutational changes. This measure corresponds to the number of neighbors \( k_i \) of a given sequence \( i \) within its neutral network. Then, the degree distribution \( p(k) \) gives the probability of finding a node of degree \( k \)

\[
p(k) = \frac{| \{ g \in S / \ g \ has \ degree \ k \} \} { |S| },
\]

for each subnetwork \( S \).

The average degree \( \langle k \rangle \) of a subnetwork is

\[
\langle k \rangle = \frac{\sum k_g}{|S|},
\]

These two local measures are related to the robustness to mutational changes. Then for larger values of \( k \) it is less likely that a random mutation causes a different phenotype. Degree is this a first indicator of the functional stability of a given genotype.

Another measure of robustness is the average shortest path between any pair of nodes. The distance between an arbitrarily chosen pair of sequences in
a subnetwork could be large, showing that it is possible a large accumulation of point mutations before a change in phenotype arise.

The shortest path, \( \langle d \rangle \) of each subnetwork is computed as the average of the shortest path length \( d_{ij} \) between all pair of nodes \( i \) and \( j \) belonging to the same subnetwork:

\[
\langle d \rangle = \frac{\sum_{i,j} d_{ij}}{|S||S-1|},
\]

There is a novel measure [22] which is called Accessibility and it is defined as

\[ A_i = \sum_j f_{ji}, \text{ where } f_{ij} = \frac{v_{ij}}{\sum_{k \neq i} v_{ik}} \]

\( v_{ij} \) is the number of point mutations to genotypes in the neutral network for phenotype \( i \) that create a genotype in the neutral network for phenotype \( j \), and \( \sum_{k \neq i} v_{ik} \) is the total number of non-neutral point mutations to genotypes in the neutral network for phenotype \( i \). Then, \( f_{ij} \) is the fraction of non-neutral point mutations to genotypes in the neutral network for phenotype \( i \) that create genotypes in the neutral network for phenotype \( j \).

Large values of the fraction \( f_{ij} \) indicate that phenotype \( j \) is relatively easy to find via random mutations from phenotype \( i \). then Accessibility is a measure of the overall accessibility of phenotype \( i \) from other phenotypes. Large values of \( A_i \) indicate that phenotype \( i \) is relatively accessible from throughout the space.

Another measure which is also developed in this article [22] is called Evolvability, which quantifies the potential for evolution away from phenotype \( i \):

\[ E_i = 1 - \sum_j f_{ij}^2 \]

This measure indicates that the diversity of other phenotypes that can be easily produced by mutations from a given phenotype, and this may indicate the potential for further adaption away from that phenotype. Then, if we select at random two non-neutral point mutations from a neutral network, it gives the probability that they will result in the same phenotype. So, this measure gives large values for phenotypes that are adjacent to many other phenotypes, and its non-neutral mutations are quite evenly divided among the adjacent phenotypes. Also, it is small for phenotypes that are likely to mutate to one or a very few number of alternate phenotypes.

A novel measure is here proposed to study how likely is that a non-neutral point mutation scapes from the new phenotype. I have called it Neighbour abundance and it is defined for every neutral network \( i \):

19
\[ B_i = \sum_g G_g / |N_i|, \text{ with } G_g = \sum_j \frac{k(j)}{|S_j|}, \]

where the first sum is done for every genotype belonging to the neutral network \( i \) with \( N_i \) the size of the neutral network and the last sum is done for every non-neutral point mutation \( j \) from genotype \( g \), \( k(j) \) is the degree of node \( j \) and \( |S_j| \) is the size of the neutral network \( j \) belongs to.

This measure is independent of the neutral network size and it gives the likelihood of a phenotype to not go back once a non-neutral point mutation is done. Large values of this measure means that once a point mutation is done and gives another phenotype, it is not likely that another mutation goes back to the previous phenotype, to the first neutral network. So it is a measure of the robustness of the positions in genotype space of the non-neutral point mutations.
3 Network Properties of RNA Genotype-Phenotype mappings

An interesting result of the distribution of phenotypes on RNA neutral networks is that there is a few abundant RNA shapes and a large number of rare ones. Evolution can explore most or all of the common structures by one-step mutations that preserve structure while moving the population on a neutral path, until it bumps into a novel phenotype. Most genotypes turn out to be located within a few mutational steps from most of the common phenotypes, making it predictable that such phenotypes will in fact be found by natural selection in a relatively short period of time. It is also found punctuated equilibria in simulations of RNA folding and it results from the fact that the population divides itself into smaller chunks, each of which explores a portion of the neutral network.

From different models of neutral networks we can extract some conclusions by comparing the results. In order to illustrate the network structure, we compare at first how RNA folding models yield to a rank distribution of phenotype abundance. In studies, RNA Neutral Networks are constructed such that all sequences that fold into the same structure form the neutral network of that structure. In the study of RNA Neutral Networks [19-22], the resulting structure is a large number of subnetworks in which each subnetwork is formed by genotypes that give rise to the same secondary structure and are linked by one point mutation. Then, we have, for every phenotype or secondary structure a number of subnetworks.

Different RNA sequences of length \( l=12 \) are folded \textit{in silico} by using the minimum free energy secondary structure predicted by the program RNAfold. In general measures are computed for sequence length \( l=12 \) in the article [22]. In [19] the \( 4^{12} \) molecules fold into 57 different secondary structures. In total there are 645 different subnetworks found for the 57 structures. Similar results are found for [19]. Note that in [22] measures are computed for different sequences lengths, which are denoted by k-mers.

In figures 2 and 3, it is shown by tables the structure of the different neutral networks formed by the RNA folding. For both tables we have a ranking succession of data which is ordered by phenotype abundances. Each neutral network is constructed such that is formed by all genotypes that fold into the same secondary structure. Then, every neutral network represents a phenotype and it is defined the phenotype abundance for every neutral network (or phenotype) as the number of genotypes that produce a particular phenotype; it is the size of the neutral network. Phenotype abundance is also termed as frequency in [19], which associated data structure is given in figure 3. Then, each rank index represents a neutral network versus its neutral network size. Abundance distribution is not dependent of the size of the genome length (n-mers). A similar distribution is showed for figures 2 and 4: there are a few highly abundant phenotypes and many rare ones. Also, the insets show the relation between such subnetwork sizes and the size of the network they belong to depending on the number of base pairs in the structure.
Figure 2: **Phenotype abundance distributions** [22]. The table which is at the left of the figure shows the distribution of abundances for the 12-mer (for sequences of length 12). There is a rank ordering of such abundances together with some measures: Nc is the number of subnetworks that are formed by the neutral networks, Dhamm is the maximum Hamming distance between a pair of genotypes in its neutral neutrok and Dspl is the maximum shortest path length between a pair of genotypes. The graph shows the phenotype abundances (y-axis) for each phenotype, ranked in order of abundance (x-axis). The most common phenotype is rank 1, the second most common is rank 2, and so on. Colors are representative for the corresponding sequence length computed in the model. The case of rank 1 is determined in a null distribution (nd) of abundances from random sequences. To generate the null distributions, each sequence is randomized in a family 500 times (preserving nucleotide composition).
Figure 3: Ranking of phenotype frequency (abundances)\cite{19}. In a rank ordering phenotype frequencies are shown for each neutral network together with the number of subnetworks that are formed by its structure.
Figure 4: **Distribution of subnetwork sizes** [19] In linear-logarithmic scale, it is shown the ranking distribution of subnetwork sizes. Colors indicate the number of base pairs considered in the secondary structure: one pair (black), two pairs (red), three pairs (green) and four pairs (blue). The solid line corresponds to an exponential fitting. Insets show for each group of structures the size of the subnetworks (in the y-axis) that belong to the same neutral network as a function of the corresponding neutral network size (in the x-axis). Note changes of scale in both axes.

Degree distribution and average degree are measured in order to study robustness to mutational changes. For large values of the degree $k$ it is less likely that a random mutation causes a different secondary structure. In figure 5, it is plotted the degree distribution $p(k)$ of fifteen subnetworks of different sizes [19]. These distributions cannot be approximated by a Poisson or binomial distribution. They are single-peak and it is shown that high-degree nodes are more frequent because the peak is presented close to the maximum degree value approached by each subnetwork. However, the maximum degree value possible is never reached.
Figure 5: **Degree distribution** $p(k)$ [19]. Degree distribution $p(k)$ of fifteen subnetworks. They are the five largest $N \sim 5 \times 10^4$ (black curves), five of intermediate size $N \sim 5 \times 10^3$ (brown curves) and five small subnetworks $N \sim 5 \times 10^2$ (blue curves) smaller. Here $N$ denotes subnetwork size and colors represent magnitude of the computed subnetwork sizes.

The average degree $<k>$ is shown in figure 6 as a function of subnetwork size. It grows with size approximately in logarithmic-linear scale $<k> (N) \sim 1.79 \ln N$.

Figure 6: **Average degree** $<k>$ [19]. Average degree as a function of the subnetwork size $N$. Colors correspond to one (black), two (red), three (green) and four (blue) base pairs in the secondary structure. The solid line corresponds to the numerical fitting. Note the logarithmic-linear scale. Dashed black lines are an analytical approximation, and the upper and lower bounds.

The average shortest path is measured in figure 7 for each subnetwork as a function of the network size. The average shortest path length is far away from the one estimated for random networks: $<d> \sim \ln N/\ln <k>$ [20]. In the model [19], $<d>$ scales with the logarithmic of the network size as $<d> \sim 0.63 \ln N$, which is represented by the solid black line. In the inset (7A), it is shown the shortest path length as a function of the average Hamming distance $<H>$ of each subnetwork. Both values are very close to each other. A similar situation is happening in the inset (7B) where is plotted the diameter of the network $d_{max}$, which corresponds to the number of steps between the most
distant nodes. It is given as a function of the lower bound $H_{\text{max}}$. The more abundant structures are those with the highest average connectivity. Then, these structures are embedded in large regions robust to mutations. Also, a large value of connectivity diminishes the fragmentation of the neutral network and thus facilitates large areas of robustness and it is more likely a punctuated equilibrium pattern in evolution. But connectivity in RNA folding models is not too much large, letting pathways for evolvability.

Figure 7: **Average shortest path** $<d>$ [19]. Dependence of the average shortest path on the subnetwork size $N$ for all folded neutral networks (colored circles), equivalent random networks (black squares) and theoretical predictions with a classical random model (green stars). Circle colors correspond to the number of base pairs of each subnetwork. The numerical fitting is plotted as a solid black line, while the analytical approximations correspond to the long-dashed black lines. Inset (A): relation between the average shortest path and the average Hamming distance of the subnetworks. Inset (B): relation between the longest distance between any pair of nodes of the network $d_{\text{max}}$ and the maximum number of different bases between sequences $H_{\text{max}}$ (maximum Hamming distance). In the insets, the dashed lines correspond to the lower bounds of $<d>$ and $d_{\text{max}}$, respectively.

Now we go to some measures related to Accessibility and Evolvability developed in the RNA folding model [22], which are given in figure 8. It is shown that Accessibility is higher for more abundant phenotypes, which means that the likelihood that point mutations over the entire sequence space yield to phenotypes that belong to large neutral networks is higher than the likelihood that point mutations yield to phenotypes that belong to small neutral networks. At the other hand, Evolvability decreases with the the phenotype Abundance and consequently point mutations to sequences in large neutral networks are less likely to yield novelty than point mutations to sequences in small neutral networks.
In other words, it is more difficult to evolve away from large neutral networks than small neutral networks.

These observations suggest that abundant phenotypes may be easy to find from the entire genotype space but it is difficult to escape from them. This supports the idea that large neutral networks enhance evolvability, and then robustness and evolvability are not contrary concepts: one helps the another one and both are necessary for evolution.

Figure 8: **Accessibility and Evolvability** [22] Accessibility and Evolvability (y-axes) as a function of Phenotype Abundance (x-axis). With Accessibility (Top) is shown the likelihood that a given phenotype will arise through point mutation. Random mutations are more likely to hit upon larger neutral networks that smaller neutral networks. With Evolvability (Bottom) is shown the likelihood of given phenotype will produce diverse alternative phenotypes upon mutation. Point mutations to sequences in large neutral networks are less likely to yield novelty than point mutations to sequences in small neutral networks. Black lines corresponds to fitted data. Accessibility and Evolvability have a log-transformed data in both axes.
4 Network Properties of a Genotype-Phenotype mapping based on a cellular automata

4.1 Cellular automata

4.1.1 Introduction

A cellular automaton (CA) is a discrete dynamical system, which was introduced by John von Neumann in the 1950s, and have been studied extensively as models of real-world systems and also in their abstract mathematical computational systems.

CAs are examples of systems with emergent properties: the global behaviour of a CA is not designed into its components, but arises from the complex interaction between these components. Because of this, it is useful to use these systems to extract information of how nature acts.

Wolfram [27] asserts that the long-term qualitative behaviour of a CA falls into one of four classes: homogeneous, periodic, chaotic, or complex.

4.1.2 Definition and dynamics

A cellular automaton (CA) which we are going to consider in this work is a tuple $\langle S, L, T, f \rangle$ of four components:

1. The set of states, denoted by $S$, which is composed by two Boolean states $S = \{0, 1\}$.
2. The lattice, $L$, which elements are the cells of the CA. We consider for our purpose a one-dimensional lattice with periodic boundary condition.
3. The neighbourhood template $T = \langle \eta_1, .., \eta_v \rangle$ is a sequence over $L$.
4. The local rule is a function $f : S^{|T|} \rightarrow S$, mapping neighbourhood states to cell states.

A configuration of the CA is a function $c : L \rightarrow S$. Each configuration assigns a state to each cell. The set of all configurations is denoted by $S^L$.

The state of cell $i$ in configuration $c$ is denoted $c[i]$. Let $i \in L$ be a cell. For the moment, we consider $v=3$, and consequently $|T| = 3$. The neighbourhood of $i$ is the sequence of cells $\langle \eta_{i-1}, \eta_i, \eta_{i+1} \rangle = \langle c[i-1], c[i], c[i+1] \rangle$.

The dynamics of the CA proceeds in discrete time steps $t=0, 1, 2, ...$, with the current configuration being updated on each time step. Let $c_t$ denote the configuration at time $t$. The state of cell $i$ at time $t+1$ is obtained by applying the local rule to the states of $i$’s neighbourhood at time $t$:

$$c_{t+1}[i] = f(c_t[i-1], c_t[i], c_t[i+1])$$
This process yields a function from configurations to configurations, denoted $F : S^L \rightarrow S^L$, called the global map of the CA. CAs are examples of discrete dynamical systems: they operate in discrete time steps, on a discrete space (the lattice), with a discrete state at each point in space (cell).

A configuration of the CA is a state of the dynamical system (Note that a state of the CA is not the same as the state in dynamical systems terms.) A sequence of configurations visited by the CA is a trajectory, which will reach an attractor cycle (a repeating sequence of configurations). Because we are applying a rule over a finite string (initial condition), the transient from the first time step to the $2^N + 1$ time step will return at least one repetitive string, and consequently an oscillating pattern would arise from the first repetitive string.

**Example:** A local rule acting over the current configuration,

<table>
<thead>
<tr>
<th>xyz</th>
<th>111</th>
<th>110</th>
<th>101</th>
<th>100</th>
<th>011</th>
<th>010</th>
<th>001</th>
<th>000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(x,y,z)$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Following this particular rule for every cell state $i$, for the first configuration in the table:

if $c_t[i - 1] = 1$, $c_t[i] = 1$, $c_t[i + 1] = 1$, then in the next time step $c_{t+1}[i] = 0$, by applying the function $f(1,1,1) = 0$.

### 4.2 Model

In order to reproduce the many-to-many relationship between genotype and phenotype we use the CA, from which we can define a set of initial conditions, a set of rules associated to external signals and factors acting over the genotype space and the attractor set, which are the phenotypes obtained from the process. By this way, it is possible to obtain dramatically different phenotypes from the same genotypes only by applying the different rules to the different initial conditions. In this sense, rules act as environmental and external factors over the genotype space. We are going to show that CA serves as an abstract model that has nothing to do with reality but its neutral network gives similarities with actual neutral networks and by which, without taking into account the several variables affecting the genome, it is possible to study mathematically, statistically or computationally this relationship. The model is chosen because of its simpler form that mimics the many-to-many genotype-phenotype relationship and their sensibility to initial conditions. In the model [21] where cellular automata is also used to map genotypes authors define a bipartite network such that genotypes are the Boolean update functions (rules) and attractors reached by the CAs are the phenotypes resulting from genotypes and initial conditions. Therefore, the same attractor can be reached by different genotypes and a single genotype can result in different attractors depending on the initial conditions. In this work there is clearly a many-to-many relationship between genotype and phenotype, whereas in our model this relation is hidden by considering genotype space as the combination of initial conditions and rules. In other words, we get a many-to-one relationship in our model. This selection of genotype space is done by this way because of the direct comparison with RNA folding models.
The genotype-phenotype relation can be treated as a relationship between information and environment. CAs gives to us an abstract space by which we can study biological properties arising from a neutral network. Since the relation between genotype and phenotype spaces is not one-to-one, we do a simulation by which one phenotype can be reach from different genotypes. The main goal is to map genotype space to phenotype space by a no bijective funtion, and analyse some biological properties of the neutral network that arises from this map.

To map this genotype-phenotype relation, we construct a cellular automaton, CA(N,v), which has only dependence on the one-dimensional lattice length and the neighbourhood length. For simplicity we call as strings the one-dimensional lattices.

• **Initial conditions:** Initial conditions, \( CI \), are the configurations from which the cellular automata is going to evolve. We consider all possible strings (configurations) of size \( N \), consequently we have \(|CI|=2^N\) initial conditions, due to its Boolean characterization.

• **Rules:** We consider the set of all possible rules, \( R \) to act over initial conditions. There is a total of \(|R|=2^{2^v}\) rules, being in this work \( v=3 \) fixed the neighbourhood length.

• **Genotype space:** The genotype space, \( G(N,v) = \{g \in G(N,v)\mid g = (g_{CI_1}, \ldots, g_{CI_N}, g_{R_1}, \ldots, g_{R_{2^v}})\} \) is the set of all possible genotypes, where each genotype is a boolean string composed by one initial condition and one rule. The size of the genotype space is \(|G| = 2^N \times 2^{2^v}\), which is the total number of genotypes.

• **Evolutionary process:** By applying, for every genotype, the rule \( f_g \) over the initial condition, we will get across transient configurations a fixed point (only one configuration) or a limit cycle (a finite subset of configurations cycling deterministically). Because of its deterministic nature, the CA will get one of the both possibilities at most in \( 2^N \) time steps. In both cases we call the final subset of configurations of a genotype an attractor.

• **Phenotype space:** The phenotype space \( P(N,v) \) is the set of attractors reached by the genotype space. It is clear that \(|P| < |G|\), since we expect to get a many-to-one relation.

We construct a neutral network to capture biological properties arising from point mutations. Every genotype is a combination of a rule with an initial condition. Genotypes which give rise to the same phenotype (attractor) are the nodes of a given phenotype. A neutral network depends only on the value of \( N \) (\( v \) has been fixed), and it is constructed such that the nodes are the genotypes and two nodes are connected if, and only if, both map the same phenotype and they only differ in one point mutation; their strings only differ in one cell (also called nucleotide). In literature we can see the definition of the Hamming distance of two genotypes as the number of different nucleotides between them. Following this definition, two nodes are connected if they map the same phenotype and
are at a Hamming distance of one.

To capture good statistical data, we compute every entire network for different values of N: from N=3 to N=8. With an increase of N, the network size is greater and more realistic mapping is got. The structure of every neutral network is a disconnected and undirected network. Then, there exists a set of neutral subnetworks, with most of the nodes totally disconnected. We denote by the letter S a subnetwork. Then, the total number of subnetworks (connected and disconnected) of a neutral network is denoted by $|N_S|$, the total number of subnetworks formed by the entire genotype space is denoted by $|G_S|$ and the size (number of nodes) of a given subnetwork is denoted by $|S|$.

4.3 Results

The RNA folding models have in general some similar properties with the mathematical model which is constructed by using a cellular automaton (CA). A CA model constructs a bipartite network where genotypes are linked if they share a common phenotype. In this model we have $2^v$ different genotypes and the computation is always done for $v=3$ and it gives at most $2^v \times 2^N$ possible phenotypes, where $N$ denotes the length of a binary string by which genotypes and phenotypes are represented [21]. A very different upper bound is found by the RNA folding model, in which the number of different secondary structures $S_l$, or phenotypes, is also dependent of the sequence length $l$, and it is found an upper bound of $S_l$ that is valid for sufficiently large sequences: $S_l = 1.4848 \times l^{-3/2} (1.8488)^l$ [18]. Consequently, $S_l \ll S_N$ the two bounds are very far away from each other with $l=2^v$. However, we cannot extract from here the actual distribution of neutral networks sizes, because it has not a well defined average.

Before analyzing statistical measures of neutral networks, it is interesting to show that we get a similar structure of the genotype-phenotype mapping with the CA model [19]. These models constructed by binary strings are abstract and its implementation has nothing similar the actual structure of the mapping extracted by the RNA folding method. In the following table we have, for each N: the total number of genotypes $|G|$, the number of phenotypes $|P|$ which has been mapped from the genotype space (therefore the number of neutral networks) and the resulting number of subnetworks of the entire phenotype space $|G_S|$.

| N  | $|G|$ | $|P|$ | $|G_S|$ |
|----|------|------|-------|
| 3  | 2048 | 799  | 815   |
| 4  | 4096 | 4    | 4     |
| 5  | 8192 | 1992 | 1992  |
| 6  | 16384| 8569 | 8667  |
| 7  | 32768| 15508| 15589 |
| 8  | 65536| 31420| 31681 |
These values gives a similar structure for every length scale N. Because of the CA nature and because we have assumed the genotype space as the combinatorial set of initial conditions and set of rules, we get that the number of phenotypes is large and the mapping process gives most of the structures full connected in the resulting neutral network. This is because from the mapping we have linked two genotypes if both give rise the same phenotype and both differ at a Hamming distance equal to one, i.e., they differ un one character. Figure 9 shows a qualitative picture about how this subnetwork structure is obtained. Consequently, $|P|$ is close to the total number of subnetworks and this is a signal that we are going to obtain robustness properties in the network. This is a very different scenario that it is shown by the RNA folding model which shows that the total number of subnetworks is very much greater than the number of phenotypes. Specifically, they have $4^{12} = 2451912$ number of genotypes, which corresponds to number of folded sequences (discardding the ones that are not folded), a number of 57 different secondary structures (phenotypes) and the resulting number 645 of subnetworks that have been decomposed. Then, we know that we are going to find larger areas of neutrality in our model than the found in the RNA folding in which it is also likely a large number of clusters with less range of robustness and more accessibility that actually give rise to the same phenotype [19]. However, similar results are shown in [21].

Figure 9: Qualitatively, this figure represents the two main processes: in the mapping process we have circles which represents phenotypes and red ones inside which are the genotypes that give rise the same phenotype. Now, to construct the neutral network, a decomposition of these structures is done, because firstly is not assumed that genotypes that are in the same circle are at a Hamming distance one.

We have ordered similarly to RNA folding models neutral networks as a function of its size. Then we have compute the phenotype abundances for each phenotype, given in figure 10 and ranked in order of abundance. We get similar results with the RNA folding models [22-19]. The distributions of phenotype abundances appear similar across all genotype lengths and we can conclude that
few of these phenotypes are quite common with many genotypes folding into them, while many others are quite rare, formed by few genotypes. This result is similar to the phenotype abundance of the RNA folding model showed in figure 2.

Figure 10: **Phenotype abundance distribution** Phenotype abundance as a function of the rank ordering of phenotype abundance. Colors differentiate the sequence length at which the measure is computed. There are a relatively few highly abundant phenotypes (a few highly sets of genotypes that give rise to the same phenotype) and many rare ones.

Also related with the size of neutral networks, we show in figure 11 the relation between the subnetwork sizes that belong to the same neutral network as a function of the corresponding neutral network size. It has similar results with [19], which is illustrated in the insets of figure 4. Here there is a difference between the two models: For the RNA folding model, most neutral networks are divided into different subnetworks because in the diagonal there are no so many points as in our model.
Figure 11: **Subnetwork size** Each subnetwork size as a function of the network size at which it belongs. There are many neutral networks which are connected and therefore they formed a unique subnetwork (diagonal straight line drawn by points) and fewer neutral networks which are divided into subnetworks. Colors differentiate the sequence length.

In the following table different size measures are shown and are related to subnetwork properties together with a simple measure of connectivity and genotypic robustness. For every sequence length N it is computed the total number of nodes (genotypes, which are formed by one initial condition and one rule associated) \( |G| \) of the network, the total number of connected subnetworks, \( |S_c| \), considering only the nodes that have at least one link. The number of connected nodes, \( |G_c| \), which is the total number of nodes belonging to some connected subnetwork. Also we compute the total number of disconnected nodes, \( |G_d| \), and the total number of subnetworks of the total network \( |G_S| \), which is formed by the connected and disconnected nodes. With this information it is also computed the connectivity \( |C| \) and the genotypic robustness \( |GR| \), which is the fraction of the total number of possible point mutations to a given genotype that are neutral. In other words, is the average degree of the network.

| N  | \( |G| \) | \( |S_c| \) | \( |G_c| \) | \( |G_d| \) | \( |G_S| \) | \( |C| \)   | \( |GR| \)  |
|----|-------|-------|-------|-------|-------|-------|-------|
| 3  | 2048  | 17    | 1250  | 798   | 815   | 0.6103| 0.3581|
| 4  | 4096  | 4     | 4096  | 0     | 4     | 1     | 0.6354|
| 5  | 8192  | 8     | 6208  | 1984  | 1992  | 0.7578| 0.4173|
| 6  | 16384 | 115   | 7832  | 8552  | 8667  | 0.4780| 0.2512|
| 7  | 32768 | 89    | 17268 | 15500 | 15589 | 0.5269| 0.2736|
| 8  | 65536 | 240   | 34095 | 31441 | 31681 | 0.5202| 0.26676|

It is clear that there is an increase of the number of subnetworks with N, despite of the case N=4 in which CAs properties give rise a total conexion of the network and a classification in four neutral subnetworks. Subnetworks represent
groups of genotypes that are going to give the same biological functionality because they share the same attractor, the same phenotype and this means that if genotypes are information source, this information is acting equally to the environment.

The connectivity is high for low values of N, showing that for shorter genome lengths there is more resistance to phenotypic changes from mutations. Most of genotypes have the same functionality if mutate for short genomes. In contrast, this connectivity seems to relax for high values of N. High connectivity is shown for the neutral networks N=4 and N=5 (being N=4 a special case), and then there exists a set of common structures, i.e, phenotypes which are related to community properties. We have no information from the connectivity of the structure of these hubs, so we can say nothing about the robustness present in the network. We need more statistical measures for that. This value for N=5 is very close to the one found in another model of CA for a neutral network which maps also for N=5 and k=1. Specifically, they find $C = 0.816$ [21], which also shows a high connectivity.

The genotypic robustness of the network seems to not give to us any information about the topological structure and consequently we can no extract any biological conclusion. However, it can be seen as a key to differentiate a general property of networks of different sequence length: Since for N=6,7,8 the average degree is similar, it is possible to do a good comparison between them. The three networks show low robustness but sufficiently high values to let the genotype explore new functions by point mutations.

For every sequence length N, a histogram of the number of subnetworks of size $|S|$ is represented in figure 12. There are few subnetworks of large size and most of them are composed by a small number of genotypes. This means that there are few common structures within large neutral regions, where there are many possible pathways by point mutations keeping the structure (phenotype). This characterize the network with few regions of high robustness. At the other hand, since the majority of the subnetworks are of small size, they are not common structures and are formed by rare phenotypes. These subnetworks with low robustness also help to evolvability, letting pathways passing through them and change more easily phenotype by point mutations. The data fit is found to be around $|S|^{-1.8}$.
As a first step, robustness is measured by using the degree distribution. In figure 13, it shown that there is a spread of the distribution for higher values of sequence length N, keeping a lower but necessary grade of robustness due to the most abundant neutral networks. In order to compare this measure with another one extracted by RNA folding and represented in figure 5 [19], it is also computed the degree distribution for subnetworks in figure 14. The degree distribution appears to be larger for high values of the degree k, so there exists large regions of neutrality, where robustness is present. This is suggested because robustness is relaxed for high values of N and there is always a single-peaked in all cases, with the maximum value approaching higher values of k. At least, this happens for sufficiently large values of sequence length N. Note that in the representation of our model N represents the associate sequence length whereas in the compared result N is the exact number of the subnetwork size, in our model denoted by $|S|$. 

Figure 12: **Histogram of subnetwork sizes** Number of subnetworks of size $|S|$ (y-axis) as a function of subnetwork sizes for every sequence length N, which are indicated by colors. There are few abundant subnetworks and many of small size. The black curve is the fit of the data.
Figure 13: **Degree distribution p(k)** Degree distribution of genotype space for every sequence length N denoted by colors. It suggest a spread with higher values of N. This means that for genotype spaces smaller (N=3,4,5), there are more common structures with similar average degrees whereas for big genotype spaces (N=6,7,8) there are more variability, where we find nodes with high degree belonging to abundant neutral networks and another nodes belonging to smaller neutral networks.
Another local measure of robustness is the average degree of each subnetwork, which is computed and represented in figure 15 as a function of the subnetwork size it belongs to. Average degree grows with subnetwork size, with the approximated fit \( \langle k \rangle(N) \sim 1.05 n(N) \), which is a similar result with the one obtained [19] in figure 6 \( \langle k \rangle(N) \sim 1.79 n(N) \).
Figure 15: **Average degree** $<k>$ Average degree of subnetworks (y-axis) as a function of the subnetwork size they belong to (x-axis). Colors denote different sequences length $N$ for which this measure is computed. The black line is the fit data. Note the linear-logarithmic representation of x-axis data.

It is shown in figure 16 the average shortest path for every subnetwork and for every neutral network as a function of the subnetwork size. It is a quantification of the navigability of neutral networks. It grows with the subnetwork size and its functional dependence if far from the analytical result in random networks which predicts the fit $<d> \sim \ln N/\ln <k>$. In our model the fit is found to be $<d> \sim 0.504\ln N$, whereas in the RNA folding model the fit showed in figure 7 [19] is found to be $<d> \sim 0.63\ln N$. For $N=3,4,5$, the results indicate large regions of neutrality where navigability is also large and consequently pathways by point mutations will keep the structure of its phenotype for a while. Therefore, it is for these sequences lengths the most probable a punctuated equilibria. However for $N=6,7,8$, there is a grow of the average shortest path of subnetworks as a function of its size and consequently large regions of navigability are found to be in abundant subnetworks whereas there are many subnetworks with low navigability, allowing changes in phenotype with a few number of point mutations.
Figure 16: **Average shortest path** $< d >$ Average shortest path as a function of the subnetwork size $|S|$. Colors denote the different sequence lengths at which this measure is computed and the green line is the data fit.

It is also found in figure 17 (Top) that for large neutral networks, represented as Phenotypic Abundances, the CA model gives high accessibility values for large neutral networks, and consequently abundant phenotypes are most likely to find through point mutations applied to the entire genotype space. The fit $\log A \sim 0.0901 \log N$ is done for the sequence length $N=6$ because it is the network which gives enough points to get statistical conclusions. For evolvability it is found that the likely to yield new phenotypes by point mutations is lower in large neutral networks. Therefore, for small neutral networks there is more variability through point mutations. The fit is also calculated for the network $N=6$, $\log E \sim -0.0973 \log N$. These measures suggest that pathways from the entire genotype space will go to large neutral networks, which have the capacity to keep the structure by many pathways through point mutations and it is difficult to escape from them. These results are the same than the ones found in figure 8, with Accessibility growing with neutral network size and Evolvability decreasing with neutral network size.
Figure 17: **Accessibility A and Evolvability.** (Top) Accessibility as a function of the neutral network size and (Bottom) Evolvability also as a function of the neutral network size. Both are represented for different sequences length N denoted by colors. The green lines corresponds to the fit of data belonging to the case N=6. Both axes are in logarithmic transformed data.

In figure 18 it is plotted Neighbour Abundance B, a proposed measure for the model. For phenotypes which are formed by only one genotype, there is variability about the Neighbour Abundance, suggesting that the robustness of
the new phenotypes reached by point mutations is very variable. In other words, it is very uncertainty for these neutral networks the stability of phenotypes that can be reached by pathways through point mutations. From very small neutral networks there is an increasing of $B$, so in this region the increasing of the neutral network size means the increasing of its unstability. This is because only in this region neutral networks with higher robustness become to appear as possible outcomes through point mutations, and therefore these neutral networks with higher robustness are more likely to be reached by point mutations. By increasing the size of neutral networks there is a critical neutral network size, around $|S| = 10$, which indicates that there is a maximum unstability reached. From this peak $B$ becomes to decrease, and it is in this region where is more likely for larger neutral networks that non-neutral point mutations take a genotype to another one belonging to a neutral network with lower robustness.

![Figure 18: Neighbour Abundance $B$](image)

**Figure 18:** Neighbour Abundance $B$ Neighbour Abundance as a function of the Neutral Network size for different sequences lengths $N$ denoted by colors. Both axes are in logarithm transformed data. It showed for small networks that once a genotype scapes it is less likely that it goes back to its original phenotype. In other words, for small neutral networks we find that it is more easy to scape and non-neutral point mutation will take it to a most robust phenotype.

Note that in general, for $N=6,7,8$ we get sufficiently data for comparing the results with the ones obtained in the RNA folding models and for $N=3,4,5$ it is more difficult to get good conclusions.
5 Conclusions

The main difference of the CA from the RNA folding models is that the genotype-phenotype mapping give rise many common phenotypes, whereas in RNA folding models there are a few number of common structures and many rare ones. Also, the neutral networks in CA are very connected, a property which is not found for the RNA folding models. This is not represented by connectivity, because it is a general measure of the entire genotype space. But it is shown in the first tables where we have the number of different phenotypes (number of different neutral networks) and its abundance (or also called frequency). Although there is this difference, the general structure is similar to each other in the different models in the sense that we get similar network properties: high degree nodes are more frequent and the average degree grows with neutral sub-network sizes. Also, abundant structures which define the common phenotypes as hubs, are very much robust to mutations than small neutral networks. But this property facilitates evolvability because we find that the presence of small networks are needed to variability of phenotype with time. Since it is measure Evolvability, we know that there is more variability for small neutral networks; small neutral networks give any non-neutral point mutation to a larger types of phenotypes than large neutral networks. This means that pathways through point mutations can pass through small networks and although are not going to spent to much time in them, the pass through them leads to changes in phenotype because small networks have also more variability, i.e., point mutations give rise to a large range of different possible phenotypes. This general observation is suggested in both models: the RNA folding model and the CA.

Although the RNA folding model is more realistic because of the parameters introduced in the model as temperature, minimum free energy, and so on, it is interesting to the scientific community to have the possibility of work on this abstract and mathematical model to study biological properties arising from the neutral network.

To finalize, note that for N=4 we have a special case of regularity, where we have only four neutral networks (four different possible phenotypes) and it is not the best sequence length for studying genotype-phenotype mapping. The first N which seems to show statistical similar data to other models is N=6. RNA folding models use sequences length of at least l=12, and if we would to compare our results in the same length scale, we should use $N \sim 16$ because of the comparison of the total number of genotypes in genotype space. This N=16 is too computationally expensive. The higher N it has been computed is N=8 and for the three lastest measures N=7 is too much cost of computationally time.
Appendix

Genotype-Phenotype mapping

The Cellular Automaton together with the neutral networks has been constructed by using the C++ language programming. For constructing the genotype-phenotype mapping, there are two main parts in the program:

- **Genotype Space**
  
  Genotype space is composed by all combinations of initial conditions and rules. Therefore, one genotype is considered as a given initial condition and a given rule. We store all possible configurations in two matrices: one for keeping initial conditions and another one for keeping rules. The matrix which contains initial conditions is formed by all possible binary strings of length $N$. This matrix has a two dimensional $2^N \times N$ size.

  A matrix template is used to keep all possible neighborhoods of length $v$ ($v=3$ fixed), being $2^v$ the number of rows of the matrix. Every row of this template is a neighborhood of length $v$. For example, in figure 19 the matrix template constructed by our program is shown (it serves for all $N$ and $v=3$). This template is used to define the rules such that a given rule is defined by associating a binary number to each row. So, there are $2^{2^v}$ total rules. In figure 19 we have only one rule defined over the template. When the program looks for the attractor through the transient process from the initial condition, the specific rule acts over a binary string such that for every neighborhood given in the template, the character located at the middle of the neighborhood is changed to the binary number defined by the rule.

![Figure 19: Template matrix of all possible neighborhoods and a given rule defined over the template](image1.png)

Rules are defined over the template such that column number 'x' contains the binary number associated to the neighborhood kept in the row number 'x' of the template.

Rules are stored in a new matrix where each row is formed by one rule and each column of the matrix is defined for a specific neighborhood. We
get a two dimensional matrix of size $2^{2v} \times 2^v$ where each row is a defined rule.

In this way, each genotype has length $N \times 2^v$ (N because of the initial condition and $2^v$ because of the rule). Since we've fixed $v=3$, the program has been runned from $N=3$ to $N=8$ and for each case figures have their representation by using the 'N' notation, because it is the only variable changed and the source of the different obtained data. So this notation is not a direct representation of genotype space sizes.

Since we have two separate matrices to define the genotype space, there is an enumeration to organize genotypes: A genotype has a natural number associated to it and by which it is represented. This is showed in figure 20 where we denote a given initial condition by CI(x) with 'x' being the number of the row of the matrix which defines initial conditions. Also, a given rule is defined by R(y) where 'y' is the number of the row of the matrix which defines rules.

![Figure 20: Genotype enumeration](image)

Genotypes are associated to a natural numbers following the showed order. This numbers are used in all program to keep and look for genotypes in a simple way.

For each genotype, it is kept in a new matrix its natural number associated, its initial condition (by using the number of the row of the first matrix in which initial condition of the genotype belongs to), its rule (by using the number of the row of the second defined matrix in which rule of the genotype belongs to), and the decimal number which is obtained by the binary string of the initial condition followed by the binary string of
the rule. This last decimal number is kept to facilitate the searching of genotypes.

- **Phenotype Space**

Following the order of the stored genotypes (genotype '0', genotype '1', and so on) in the previous classification, for each of them it is calculated and kept in a new matrix the resulting phenotype. It is also used for reducing memory the period of cycles of each phenotype. Phenotypes are obtained by applying the rule over the initial condition of the genotype: Rule is applied over a binary string at once, in only one time step, and therefore neighbourhoods which are the dependence of the local rule applied are all actualize at the same time. Note that there are boundary conditions. So, in the first time step rule is applied over the initial condition, in the second time step rule is applied over the resulting binary string of the previous step, and so on. There is a thermalization process in which we let running the process $2^N$ times to be sure that we have got the resulting phenotype, the attractor of the CA. Once we have this attractor, applying again the rule do not change the result. Only in the case of cycles with period greater than one, if you apply once again the rule the set of strings defining the attractor would only change their row positions but not its composition.

**Neutral Networks**

In order to increase the computational efficiency, only phenotypes with the same period are compared. By comparing phenotypes, it is store all genotypes that give rise to the same phenotype or attractor. Previously, it has been calculated the total number of different attractors to define the size of the matrix where is going to be stored all genotypes. Rows of this matrix are composed by genotypes that give rise to the same attractor, and therefore they are the neutral networks of the genotype-phenotype mapping. It is true that by this procedure the matrix is going to have empty cells, but it is easier to store all data. Finally, to obtain the sub-networks, it is compared genotypes belonging to the same neutral network such that pairs differing in only one digit (of their initial conditions and rules) are connected.
References


an alternative to phyletic gradualism.” (1972): 82-115.


