

Chapter 9

Decoding genomic information

Giuditta Franco and Vincenzo Manca

9.1 Introduction

Science has many important challenging open problems. Some of the approaches to solve them are still intractable using conventional computing, which may be briefly identified with the Turing/von Neumann paradigm. Even if processors become faster and more compact, and memory storage larger, standard computer science has clear limits of miniaturization, speediness, and parallelism scalability. Moreover, only a specific model of computation is investigated (albeit on a massive scale), with well known intrinsic limits (of decidability, of complexity) which cannot be overcome by any technological advancement (Conrad, 1988).

Research in *Unconventional Computing* (UCOMP) takes a different view, in exploring alternative computational approaches (with properties such as massive parallelism, approximation, non-determinism, adaptation, redundancy, robustness, learning, self-organization, reproduction, competition), in order to increase the range and power of computation available to us.

A main difference between the two worlds is the (digital) data format. In conventional computing it consists of strings of digits, or formal languages, while in unconventional/biological computing data populations or multisets of strings (e.g., molecules, bacteria) are processed by the computation (Manca, 2013). Beside a redundant and undefined number of copies for each string, different notions arise of unconventional complexity (Almirantis et al., 2014; Lynch and Conery, 2003) with specific properties of uniformity and confluency of solutions (Păun, 2016).

UCOMP is often assumed to be any way to compute that is different from Turing's model. Crucial aspects of Turing Machines, which are missing in many unconventional computing approaches, are the existence of a global clock, and of a program to execute as a list of instructions. Some

Dipartimento di Informatica, Università di Verona, Italy

examples of a different approach, where the computation program is not centralised, include: distributed, cloud, network computing (namely VPN: virtual private network), machine learning in the context of artificial intelligence (e.g., computer learning games), evolutionary computing (including genetic algorithms), light/optical and quantum computing (also exhibiting an unconventional means to store information), molecular computing (employing unconventional means and instructions to compute).

Bioalgorithms with interesting massive parallel strategies have been developed in the context of DNA computing recently, for example (Ratner et al., 2013; Rothmund et al., 2004), while laboratory biotechniques have been improved in terms of computation precision and efficiency (Franco, 2005; Franco and Manca, 2011a; Manca and Franco, 2008). Network-based algorithms, for example simulating metabolism (Castellini et al., 2011; Franco and Manca, 2011b; Manca et al., 2013) or immunological processes (Castellini et al., 2014; Franco et al., 2008; Franco and Manca, 2004), have been developed in the context of systems biology and membrane/cell computing, to better understand natural processes, as well as to formulate new computational models. As attested by numerous dedicated series of books, international conferences, and journals, UCOMP includes *bioinformatics* (bio-computer science), which is the “closest upper envelope of the computability inspired by biology” (Păun, 2016).

In search of solutions for current (often urgent) questions, for example from the bio-medical area, (applied) mathematics and computer science often develop and provide *ad hoc* models, techniques, and tools to tackle problems. This approach is related to *mathematical biology*, where mathematical modelling and simulation are applied to biological, biomedical and biotechnology research.

A marvellous operative system working in nature is the genome, carrying the main information generating life of organisms and their evolution, and having a system of molecular (reading, writing and signal transmission) rules, orchestrating all cell functions and information transmission to cell daughters. Most of these rules and especially the ways they cooperate are unknown. This is a problem of great scientific and medical interest, due mainly to currently incurable genetic diseases.

After the revolutionary human genome sequencing project, and the subsequent decade-long joint project ENCODE, involving 440 scientists from 32 laboratories around the world¹, sequences of genes, chromosomes, and whole genomes from numerous species are downloadable by freely accessible databases². Also, the ENCODE project has systematically mapped regions of transcription, chromatin structure, transcription factor association, and histone modification, providing new insights into the mechanisms of gene

¹ at MIT, Harvard, Stanford, and SUNY in the USA, and at universities in Germany, the UK, Spain, Switzerland, Singapore, China, and Japan

² such as the NCBI at www.ncbi.nlm.nih.gov/sites/genome, UCSC at hgdownload.cse.ucsc.edu/downloads.html, and EMBL-EBI at www.ebi.ac.uk/genomes/

regulation (Neph et al., 2012; Spivakov et al., 2012), and a biological annotation of the human genome (for more details, see for example (Dunham et al., 2012; Franco, 2014)). However, such an avalanche of data would be difficult to handle and understand without the use of powerful mathematical and computational tools.

Our work here outlines and follows some trends of research which analyze and interpret (i.e., decode) genomic information, by assuming the genome to be a book encrypted in an unknown language. This analysis is performed by sequence alignment-free methods, based on information theoretical concepts, in order to convert the genomic information into a comprehensible mathematical form and understand its complexity (Almirantis et al., 2014; Lynch and Conery, 2003; Vinga, 2013).

Sections of this chapter are adapted from Bonnici and Manca (2016), Castellini et al. (2012), Franco (2014), Manca (2015), and Manca (2016), to be considered foundational references of the bibliography, while relevant related papers which pursue genomics investigations by the same aim are Chor et al. (2009), Fofanov et al. (2008), Li et al. (2016), Sadovsky et al. (2008), Sims et al. (2009), Vinga (2013), Zhang et al. (2007), Zheng et al. (2017), and Zhou et al. (2008). After a nutshell of the state of the art given as a brief overview of approaches in the area, we present our viewpoint and results on genomic wide studies by means of mathematical distributions and dictionary-based analysis inspired by information theory, that we call *Infogenomics*.

9.2 Overview

The role and the contribution of Shannon Information theory to the development of molecular biology has been the object of stimulating debates during the last fifty years (Gatlin, 1966). The concept of information itself, if viewed in a broader perspective, is very pervasive and far from being completely defined (just like the concept of energy in physics (Manca, 2013)), while classical information theory has been conceived at a high technical level (Fabris, 2002; Thomas and Cover, 1991). The concept of information (and complexity (Almirantis et al., 2014; Lynch and Conery, 2003)) in biology is still a debated problem, so the application of information theory has often to be adapted to the context, namely when outside of standard computer science (Vinga, 2013).

Information science was born with Norbert Wiener from a philosophical viewpoint, with John von Neumann from a more pragmatic viewpoint, and finally with both Claude Shannon, who defined a mathematical measure of information (in the processes of representation, communication, and transmission), and Alan Turing, who set down its famous mathematical model of computation machine in terms of data storing and program execution.

Information is related to probability (an event is as informative as it is rare to happen), and establishes mathematical relationships between digital and probabilistic concepts definable over strings. Here we work on formal languages or codes, composed by words appearing on a given genome, then we develop (genome-wide) information theoretic methods along the perspective of alignment-free methods of genome analysis (Vinga, 2013; Vinga and Almeida, 2003).

Genomes are sequences of nucleotides from hundreds to billions of base pairs long. As sequences of symbols they determine dictionaries, that is, formal languages constituted by words occurring in them. They encode the language of life, as dictating the functioning of all the organisms we consider living beings. A main open problem in science is to find any key to understand such an encrypted language, which directly effects the structure and the interaction of all the cellular and multicellular components. It is like having a book in an undeciphered language (Manca, 2013; Manca, 2015; Manca, 2016; Mantegna and al., 1994; Percus, 2007; Searls, 2002). A genome is however a special book, being diachronic (rather than synchronic): it reports in its own sequence all developments it had passed through during evolution. All fragments which were mutated, duplicated, assembled, or silenced are still present in the genomic sequence to some extent, and could tell us the paths which evolution has followed to generate modern organisms.

We focus on genome-wide numerical properties, by computing, analyzing, and comparing informational indexes, with the aim to discover which of them can be relevant to identify characteristics of genomes that are of biological or clinical interest. Related previous dictionary-based studies of genomes may be found in Crochemore and V erin (1999) and Vinga and Almeida (2007), where entropy measures are employed to estimate the randomness or repeatability of DNA sequences (Holland, 1998; Kong et al., 2009), even in function of their different ‘biological complexity’ (Annaluru et al., 2014; Castellini et al., 2015; Deschavanne et al., 1999; Franco and Milanese, 2013; Spivakov et al., 2012).

In the post-genomic era, several attempts are emerging to understand genomic complexity. Works on modelling biological sequences by means of formal languages have been proposed, along with an extensive investigation (Searls, 2002), including a linguistic semantics inspired approach (Mantegna and al., 1994; Neph et al., 2012; Searls, 2002). A recent development is the introduction of context-free grammars formalizing design principles for new genetic constructs, by starting from a library of genetic parts already organized according to their biological function (Y.Cai and al., 2007).

In general, the definition, computation, and analysis of a few informational indexes have highlighted some properties of genomic regularity and specificity that may be a basis for the comprehension of evolutionary and functional aspects of genomes.

9.2.1 Dictionary based indexes

A *genomic dictionary* is a set of strings occurring in a given genome G . We denote by $D_k(G)$ the dictionary of all k -mers occurring in G . A word α may occur in G many times, and we call *multiplicity* of α its number of occurrences. It is easy to verify that the number of occurrences of k -mers in G corresponds to the maximum cardinality reachable by a dictionary of k -mers within genomes of the same length, and that the multiplicities average decreases with the k -value.

A word with multiplicity greater than one is called a *repeat* of G , whereas a word with multiplicity equal to one is called a *hapax*. This term is used in philological investigation of texts, but it is also adopted in document indexing and compression (Giancarlo et al., 2009; Sadovsky et al., 2008). A nullomer (Hampikian and Andersen, 2007) or forbidden word is a sequence that does not appear in the genome. Franco and Milanese (2013) propose a bioinformatic investigation on genomic repeats that occur in multiple genes, of three specific genomes, thus providing non-conventional graph based methods to abstractly represent genomes, gene networks, and genomic languages. By normalising multiplicities one obtains frequencies, and a consequent discrete probability distribution over genomic words.

Recent approaches may be pointed out, based on the empirical frequencies of DNA k -mers in whole genomes (Chor et al., 2009; Wang et al., 2015; Zhou et al., 2008). However, any set of words (factors) occurring in a genome provides a genomic dictionary, and some indexes related to characteristics of dictionaries may be defined on genomes. For example, $MRL(G)$ is the length of the longest repeat of G ; MRL is the minimum length such that k -mers in the dictionary with k greater than MRL are all hapaxes; $MHL(G)$ is the minimal length for hapaxes in the genome G ; $MFL(G)$ is the minimal forbidden length, that is, minimal length of words that do not occur in G (Fici et al., 2006; Herold et al., 2008).

When genomic complexity is considered, it cannot be easily measured by parameters such as genome length, number of genes, CG-content, basic repeatability indexes, or their combinations. An information theoretical line of investigation based on k -mer dictionaries and entropies may be found for example in (Bonnici and Manca, 2015b; ~~Bonnici and Manca, 2016~~; Manca, 2013; Manca, 2015; Manca, 2016; ~~Manca, n.d.~~), which is aimed at defining and computing more complex informational indexes for a representative set of genomes. In this context, it is natural to assume that the complexity of a genome increases with its distance from randomness (~~Chor et al., 2009; Fabris, 2002; Holland, 1998; Kong et al., 2009; Sims et al., 2009; Zhang et al., 2007~~), as identified by means of a suitable comparison between the genome under investigation and random genomes of the same length. So the identification of appropriate genomic distributions is crucial for looking at the genomic information.

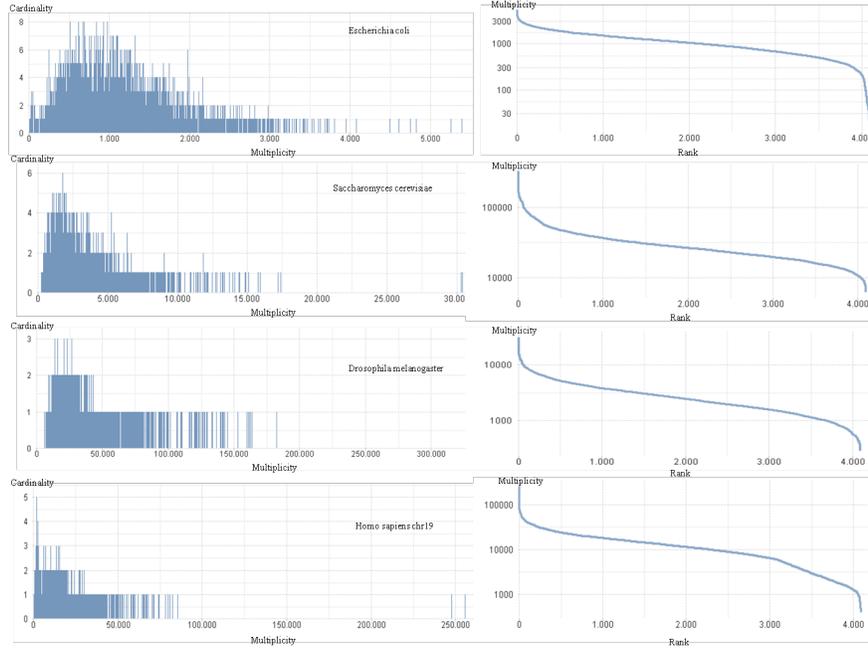


Fig. 9.1 Multiplicity-cardinality and rank-multiplicity Zipf distributions of some organisms are reported. [Reproduced by courtesy of the authors of (Castellini et al., 2012)]

9.2.2 Genomic distributions

For any numerical index I_k with parameter k , the distribution $k \mapsto I_k$ can be defined on a genome, and its classical statistical parameters (mean, standard deviation, median, mode, etc.) may be derived as further indexes (Castellini et al., 2012; Manca, 2016).

Word distribution in a genome may be represented along a graphical profile, which measures the number of k -words having a given number of occurrences. We call such curves the multiplicity-cardinality k -distribution of a genome, having the same information of a rank-multiplicity Zipf map as usually employed to study word frequencies in natural languages (Mantegna and al., 1994); see Figure 9.1. Several other nice representations of genomic frequencies may be found in the literature, for example by means of images; in Deschavanne et al. (1999) distance between images results in a measure of phylogenetic proximity, especially to distinguish eukaryotes and prokaryotes.

An intriguing genomic distribution, called the recurrence distance distribution (RDD), has been computed for several genomic sequences by Bonnici and Manca (2015b). For a given word α (say a 3-mer or a 6-mer), RDD associates a distance-value n to the number of times that α occurs at distance n from its previous occurrence in G . The well-known peak 3-periodicity has

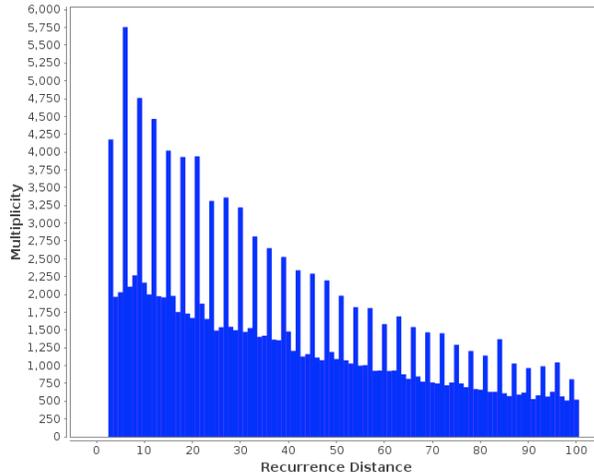


Fig. 9.2 RDD related to the word AGA, computed on the human exome. [Reproduced by courtesy of the authors of (Bonnici and Manca, 2015b)]

been confirmed by Bonnici and Manca (2015b), and is easily visualized by means of RDD plots; see a simple example in Figure 9.2. The same periodicity has been observed in bacteria whole genomes, human protein coding exon regions, and exons of ncRNAs (non-protein coding RNA). More interestingly, a connection has been established between distance peaks (on k -mers with $k > 3$) and (approximate) repetitive elements between the corresponding recurrent k -mers.

9.3 Contribution to UCOMP Aspects

Our topic focuses on an informational analysis of real genomes, and may be framed within a new trend of computational genomics, lying across bioinformatics and natural computing, depending on which type of methods are employed, both to analyze high-throughput biotechnology genomic data and to develop a mathematical modelling of basic laws underlying information structured in genomes. This approach enters the field of UCOMP, as it is outside the standard model of conventional computing that underlies the implementation of commercially available devices. Conventional computation, as nowadays implemented by parallel algorithms and parallel architectures for big data mining, are important in that they need to be employed to simulate our data-driven mathematical models, as data rich information sets or genomic databases. In terms of traditional computing architectures, big data

and massive parallelism will be involved and developed alongside research lines based on infogenomic interests.

9.3.1 Speed

One of the current challenges is to find good genome representation to speed up the analysis of interest. However, distributed and parallel computing will be necessary to successfully handle big volumes of variable data in practice, and the capabilities of existing big-data frameworks should be combined with bringing the computation as close as possible to the data (Pan-Genomics Consortium, 2016).

Advantages of k -mer-based representations include simplicity, speed and robustness (Wang et al., 2015; Zhou et al., 2008). Among alternative indexing methods, we mention as an example the Burrows–Wheeler based approaches, which append the extracted contexts around variations to the reference genome.

9.3.2 Resource(s)

The advent of rapid and cheap next-generation sequencing technologies since 2006 has turned re-sequencing into one of the most popular modern genome analysis workflows. An incredible wealth of genomic variation within populations has already been detected, permitting functional annotation of many such variants, and it is reasonable to expect that this is only the beginning.

9.3.3 Quality

Next-generation short-read sequencing has contributed tremendously to the increase in the known number of genetic variations in genomes of many species. However, the inherent limitations often provide us with error prone and uncertain data.

The most promising developments in sequencing technology involve single-molecule real-time sequencing of native DNA strands, which is widely used for variation discovery and genome assembly. The MinION device (Oxford Nanopore Technologies) provides even longer reads of single DNA molecules, but has been reported to exhibit GC biases. Data generated on the MinION platform have been successfully used for assembly of small genomes and for unravelling the structure of complex genomic regions.

Despite this progress, sequencing reads are not yet sufficiently long to traverse and assemble all repeat structures and other complementary technologies are necessary to investigate large, more complex variation (Pan-Genomics Consortium, 2016).

One of the computational (and modelling) current challenges is indeed to find the knowhow dealing with data uncertainty propagation through the individual steps of analysis pipelines (Castellini et al., 2011; Cicalese, 2016), which need to be able to take uncertain data as input and to provide a level of confidence for the output made.

9.3.4 *Embeddedness*

Clear applications of the above approach may be identified for metagenomics (i.e., the genomic composition of microorganisms sampled from an environment) and viruses (which are notoriously mutation executors), apart than on human genetic diseases, as cancer. Besides, metagenomics can be applied as well to gain insights on human health and disease.

Personalized medicine is one the main goals (Ginsburg and Willard, 2017), having a notable social and economical impact, while political and ethical/privacy issues should be discussed and regulated for (genomic) data sharing.

9.3.5 *Programmability/Programming*

Ad hoc developed methods to analyse genomes belong to a computing model that is universal (in the Turing sense), ~~which is of course also programmable. In this case, a biological substrate is the starting point (Consortium, 2001), to which sequencing algorithms are applied to get the final genome sequence. Then, analysis are performed with~~ powerful software developed for the scope.

~~An alternative informational concept of (molecular) programmability may be developed,~~ as in Bonnici and Manca (2016) and Conrad (1988).

9.3.6 *Formalism*

Infogenomics employs information theoretical analysis of well-characterized genomic features, such as indices, distributions, entropies, representations (and visualizations). The main formalisms for this approach, and in general for computational genomics, are:

- Algorithms on strings and related structures (suffix arrays, hash tables, dictionaries, multisets of strings) and efficient massive computation (Bon-

- nici and Manca, 2015a; Cicalese, 2016; Cicalese et al., 2011; Fici et al., 2006; Herold et al., 2008; Lothaire, 1997) ;
- Strings representation and reconstruction, dictionaries, factorization, localization, articulation and assembly, variability, similarity, networks (Bonnici and Manca, 2015b; Castellini et al., 2015; Franco and Milanese, 2013; Li et al., 2016; Manca, 2013; Manca, n.d.; Percus, 2007);
 - Discrete probability (probability distributions, random variables, purely random processes, Montecarlo methods) (Fofanov et al., 2008; Li et al., 2016; Sims et al., 2009; Zhang et al., 2007);
 - Information theoretic concepts (information sources, codes, entropy, entropic divergences, mutual information) (Fabris, 2002; Manca, 2015; Manca, 2016; Manca, n.d.; Sadovsky et al., 2008; Thomas and Cover, 1991);
 - Specialized software, that is a computational platform for massive computations of genomic informational indexes. For example, an open-source suite for the informational analysis of genomic sequences has been developed (Bonnici and Manca, 2015b) and proposed in (Bonnici and Manca, 2015a).

9.3.7 Applications

Recent experiments on minimal bacteria (Gibson et al., 2010; Gibson et al., 2014; Venter and al., 2016) are based on the search for genome sequences obtained by manipulating and reducing some real genomes. It has been proved that after removing some parts of the *M. mycoides* genome, the resulting organism (with 531 kilobase pairs, 473 genes), is able to survive and has a genome smaller than that of any autonomously replicating cell found in nature (very close to *M. genitalium*). In this manner a better understanding of biological basic functions is gained, which directly relates to the investigated genome (removing essential portions results in life disruption).

On the basis of this principle, Bonnici and Manca (2016) consider *M. genitalium* and remove some portions of its genome through a greedy exploration of the huge space of possibilities. At every step of their genome modifications (of many different types), they check the validity of their genomic laws, and the number of genes to be possibly eliminated (by keeping the holding of the laws) is comparable with the actual recent experiment in the lab (Venter and al., 2016). This is an example about the applicability of computational experiments, based on informational indexes and laws (possibly after suitable improvements to support and complement the development of genome synthesis and analysis), in the spirit of emergent trends in synthetic biology.

The InfoGenomics project aims at proving an innovative systematic approach for analysis of genomic diseases, and comparative analysis between “ill” and “healthy” genomes, and between species. Other areas commonly

face the challenge of analyzing rapidly increasing numbers of genomes, such as microbiology and virology. Identification of genomic markers takes to the development of individual pharmacogenetics as well as the so-called personal(ised) and precision genomics and medicine (Ginsburg and Willard, 2017).

Genomic rearrangements and structural variants are of fundamental importance in medicine, namely chromosomal rearrangements and structural variations do in chromotripsis. If we think of cell receptors, of antibody equipment, of viral loads, of genomic variations in microRNA, as bags of words to be designed or analysed, then we may see that computational genomics is an important part of future medicine.

9.3.8 Philosophy

The computer is a digitalisation of mathematics, as DNA is a digitalisation of life. Computational genomics aims/points at extracting principles of organisation and phenomena of regularity in genomic sequences, by means of algorithms, information theoretic concepts, and formal language notions. This perspective is a modelling attitude typical of physicists, with some important differences. The mathematics underlying this approach is mainly of discrete nature (Lothaire, 1997; Manca, 2013); the goal of the investigations is focused on the discovery of general principles of aggregation, and well-formedness of genomic structures, rather than on the determination of equations or invariants of temporal dynamics. Evolution is an essential characteristic of genomes, but there is no specific interest in the predictive analysis of genome evolution; rather, a crucial research perspective is how random processes and mechanisms of structural control in genomes can cooperate to ensure evolvability and programmability. Understanding the interplay of these two apparently conflicting aspects is one of the most difficult conundrums emerging in all the cases where new notions of calculus are considered, especially inspired by natural systems.

A classical computing agent is neutral with respect to the program that is called to execute, and remains unaffected by the computations that it performed in the past. Natural systems, however, especially in situations of great complexity, have an intrinsic relationship with their historical background. Nevertheless, many processes are realised with perfect uniformity, and the individual variability of some natural agents performing computation does not compromise the precision; rather it often enriches the ability to reformulate problems and find solutions (adaptivity, typical of biological systems). The computational and mathematical analysis of these competences, starting from genomes, which are a kind of “operating system” of cells, has a deep relevance not only for genomics and its applications, but also for suggesting new perspectives in the extension of classical paradigms of calculus.

9.3.9 *Scaling up*

Simply scaling up established bioinformatics pipelines will not be sufficient for leveraging the full potential of such rich genomic data sets. Instead, novel, qualitatively different computational methods and paradigms are needed.

9.4 Main achievements so far

The analysis of genomes by means of strings of length k occurring in the genomes, that is by means of genomic dictionaries of k -mers, has provided important insights into the basic mechanisms and design principles of genome structures (Bonnici and Manca, 2015b; Castellini et al., 2012; Chor et al., 2009; Franco, 2014; Li et al., 2016; Manca, 2013; Sims et al., 2009; Wang et al., 2015; Zheng et al., 2017; Zhou et al., 2008).

Castellini et al. (2012) individuates a relevance in the distinction of hapaxes (once-occurring words) versus repeats (multi-occurring words). Hapax/repeat ratio, minimal length of non-appearing factors, maximal repeat length, and repeat distributions, with respect to their lengths, are defined, and specific genome characters are investigated by means of them. In general, a methodology based on dictionaries has been discussed, where k -mer distributions are integrated with specific features depending on the internal organisation of genome structure.

Many studies have approached the investigation of genomes by means of algorithms, information theory and formal languages, and methods have been developed for genome wide analysis. Dictionaries of words occurring in genomes, distributions defined over genomes, and concepts related to word occurrences and frequencies, have been useful to characterise some genomic features relevant in biological contexts (Bonnici and Manca, 2015b; Castellini et al., 2015; Chor et al., 2009; Fofanov et al., 2008).

Bonnici and Manca (2016) and Manca (n.d.) propose the proper choice of the value k for applying information theoretic concepts that express intrinsic aspects of genomes. The value $k = \lg_2(n)$, where n is the genome length, allows the definition of some indexes based on information entropies, helpful to find some informational laws (characterizing a general informational structure of genomes) and a new informational genome complexity measure. Bonnici and Manca (2016) compute this by a generalised logistic map that balances entropic and anti-entropic components of genomes, which are related to their evolutionary dynamics. Figure 9.3 shows the localisation of some organisms according to such a numerical complexity.

Figure 9.4 shows a chart of the main informational indexes investigated by Bonnici and Manca (2016) over seventy different genomes. The two quantities EC and AC correspond to informational measures of evolvability (a random

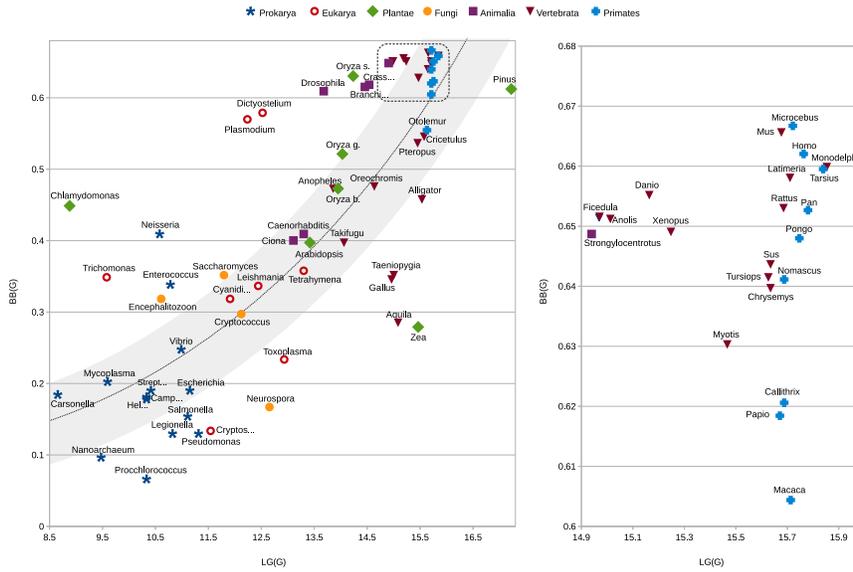


Fig. 9.3 Biobit computed for seventy genomes from different species. [Reproduced by courtesy of the authors of (Bonnici and Manca, 2016)]

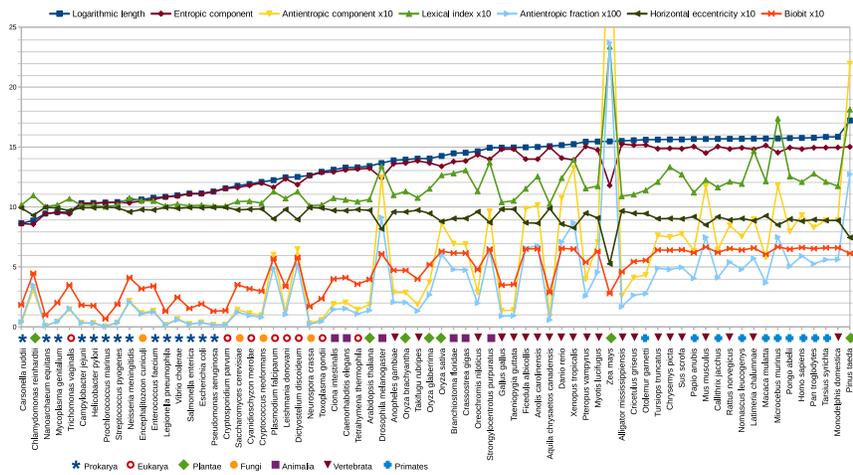


Fig. 9.4 A chart main informational indexes computed over seventy genomes. [Reproduced by courtesy of the authors of (Bonnici and Manca, 2016)]

component) and programmability (order conserved during evolution) (Conrad, 1988).

We refer to (Bonnici and Manca, 2016; Manca, n.d.) for a formal definition of EC, AC and of the related indexes of Figure 9.4. However, $EC(G)$ is the difference $E_{\lg_2 n}(G) - \lg_4 n$, where n is the length of the genome G and $E_{\lg_2 n}(G)$ is the logarithmic entropy of the genome computed for k -mers with $k = \lg_2 n$. The index $AC(G)$ is given by $\lg_2 n - E_{\lg_2 n}(G)$, which is always positive because $\lg_2 n$ is an upper bound of any empirical entropy of the genome, essentially coinciding with the maximum entropy reachable by a random genome of length n . Of course, $EC + AC = \lg_4 n$ (we omit the explicit mention of G) and this value is an index denoted by LG. Three other indexes are $EH = EC - AC$, $AF = AC/LG$, and LX that is the average multiplicity of the logarithmic k -mers of the genome. Finally, a more complex index BB is defined by means of Euler's beta function $\beta(AF, a, b)$ for two suitable parameters a, b ; see (Bonnici and Manca, 2016; Manca, n.d.) for the motivation of this definition. The interest of these indexes is given by some informational laws (Bonnici and Manca, 2016; Manca, n.d.) expressed by means of them. These laws have been tested over hundreds of genomes, including prokaryotes, algae, amoebae, fungi, plants, and animals of different types.

The specific software IGtools (Bonnici and Manca, 2015a) has been developed for extracting k -dictionaries, computing on them distributions and set-theoretic operations, and for evaluating empirical entropies and our informational indexes, for different and very large values of k -mers. IGtools is a suite (also open to developers) made on top of well-established data structures and algorithms (suffix trees and suffix arrays), adapted for real genomic sequences, and equipped by interactive graphical interfaces and CLI (for batch analyses). Figure 9.5 illustrates some computation for different genomic representations by the IGtools interface.

9.5 Current challenges

Current challenges in computational genomics undoubtedly include the development of new algorithms to process genomic data, and data structures able to efficiently handle with the huge mole of genomic variability. These should allow dynamic updates of stored information without rebuilding the entire data structure, including local modifications and dealing adequately with genomic variants. Especially owing to the huge size of generated sequencing data, extreme heterogeneity of data and complex interaction of different levels, we definitively need

A first conceptual challenge is the search of suitable representation and visualization of genomes, at different scale and with multidimensional perspectives, by providing easy frameworks within which to organize and think about genomic data.



Fig. 9.5 IGtools software interface. [Reproduced by courtesy of the authors of (Bonnici and Manca, 2015a)]

An interesting current challenge is the definition and computation/extraction of specific genomic dictionaries, giving both the key to compare different genomes and individual genomes of the same species (that would mean to be able to efficiently dominate the species variants).

Specific analysis in this respect could focus on computation of dictionary intersections, to systematically find evolutionarily conserved motifs among genomes (UCE) (Franco, 2014).

In conclusion, this field still expects considerable progress in both algorithmic and software engineering aspects, to face questions about efficient data structures, algorithms and statistical methods to perform complex and integrated bioinformatic analyses of genomes.

More related to the Infogenomics project, a current challenge is to apply informational indexes as biomarkers in specific pathological situations, and suitable distributions in order to discriminate genome regions and their internal organization.

9.6 What could be achieved in ten years

Evolution is the secret of life and the genomic perspective provides a more precise formulation of Darwinian theory of natural selection. However, even though this theory is a cornerstone in the interpretation of life phenomena, it remains a qualitative theory. A challenge of inestimable importance for a

deep comprehension of life is the discovery of quantitative principles regulating biological evolution. Computational genome analyses where specific informational concepts are massively investigated could unravel the internal logic of genome organization, where rigorous mechanisms and chance are mixed together to achieve the main features that are proper of living organisms.

It is not easy to tell now what are the detailed steps of this path, but surely such a kind of enterprise will shed a new light in the interplay between chance and computation and new computing paradigms will emerge that are inherently involved in the deepest mechanisms of natural evolution. One example of quantitative analysis related to this scenario is the “Fundamental Theorem of Natural Selection” proved by Ronald Fisher (Fisher, 1958). Informally, this theorem tells us that the evolutionary change of a population is directly related to the degree of gene variability within the population. This explains in rigorous terms why nature introduces mechanisms of genomic variability within species: the more the individuals present genomic variability, the more rapidly the species can evolve. This theorem is an example of mathematical analysis explaining biological evidence.

In the genomic era, mathematical rigour will be conjugated with genomic data and with the computational power of bioinformatics. We could hope for the achievement of important results that not only will explain to us some secrets of life, but will suggest to us new computational mechanisms with abilities typical of living organisms and of the evolution directing them. This would naturally have spin-offs in biotechnology, health science, synthetic biology, and all the life sciences. We can then expect to witness amazing development in the understanding of the nature of evolution in the mid-term future (Pan-Genomics Consortium, 2016).

References

- Almirantis, Y., P. Arndt, W. Li, and A. Provata (2014). “Editorial: Complexity in genomes”. In: *Comp. Biol. Chem.* 53, pp. 1–4.
- Annaluru, N., H. Muller, L. A. Mitchell, and et al. (2014). “Total synthesis of a functional designer eukaryotic chromosome”. In: *Science* 344.6186, p. 816.
- Bonnici, V. and V. Manca (2015a). “Infogenomics tools: a computational suite for informational analysis of genomes”. In: *Bioinform. Proteomics Rev.* 1.1, pp. 7–14.
- (2015b). “Recurrence distance distributions in computational genomics”. In: *Am. J. Bioinformatics and Computational Biology* 3.1, pp. 5–23.
- (2016). “Informational laws of genome structures”. In: *Nature Scientific Reports* 6.28840. DOI: 10.1038/srep28840.

- Castellini, A., G. Franco, and V. Manca (2012). “A dictionary based informational genome analysis”. In: *BMC Genomics* 13.1, p. 485. DOI: 10.1186/1471-2164-13-485.
- Castellini, A., G. Franco, V. Manca, R. Ortolani, and A. Vella (2014). “Towards an MP model for B lymphocytes maturation”. In: *Unconventional Computation and Natural Computation (UCNC)*. Vol. 8553. LNCS. Springer, pp. 80–92.
- Castellini, A., G. Franco, and A. Milanese (2015). “A genome analysis based on repeat sharing gene networks”. In: *Natural Computing* 14, p. 403. DOI: 10.1007/s11047-014-9437-6.
- Castellini, A., G. Franco, and R. Pagliarini (2011). “Data analysis pipeline from laboratory to MP models”. In: *Natural Computing* 10.1, pp. 55–76.
- Chor, B., D. Horn, N. Goldman, and et al. (2009). “Genomic DNA k -mer spectra: models and modalities”. In: *Genome Biology* 10, R108.
- Cicalese, F. (2016). *Fault-tolerant search algorithms: reliable computation with unreliable information*. Springer.
- Cicalese, F., P. Erdős, and Z. Lipták (2011). “Efficient reconstruction of RC-equivalent strings”. In: *IWOCA 2010*. Vol. 6460. LNCS. Springer, pp. 349–62.
- Conrad, M. (1988). *The price of programmability. The Universal Turing Machine A Half-Century Survey*. Oxford University Press.
- Consortium, International Human Genome Sequencing (2001). “Initial sequencing and analysis of the human genome”. In: *Nature* 409, pp. 860–921.
- Crochemore, M. and R. Verin (1999). “Zones of low entropy in genomic sequences”. In: *Computers & chemistry* 23, pp. 275–282.
- Deschavanne, P.J., A. Giron, J. Vilain, G. Fagot, and B. Fertil (1999). “Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences”. In: *Mol. Biol. Evol.* 16.10, pp. 1391–99.
- Dunham, I., A. Kundaje, S. Aldred, and et al. (2012). “(the ENCODE Project Consortium): An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489, pp. 57–74.
- Fabris, F. (2002). “Shannon information theory and molecular biology”. In: *J. Interdisc Math* 5, pp. 203–220.
- Fici, G., F. Mignosi, A. Restivo, and et al. (2006). “Word assembly through minimal forbidden words”. In: *Theoretical Computer Science* 359, pp. 214–230.
- Fisher, R.A. (1958). *The Genetical Theory of Natural Selection*. 2nd edn. Dover.
- Fofanov, Y., Y. Luo, C. Katili, and et al. (2008). “How independent are the appearances of n -mers in different genomes?” In: *Bioinformatics* 20.15, pp. 2421–28.
- Franco, G. (2005). “A polymerase based algorithm for SAT”. In: *ICTCS*. Vol. 3701. LNCS. Springer, pp. 237–50.

- Franco, G. (2014). “Perspectives in computational genome analysis”. In: *Discrete and Topological Models in Molecular Biology*. Ed. by N. Jonoska and M. Saito. Springer. Chap. 1, pp. 3–22.
- Franco, G., N. Jonoska, B. Osborn, and A. Plaas (2008). “Knee joint injury and repair modeled by membrane systems”. In: *BioSystems* 91.3, pp. 473–88.
- Franco, G. and V. Manca (2004). “A membrane system for the leukocyte selective recruitment”. In: *Membrane Computing*. Vol. 2933. LNCS. Springer, pp. 181–190.
- (2011a). “Algorithmic applications of XPCR”. In: *Natural Computing* 10.2, pp. 805–819.
 - (2011b). “On Synthesizing Replicating Metabolic Systems”. In: *ERCIM News 85 - Unconventional Computing Paradigms*. Ed. by Peter Kunz. European Research Consortium for Informatics and Mathematics. Chap. 21, pp. 21–22.
- Franco, G. and A. Milanese (2013). “An investigation on genomic repeats”. In: *Conference on Computability in Europe – CiE*. Vol. 7921. LNCS. Springer, pp. 149–160.
- Gatlin, L. (1966). “The information content of DNA”. In: *J. Theor Biol* 10.2, pp. 281–300.
- Giancarlo, R., D. Scaturro, and F. Utro (2009). “Textual data compression in computational biology: a synopsis”. In: *Bioinformatics* 25.13, pp. 1575–86.
- Gibson, D. G. et al. (2010). “Creation of a bacterial cell controlled by a chemically synthesized genome”. In: *Science* 329.5987, pp. 52–6.
- (2014). “Synthetic Biology: Construction of a Yeast Chromosome”. In: *Nature* 509, pp. 168–169.
- Ginsburg, G. S. and H. F. Willard, eds. (2017). *Genomic and Precision Medicine – Foundations, Translation, and Implementation*. 3rd edn. Elsevier.
- Hampikian, G. and T. Andersen (2007). “Absent sequences: nullomers and primes”. In: *Pacific Symposium on Biocomputing* 12, pp. 355–366.
- Herold, J., S. Kurtz, and R. Giegerich (2008). “Efficient computation of absent words in genomic sequences”. In: *BMC Bioinformatics* 9.5987, p. 167.
- Holland, J. H. (1998). *Emergence: from chaos to order*. Perseus books.
- Kong, S. G., H.-D. Chen W.-L. Fan, and et al. (2009). “Quantitative measure of randomness and order for complete genomes”. In: *Phys Rev E* 79.6, p. 061911. DOI: <https://doi.org/10.1103/PhysRevE.79.061911>.
- Li, Z., H. Cao, Y. Cui, and Y. Zhang (2016). “Extracting DNA words based on the sequence features: non-uniform distribution and integrity”. In: *Theoretical Biology and Medical Modelling* 13.2. DOI: 10.1186/s12976-016-0028-3.
- Lothaire, M. (1997). *Combinatorics on Words*. Cambridge University Press.
- Lynch, M. and J. S. Conery (2003). “The origins of genome complexity”. In: *Science* 302, pp. 1401–04.
- Manca, V. (2013). *Infobiotics – Information in Biotic Systems*. Springer. ISBN: 978-3-642-36222-4.

- (2015). “Information Theory in genome analysis”. In: *Conference on Membrane Computing (CMC)*. Vol. 9504. Lecture Notes in Computer Science. Berlin, Germany: Springer, pp. 3–18.
 - (2016). “Infogenomics: genomes as information sources”. In: *Emerging Trends in Applications and Infrastructures for Computational Biology, Bioinformatics, and Systems Biology*. Ed. by Q. N. Tran and H. R. Arabnia. Elsevier. Chap. 21, pp. 317–323.
 - (n.d.). “The principles of informational genomics,” in: *Theoretical Computer Science*. to appear.
- Manca, V., A. Castellini, G. Franco, L. Marchetti, and R. Pagliarini (2013). “Metabolic P systems: A discrete model for biological dynamics”. In: *Chinese Journal of Electronics* 22.4, pp. 717–723.
- Manca, V. and G. Franco (2008). “Computing by polymerase chain reaction”. In: *Mathematical Bioscience* 211.2, pp. 282–298.
- Mantegna, R.N. and et al. (1994). “Linguistic Features of Noncoding DNA Sequences”. In: *Physical Review Letters* 73.23, pp. 3169–72.
- Neph, S., J. Vierstra, A. Stergachis, and et al. (2012). “An expansive human regulatory lexicon encoded in transcription factor footprints”. In: *Nature* 489, pp. 83–90.
- Pan-Genomics Consortium, The computational (2016). “Computational pan-genomics: status, promises and challenges”. In: *Brief Bioinform.* DOI: 10.1093/bib/bbw089.
- Păun, G. (2016). “Looking for Computers in the Biological Cell. After Twenty Years”. In: *Advances in Unconventional Computing, volume 1: Theory*. Ed. by A. Adamatzky. Springer, pp. 805–853.
- Percus, J. K. (2007). *Mathematics of Genome Analysis*. Cambridge Studies in Mathematical Biology.
- Ratner, T., R. Piran, N. Jonoska, and E. Keinan (2013). “Biologically Relevant Molecular Transducer with Increased Computing Power and Iterative Abilities”. In: *Chemistry & Biology* 20.5, pp. 726–733.
- Rothmund, P. W. K, N. Papadakis, and E. Winfree (2004). “Algorithmic Self-Assembly of DNA Sierpinski Triangles”. In: *Plos Biology*.
- Sadovsky, M., J.A. Putintseva, and A. S. Shchepanovsky (2008). “Genes, information and sense: Complexity and knowledge retrieval”. In: *Theory in Biosciences* 127.2, pp. 69–78. DOI: 10.1007/s12064-008-0032-1.
- Searls, D. B. (2002). “The language of genes”. In: *Nature* 420, pp. 211–217.
- Sims, G. E., S.R. Jun, G. A. Wu, and S.H. Kim (2009). “Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions”. In: *PNAS* 106.8, pp. 2677–82.
- Spivakov, M., J. Akhtar, P. Kheradpour, and et al. (2012). “Analysis of variation at transcription factor binding sites in *Drosophila* and humans”. In: *Genome Biology* 13, R49.
- Thomas, A. and T. M. Cover (1991). *Elements of Information Theory*. John Wiley.

- Venter, C. and et al. (2016). “Design and synthesis of a minimal bacterial genome”. In: *Science* 351, p. 6280.
- Vinga, S. (2013). “Information theory applications for biological sequence analysis”. In: *Brief Bioinform* 15.3, pp. 376–89. DOI: 0.1093/bib/bbt068.
- Vinga, S. and J. Almeida (2003). “Alignment-free sequence comparison - a review”. In: *Bioinformatics* 19.4, pp. 513–523.
- (2007). “Local Renyi entropic profiles of DNA sequences”. In: *BMC Bioinformatics* 8, p. 393.
- Wang, D., J. Xu, and J. Yu (2015). “KGCAK: a k -mer based database for genome-wide phylogeny and complexity evaluation”. In: *Biol direct* 10.1, pp. 1–5.
- Y.Cai and et al. (2007). “A syntactic model to design and verify synthetic genetic constructs derived from standard biological parts”. In: *Briefings in Bioinformatics* 23.20, pp. 2760–67.
- Zhang, Z.D., A. Paccanaro, Y. Fu, and et al. (2007). “Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions”. In: *Genome Res.* 17.6, pp. 787–97.
- Zheng, Y., H. Li, Y. Wang, and et al. (2017). “Evolutionary mechanism and biological functions of 8-mers containing CG dinucleotide in yeast”. In: *Chromosome Research* E-pub ahead of print], pp. 1–17. DOI: 10.1007/s10577-017-9554-z.
- Zhou, F., V. Olman, and Y. Xu (2008). “Barcodes for genomes and applications”. In: *BMC Bioinformatics* 9, p. 546.