

UNIVERSITÀ DEGLI STUDI DI CATANIA

DIPARTIMENTO DI MATEMATICA E INFORMATICA

DOTTORATO DI RICERCA IN INFORMATICA

KNOWLEDGE BASES AND STOCHASTIC
ALGORITHMS FOR MINING BIOLOGICAL DATA:
APPLICATIONS ON A-TO-I RNA EDITING AND RNAi

GIOVANNI NIGITA

A dissertation submitted to the Department of Mathematics and Computer Science
and the committee on graduate studies of University of Catania, in fulfillment of the
requirements for the degree of Doctorate in Computer Science.

ADVISOR

Prof. Alfredo Pulvirenti

XXVI CYCLE

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Prof. Alfredo Pulvirenti) Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Prof. Vincenzo Cutello) Director of Graduate Studies

Approved for the University Committee on Graduate Studies.

*To my wife and
to God.*

*“The Lord has done great things for us,
and we are filled with joy.”*

Psalm 126:3.

New International Version (NIV).

Contents

Contents	v
List of Figures	xi
List of Tables	xv
List of Algorithms	xvi
Abstract	xviii
Preface	xx
I Introduction	1
1 Bioinformatics and the new era of DNA Sequencing	2
1.1 History of the Bioinformatics	3
1.2 Public Biological Databases	7
1.2.1 Genes	9
1.2.2 Sequences and protein structures	10

1.2.3	Vocabulary of the genes	11
1.2.4	Genomes	12
1.2.5	Gene expression	12
1.3	Comparison of Sequences	13
1.3.1	Statistical significance	15
1.3.2	Functional Motifs	16
1.4	Data Mining	17
1.4.1	The data mining process	18
1.4.2	Supervised Vs. Non-Supervised learning	19
1.4.3	Cluster Analysis	20
1.4.4	Hidden Markov models	22
1.4.5	Networks	23
1.4.6	The regulatory sequences of the genes	24
1.4.7	Prediction of functional binding sites	25
1.4.8	The model of networks	27
1.4.9	Where you can find the algorithms?	29
1.5	New era of DNA sequencing	30
1.5.1	Basic Methods	31
1.5.2	Second generation HT-NGS	33
1.5.3	Third generation HT-NGS	36
1.5.4	Application of sequencing technologies on human genome research	38
1.5.5	Alignment tools	39
2	Biostatistics algorithms and Markovian models	41

2.1	What is <i>Biostatistics</i> ?	41
2.2	Steps to get to reliable results	42
2.3	Application fields	43
2.4	Deterministic and Stochastic models	44
2.5	Markov models	45
2.6	Markov chains	46
2.7	Hidden Markov models	49
2.7.1	Formalization of Hidden Markov models	50
2.7.2	The <i>forward</i> algorithm	51
2.7.3	The <i>backward</i> algorithm	58
2.7.4	The <i>Viterbi</i> algorithm	62
2.7.5	The <i>Baum-Welch</i> algorithm	67
2.8	The Dirichlet distribution	77
2.8.1	Mixtures of Dirichlets	78
2.9	Estimators for estimation problems in discrete high-dimensional spaces	79
2.10	Centroid Estimation	80
2.11	Gamma Centroid Estimator	83
2.11.1	Evaluation measures defined using <i>TP</i> , <i>TN</i> , <i>FP</i> and <i>FN</i>	84
2.11.2	Formalization of γ -centroid estimators	85
3	MicroRNA Biogenesis and RNA Editing Phenomenon	88
3.1	microRNA Biogenesis	88
3.1.1	Organization of microRNA in Human Genome	90
3.1.2	MicroRNA Biogenesis	90

3.1.3	Post-Transcriptional Regulation Mediated by microRNAs	94
3.1.4	Regulation of microRNA Expression	95
3.1.5	Bioinformatics Prediction of microRNAs' Molecular Targets	97
3.1.6	Circulating microRNAs	99
3.2	RNA Editing Phenomenon	100
3.2.1	The birth of RNA Editing	100
3.2.2	RNA Editing in Different Organisms	102
3.2.3	Editing by Deamination	103
3.2.4	A-to-I RNA editing analysis	111

II Biological Databases 117

4	Biological databases and their analysis: <i>miRandola</i> , <i>miR-EdiTar</i> and <i>VIRGO</i>	118
4.1	<i>miRandola</i> : Extracellular Circulating MicroRNAs Database	120
4.1.1	Mirandola BackEnd	121
4.1.2	Sections of Mirandola	123
4.1.3	miRandola - miRò	129
4.2	miR-EdiTar: A database of predicted A-to-I edited miRNA target sites	134
4.2.1	The construction of miR-EdiTar	134
4.2.2	miR-EdiTar contents	136
4.2.3	Database implementation and web interface	138

4.2.4	Utility and discussion	138
4.3	VIRGO: Visualization of A-to-I RNA editing sites in genomic sequences	142
4.3.1	Datasets of RNA Editing sites	142
4.3.2	The creation of VIRGO	144
4.3.3	Construction and content	145
4.3.4	Utility and discussion	153
III	HMMs and their Application to miRNA Targeting	157
5	Profile HMM for microRNA Target and Design	158
5.1	Introduction to profile HMM for microRNA target	158
5.2	Forward and Backward Algorithms	161
5.2.1	Forward Algorithm	161
5.2.2	Backward Algorithm	164
5.2.3	LogSum Trick	165
5.3	Baum-Welch	165
5.4	Gamma-Centroid Decoding	169
5.4.1	microRNA Design	170
5.5	Stochastic backtrace procedure	170
5.6	Results	174

IV Motif Discovery in RNA Editing Phenomenon	177
6 Searching for motifs in RNA Editing	178
6.1 Description of the methodology	179
6.1.1 Preparation of the dataset	179
6.1.2 Searching for motifs in edited sequences	180
6.2 Preliminary results	181
V Conclusions	188
7 Conclusions	189
Bibliography	193

List of Figures

1.1	Molecular structure of the nucleotides	4
1.2	Some databases associated with <i>Entrez</i>	6
1.3	Scheme of the databases within <i>GenomeNet</i>	8
1.4	Allen Brain Atlas 3-D Brain Explorer Application	14
1.5	Electropherogram of a small portion of the DNA sequence . .	31
1.6	An example of the results of automated chain-termination DNA sequencing	32
1.7	Advanced technological features of three leading second gen- eration HT-NGS platforms	35
1.8	Advanced technological features of three leading third gener- ation HT-NGS platforms	37
1.9	Alignment of the reads in the genomic reference	39
2.1	A simple two-state Markov chain	47
2.2	(HMM) Scheme of transition from the state π_{i-1} to the state π_i	54
2.3	(HMM) Scheme of transition from the state π_i to the state π_{i+2}	58

2.4	(HMM) Scheme of transition from the state π_{i-2} to the state π_i	63
2.5	A binary matrix representation of a secondary structure of an RNA sequence	80
2.6	Multidimensional scaled distribution derived from 1,000 representative samples from Sfold	83
3.1	Model for biogenesis and activity of transcriptional repression of microRNAs	92
3.2	Representation of some of the possible mechanisms of action of the RISC complex induced by miRNA	95
3.3	Spontaneous oxidative deamination of cytosine	104
3.4	Possible effects caused by RNA editing	107
3.5	Example of C-to-U RNA editing in the Apo B gene of Human	107
3.6	Example of action of the ADAR in a double-strand region . .	108
3.7	Molecular structures of adenine and inosine	109
3.8	Inosine behavior, similar to the Guanosine one	109
3.9	Main effect of the A-to-I RNA editing	110
3.10	Transition from adenosine to inosine	112
3.11	Comparison of ADAR proteins	113
3.12	Tertiary structure of the <i>ADAR1</i> protein	114
4.1	Tables of miRNAs in <i>miRandola</i>	122
4.2	Tables of <i>mirna_converter</i> and <i>submission</i>	123
4.3	Homepage of <i>miRandola</i>	124
4.4	<i>Search</i> page in <i>miRandola</i>	125
4.5	Example of results page	126

4.6	<i>Advanced search page in miRandola</i>	127
4.7	<i>Tools page in miRandola</i>	128
4.8	<i>Link between miRandola and miRò</i>	130
4.9	<i>Page of miRò relative to diseases of has-miR-21</i>	130
4.10	<i>Page of miRò relative to functions of has-miR-21</i>	131
4.11	<i>Page of miRò relative to processes of has-miR-21</i>	132
4.12	<i>Page of miRò relative to tissues of has-miR-21</i>	133
4.13	<i>Examples of predicted miRNA binding sites potentially af-</i> <i>ected by A-to-I editing (I)</i>	139
4.14	<i>Examples of predicted miRNA binding sites potentially af-</i> <i>ected by A-to-I editing (II)</i>	139
4.15	<i>Examples of predicted miRNA binding sites potentially af-</i> <i>ected by A-to-I editing (III)</i>	140
4.16	<i>Examples of predicted miRNA binding sites potentially af-</i> <i>ected by A-to-I editing (IV)</i>	141
4.17	<i>Examples of predicted miRNA binding sites potentially af-</i> <i>ected by A-to-I editing (V)</i>	143
4.18	<i>Sequence of steps to identify putative A-to-I editing sites</i> . . .	147
4.19	<i>Clustering filter</i>	149
4.20	<i>Fourth Step of VIRGO</i>	150
4.21	<i>Example for the p-value computation</i>	152
4.22	<i>VIRGO usage example</i>	154
4.23	<i>Venn diagram concerning the number of editing sites in com-</i> <i>mon between VIRGO and DARNED</i>	155

5.1	Transition structure of the profile HMM for the microRNA targeting	159
5.2	Comparison PicTar and Profile HMM for miRNA targeting (10 <i>Baum-Welch</i> iterations)	175
5.3	Convergence of Baum-Welch algorithm	176
6.1	Upstream and downstream regions of editing site	180
6.2	Mapping of palindromic motifs on sample edited sequences . .	182
6.3	Mapping of non-palindromic motifs on sample edited sequences	183
6.4	Example of mapping of non-palindromic motifs on positive strand of the chromosome 1	185
6.5	Example of mapping of non-palindromic motifs on the negative strand of the chromosome 1	185
6.6	Example of mapping of palindromic motifs on the positive strand of the chromosome 1	186
6.7	An example of overlap between A-to-I editing sites and motifs	187

List of Tables

1.1	Comparison of next-generation sequencing methods	35
3.1	List of some of the most important predictors of miRNA targets	98
4.1	Overall Descriptive Statistics	137
5.1	States of the profile HMM for MicroRNA targeting	160
6.1	Examples of experimental validated editing sites in <i>5HT_{2C}</i> gene	181
6.2	Number of region of 4,000 nucleotides in each human chromo- some	184

List of Algorithms

1	Forward Algorithm	57
2	Backward Algorithm	61
3	Viterbi Algorithm: first part.	66
4	Viterbi Algorithm: final part.	67
5	Baum-Welch algorithm	77

Abstract

Until the second half of twenty century, the connection between Biology and Computer Science was not so strict and the data were usually collected on perishable materials such as paper and then stored up in filing cabinets.

This situation changed thanks to the Bioinformatics, a relatively novel field that aims to deal with biological problems by making use of computational approaches. This interdisciplinary science has two particular fields of action: on the one hand, the construction of biological databases in order to store in a rational way the huge amount of data, and, on the other hand, the development and application of algorithms also approximate for extracting predicting patterns from such kind of data.

This thesis will present novel results on both of the above aspects. It will introduce three new database called *miRandola*, *miReditar* and *VIRGO*, respectively. All of them have been developed as open sources and equipped with user-friendly web interfaces.

Then some results concerning the application of stochastic approaches on microRNA targeting and RNA A-to-I interference will be introduced.

Preface

In this thesis, I will present the results of the research carried out during the three-years of the PhD program in Computer Science at University of Catania. My research has been focused mainly on algorithms and systems on *Bioinformatics*.

The thesis consists of four parts: *Introduction*, *Biological Databases*, *HMMs and their Application to miRNA Targeting*, and *Motif Discovery in the A-to-I RNA Editing*.

Part I: Introduction

The first section introduces the basic knowledge needed to deal with the research topics treated through the thesis.

The first chapter presents the basic concepts related to bioinformatics and gives an in depth survey on all the research field. We start by describing the wealth of biological data then we move to the construction of biological databases and last we highlight all the data mining approaches that have been developed and are needed to extract predictive patterns from this very rich source of knowledge.

In the second chapter we will sketch key concepts on Biostatistics giving special emphasis on Hidden Markov Models. We describe the outcome of predicting algorithms as probability distribution looking always at the reliability of results measured in terms of biological soundness. Then I will introduce a new class of estimators called Centroid Estimators which are capable to overcome the limits of Maximum Likelihood for high dimensional space problems. The so called γ -centroid estimators will be then introduced stressing their capability to tune the ratio between positive predicted values against sensitivity.

The last chapter, that concludes the first part dedicated to the theoretical introduction to various arguments, consists of two sub-session. In the first one, we will analyze the biogenesis of microRNA(miRNAs), a large class of small non-coding RNAs of about 21-25 nucleotides, that negatively regulate the gene expression. Next we will introduce the RNA editing phenomenon, the process in which the nucleotide sequence of RNA is altered from the genomic code. The editing is related to the insertion/deletion of nucleotides, or the base modification. Its peculiarity is that the result of RNA editing is a change in the diversity and/or abundance of proteins expressed in the proteomes of organisms.

Part II: Biological Databases

The second part of my thesis is focused on the presentation and analysis of biological databases that I have been developed in collaboration with few colleagues. As a results of my effort, I will show miRandola, miR-EdiTar,

and VIRGO.

I will introduce miRandola, a database of extracellular/circulating miRNA. The database provides the users with a variety of information including the associated diseases, the samples, the methods used to isolate the miRNAs, and the description of the experiment. The information about the targets of miRNAs and their records are provided through links to miRò, “the miRNA knowledge base”. miRò integrates data from different sources to allow the identification of associations among genes, processes, functions, and diseases through validated and predicted targets of miRNAs. MiRandola is the first database about circulating miRNAs, where all the data are collected and maintained up-to-date in a MySQL database.

The article, submitted on August 2012, was accepted on October of the same year and published on Plos one.

Then, I will present miR-EdiTar, a database of predicted A-to-I edited miRNA binding sites. The database contains predicted miRNA binding sites that could be affected by A-to-I editing and that could become miRNA binding sites as a result of A-to-I editing. The importance of miR-EdiTar is that it contains a collection of predicted human miRNA binding sites in A-to-I edited 3' UTR sequences. The ones contained in the database can be either “current” sites, when they are predicted to be miRNA binding sites but that might be affected by A-to-I editing, or “novel” sites, when they are not predicted to be miRNA binding sites but they could become miRNA binding sites as a result of A-to-I editing. Furthermore, as in miRandola, miR-EdiTar is connected to miRò, a web environment that provides users with miRNA

functional annotations inferred through their validated and predicted targets.

The article, submitted on July 2012, was accepted on September of the same year and published on Bioinformatics.

Finally, I will focus on VIRGO, a web-based tool that maps A-to-G mismatches between genomic and EST sequences as candidate A-to-I editing sites. It is built on top of a knowledge-base integrating information of genes from UCSC, EST of NCBI, SNPs, DARNED, and Next Generations Sequencing data. The tool is equipped with a user-friendly interface allowing users to analyze genomic sequences in order to identify candidate A-to-I editing sites. VIRGO is a powerful tool allowing a systematic identification of putative A-to-I editing sites in genomic sequences. The integration of NGS data allows the computation of p-values and adjusted p-values to measure the mapped editing sites confidence. The whole knowledge base is available for download and its central purpose is to provide users with a periodically updated system storing high quality candidate editing sites. This will allow people to quickly and easily identify whether their genomic sequences are subject to A-to-I RNA Editing or not.

The article related to VIRGO was published on April 2012, on the journal BMC Bioinformatics.

Part III: HMMs and their Application to miRNA Targeting

In the fifth chapter, it is presented the application of profile HMMs to microRNA targeting. While the first applications of profile HMMs to the microRNA targeting problem. We will introduce a conditioned profile HMM properly designed to deal with this problem.

I will describe the different components that characterize our profile HMM, starting from the formalization of both *Forward* and *Backward* algorithms, subsequently integrated in the Baum-Welch algorithm for the parameter estimation. In this phase to guarantee the reliability of the results, the *MiRecoord* database has been used as training set. It contains experimental validated alignments between miRNAs and mRNAs. Finally, I will show the implementation of the decoding algorithm in order to find the most likely hidden states that determine the pairwise alignment between the two molecules. Few decoding approaches such as Viterbi, Stochastic Backtrace and γ -centroid will be introduced.

Part IV: Motif Discovery in RNA Editing Phenomenon

Despite the enormous efforts made in the last two decades, the real biological function of the RNA editing as well as the features of the substrates of the ADAR still remain unknown. I will present a preliminary methodological

workflow for the identification of predicting RNA-Editing structural motifs. The aim is to discover potential sequence signals appearing only in genomic regions subject to editing. It is well-known that A-to-I editing in human often occurs in repetitive regions, which makes motif discovery very challenging. In order to eliminate contamination of the motif by the Alu consensus we focus in non-repetitive flanking regions of the editing sites that could distinguish the A-to-I RNA editing.

In order to ensure the trustworthiness of the results, experimental validated (*EV*) editing sites have been collected by using the literature, and then divided into two categories: *true-positive* (***TP***) editing sites, and *false-positive* editing sites. For each edited gene containing *EV* editing sites, a sample of true-positive editing site has been selected. Among this sample set of editing sites, non-repetitive flanking regions, which consist of 2,000 nucleotides downstream and upstream of sample editing site, were extracted. To discovery some motifs in the selected edited regions we used the well known ***MEME*** (*Multiple EM for Motif Elicitation*) suite to find both 50 palindromic and 50 non-palindromic motifs.

Part I

Introduction

Chapter 1

Bioinformatics and the new era of DNA Sequencing

*It is like a voyage of discovery
into unknown lands, seeking not
for new territory but for new
knowledge. It should appeal to
those with a good sense of
adventure.*

Frederick Sanger

Nobel Price in Chemistry

(1958,1980)

THE word *Bioinformatics* comes from the juxtaposition of two words: “*bios*”, the Greek word for *life*, and “*informatics*”, the area of computer science. Thus, the main object of Bioinformatics is the management and the analysis of biomedical information through computers. Its main activities

relate to the construction and maintenance of a variety of databases; the development of algorithms for the alignment of sequences of DNA (*deoxyribonucleic acid*), RNA (*ribonucleic acid*) and proteins; the identification of genes and the assembling of genomes; the prediction of both, the structure and the interactions of nucleic acids and proteins; and, finally, the reconstruction and analysis of biological networks. The part of Bioinformatics that has a particular focus on *statistical/mathematical* assessing and model building, rather than on information management, is also called *Computational Biology*.

Information management is perhaps the primary activity of Bioinformatics, and it is certainly the most widely used and appreciated by the scientific community. The aim is first of all to collect the biological information in databases, then write it down, connecting it to a variety of additional information, and eventually to develop the services needed to access and use the data. Usually, the data and the analysis softwares may be used freely, except for the commercial databases (its consultation has a cost), and some of those industrial ones (its access is restricted). The best way to become familiar with the bioinformatic world and easily explore the huge amount of biological data is to enter the bioinformatics' portals, which host the databases and offer a variety of analytical tools and links to other sites.

1.1 History of the Bioinformatics

Bioinformatics was born in the late seventies, together with both the development of recombinant DNA technology and the publication of the first

sequences of nucleic acids. The DNA in a organism consists of very long sequences from an alphabet of four letters that correspond to the four possible nucleotides (see *Figure 1.1*): **A** for *Adenine*, **G** for *Guanine*, **C** for *Cytosine*, and **T** for *Thymine*.

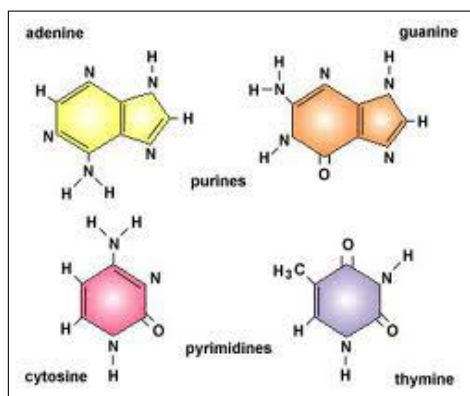


Figure 1.1: Molecular structure of the nucleotides.

These sequences are copied from generation to generation, and undergo changes within any population over the course of many generations, so this is the reason why random mutations arise and become fixed in the population. Therefore, it was immediately clear that it was impossible to decode encrypted messages in the sequences of DNA, RNA or proteins, through the implementation of descriptive algorithms of biological rules without the help of the computer technology.

Considering that it is difficult to provide an exact date that marks the beginning of the era of Bioinformatics, it might be more useful to outline the important events, distinguishing them, in particular, into two main areas, *biological databases* and *bio-computational methodologies*. Regarding the first point, although currently the *core* is formed from the databanks of DNA and

RNA, the first biological database hailed from the times of Margaret Dayhoff, an American physical chemist. In 1966, on the basis of Pauling's theories of molecular evolution, she made studies based on the analysis of protein sequences: the results were collected in an atlas on the basis of groups of homologous proteins [1]. In the seventies, it was modified into the electronic version of the database **NBRF** (*National Biomedical Research Foundation*). In the early eighties, the **EMBL** (*European Molecular Biology Laboratory*) in Heidelberg supported the construction of the EMBL datalibrary, a database of sequences of DNA and RNA [2]. The first *release* was in December 1981 and contained 519 entries relating to likewise nucleotide sequences, published and stored in an electronic document. In 1982, Walter Goad worked on the creation of a new database, from which it originated the **GenBank**, a storage similar to the European one, but produced in America [3]. In 1986 it was created the **DDBJ**, the Japanese database [4]. Later there was an international cooperation among EMBL datalibrary, GenBank and DDBJ.

It is clearly useful to have good systems for the selection and the mining of specific information collected in the biological databases. Among those systems, called *retrieval systems*, the most important are:

- The **Entrez** system, developed at the National Center for Biotechnology Information of the **NIH** (*National Institutes of Health*). In the *Figure 1.2*¹ there are the principal databases associated with Entrez:
- The **SRS** system, developed by Thure Etzold [5].

The cornerstones of bio-computational methodologies should be associ-

¹Source: <http://www.ncbi.nlm.nih.gov/Database/index.html>

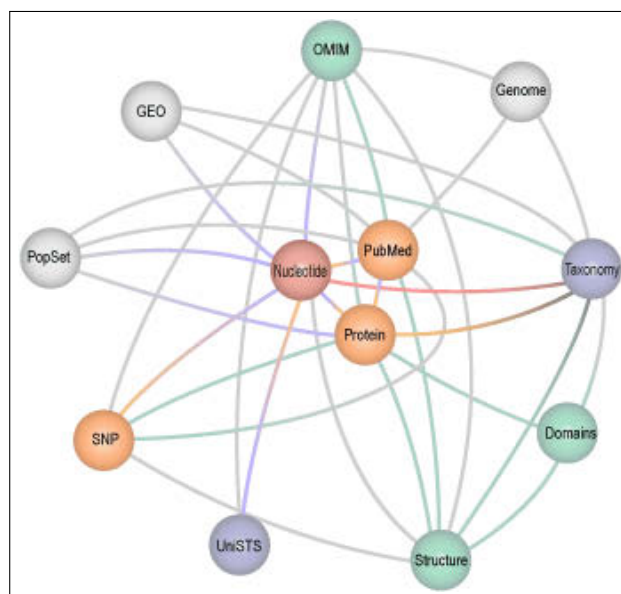


Figure 1.2: Some databases associated with *Entrez*.

ated with procedures for the comparison of biosequences to search for regions of similarity. In 1970 Needleman and Wunsch published the algorithm to search for the best global alignment between two sequences [6], and in the meantime Gibbs and McIntyre published a method based on dot-plot matrix, that allowed to display regions of similarity more or less strict and that was also used in many comparative analysis algorithms [7]. In 1981 Waterman and Smith published the algorithm for finding the best local alignment between the two sequences [8], while in 1983 Wilbur and Lipmann published an algorithm to search for the similarity inside the databases [9]. *FASTA* was published in 1985 [10] and *BLAST* in 1990 [11]. Simultaneously, numerous methods for the research of motifs and for the characterization of genomic sequences in protein coding regions were developed, such as, for example, the algorithms of Fickett and Gribskov [12, 13]. In the field of molecular evo-

lution, the publication in 1965 of Zuckerkandl and Pauling concerning the molecular clock hypothesis was a landmark [14]. This was followed by several studies by Dayhoff, the publication in 1966 [15] of the method of Maximum Parsimony analysis (then extended in 1977 by Walter Fitch [16] in the analysis of nucleotide sequences), the publication of the methods of Jukes and Cantor in 1969 [17], and in 1980 the calculation of phylogenetic distances along with the methods for the construction of phylogenetic trees by Kimura [18]. As regards the methods for structural predictions, noteworthy are the method of Zuker for the prediction of DNA structures [19, 20] and the method of Chou and Fasman for protein secondary structures [21, 22, 23].

1.2 Public Biological Databases

The *NCBI* (National Center for Biotechnology Information) was created in 1988 in the United States by the *National Library of Medicine* (*NLM* of the NIH, and maintains the largest bioinformatics portal in the world². Currently, it hosts more than thirty databases (bibliographies, genomes, sequences of nucleotides and amino acids, protein structures, and so on), that can be easily consult with the text search engine *Entrez*.

Founded in 1992, the *EBI* (European Bioinformatics Institute) is the main European center for research and bioinformatics services and maintains nucleic acids databases, proteins, macromolecular structures and biological pathways³.

GenomeNet, created in 1991, is a Japanese network for data and bio-

²For more information visit the website: <http://www.ncbi.nlm.nih.gov/>.

³For more information visit the website: <http://www.ebi.ac.uk/about/background>.

computational services. It contains the portal **KEGG** (*Kyoto Encyclopedia of Genes and Genomes*) which includes genes and proteins databases (*KEGG genes*), databases of chemical components (*KEGG ligand*), databases of molecular and biochemical reaction networks (*KEGG pathway*) [24]. In the Figure 1.3 is shown the scheme of the databases linked each other⁴:

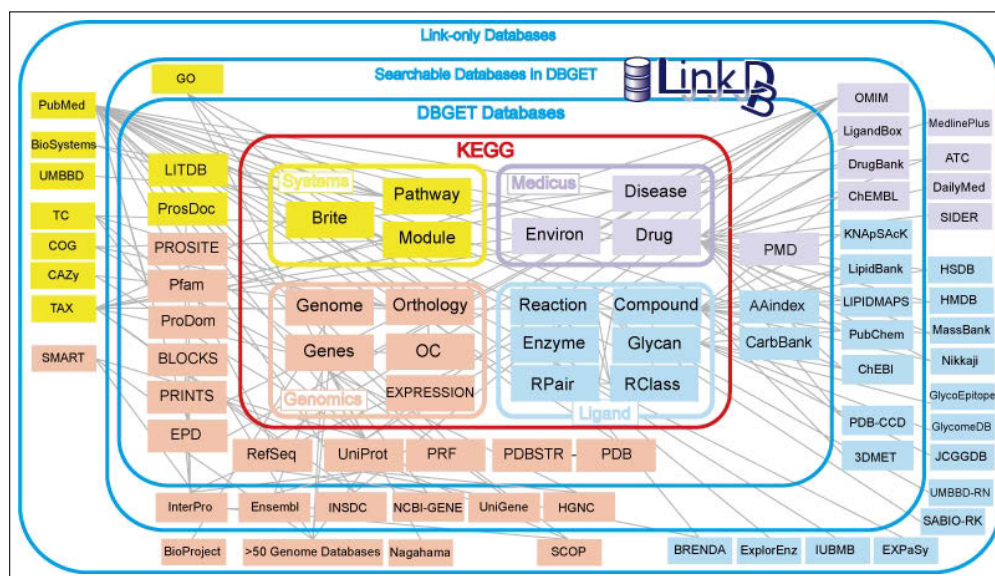


Figure 1.3: Scheme of the databases within *GenomeNet*.

ExPASy (*Expert Protein Analysis System*) proteomics, created in 1993 in Switzerland, offers a variety of tools for the analysis of data on proteins [25]. There are sequence databases, domains and protein families, proteomic data, models of protein structures and metabolic pathways.

The main public biological databases in the field of genomes are: **Ensembl**, in Great Britain, created in 1999 by the EBI and WTSI (*Wellcome Trust Sanger Institute*) [26, 27], and **UCSC Genome Browser**, in the

⁴Source: <http://www.genome.jp/linkdb/>

United States, created in 2000 by the University of California, Santa Cruz [28].

The most important laboratories and scientific journals have developed their own portals, providing information on particular biological aspects and tools to explore them. For example, *Genes to Cognition Online (G2C)*, created in 2009 by Cold Spring Harbor Laboratory, is a neuroscience portal, focusing not only on cognitive processes, but also on their related diseases and research approaches. The *Nature publishing group* contains “*Omics gateway*” for biology on a genomic level, while *The signaling gateway* is focused on signal transduction. The journal *Science* has developed “*Science signaling*”, focused on regulation and cell signaling.

1.2.1 Genes

As seen before, GenBank was one of the first nucleotide sequence database. It contains the nucleotide sequences obtained from people who deposit them there. It is part of the consortium *INSDC* (*International Nucleotide Sequence Database Collaboration*) along with the other two large databases: EMBL and DDBJ, where each archive contains over 100 million sequences.

In these databases the annotations are very limited, and there may be multiple entries for the same genes. If the genomic sequence encodes a protein, at first, the conceptual translation, called coding sequence (*CDS*), is shown, then, it receives a place in the protein database of NCBI.

The database *UniProt* (*Universal Protein resource*), managed together by the *EBI*, the *SIB* (*Swiss Institute of Bioinformatics* [29] and the *PIR*

(*Protein Information Resource*) [30], contains these sequences in the **TrEMBL section** (*Translated EMBL nucleotide sequence data library*). The NCBI's RefSeq database (*Reference Sequence*) is, instead, a collection of more restricted but also more accurate sequences. It chooses the best information available and sometimes its sequences are automatically imported from other databases. Moreover, RefSeq has about 10,000 organisms, while GenBank has sequences obtained by approximately 250,000 different organisms. When the authors publish new evidences, the **TPA** database (*Third Party Annotation*) gives them the possibility to annotate the sequences in the INSDC databases (*International Nucleotide Sequence Database Collaboration*) [31].

The **miRBase** database (*microRNA database*) is the central storage for the genomic of microRNAs, small non- coding RNA sequences of about 21 nucleotides that has a central role in the genes regulation [32, 33]. MicroRNAs control the translation of numerous mRNAs (messenger RNAs) into proteins and have a prominent part in the differentiating and cell proliferation, in the plasticity of both the synapses of nervous system and various diseases, including cancer. miRBase hosts more than 30,000 miRNAs sequences from 206 different species⁵, takes care their nomenclature and annotation, and provides programs for the prediction of the target mRNAs.

1.2.2 Sequences and protein structures

UniProtKB (*UniProt KnowledgeBase*) [34], consisting of two sections called *Swiss-Prot* and *TrEMBL*, is the most complete information source on se-

⁵At the time of the writing of the Ph.D. thesis, the *Release* of miRBase is 20: <ftp://mirbase.org/pub/mirbase/>.

quences and protein functions. Swiss-Prot is manually curated and has a very specific annotated; TrEMBL is automatically curated and contains the conceptual translation of the nucleic acid sequences that are in the databases, with little modifications.

The sequences stay provisionally in TrEMBL, waiting for a manual annotation to be transferred to SwissProt. **UniRef** (*UniProt Reference Clusters*) gathers together those sequences which are strictly connected in a single document, to speed the researches up [35]. **UniParc** (*UniProt archive*) contains, instead, both the protein sequences and all the available data).

PDB (*Protein Data Bank*), run by **RCSB** (*Research Collaboratory for Structural Bioinformatics*), hosts the structures of proteins and other biological macromolecules, and provides also a variety of resources for the study of their sequences, functions, and their possible pathological effect [36].

1.2.3 Vocabulary of the genes

Biologists use a great variety of terms to refer to genes and proteins and this variability is a restriction for an effective searching. The project **GO** (*Gene Ontology*) is the answer to the need of an unvarying terminology [37, 38]. Gene Ontology has developed an *ontology*, available through a database, that assigns three attributes to the product of each gene:

- a) the **biological process** in which it participates, such as signal transmission, pyrimidine metabolism, etc.;
- b) the **molecular function**, as, for example, catalytic activity, binding capacity, binding to a receptor;

- c) the *cellular component*, indicating its location inside the cell, such as endoplasmatic reticulum, nucleus, and ribosome.

A single gene product might have more than one molecular or biological function, and more than one location. The GO terminology facilitates the researches done by the various databases.

1.2.4 Genomes

The genomic data of individual organisms are annotated in various specialized databases, reached through the Ensembl and UCSC Genome Browser portals. Since the research focuses on the analysis of genomes, the graphical presentation of the sequences is very important. The genomic portals developed navigation tools, providing a quick view of any portion of genomes at any scale, with elaborate formatting options.

The aim of the *ENCODE* project (*Encyclopedia of DNA Elements*), launched in 2003, is to identify all the functional elements in the human genome sequence. It had an initial pilot phase, focused on a portion of the genome, and the results were published in June 2007⁶; after this, the goal was to compose the encyclopedia of the entire genome [39].

1.2.5 Gene expression

The huge amount of data obtained with the *high-throughput* technologies caused the need of databases that are able to retain them and make them accessible. In particular, the *DNA microarray* technology⁷ (commonly known

⁶The *ENCODE project consortium* 2007: <https://genome.ucsc.edu/encode/>.

⁷A *DNA microarray* is a collection of microscopic DNA spots attached to a solid surface.

as *DNA chip* or *biochip*)), or DNA *GeneChips*®⁸, was used to generate thousands of global gene expression profiles, obtained by measuring the amount of mRNA of a large number of genes in various conditions.

GEO (*Gene Expression Omnibus*) at *NCBI* [40, 41] and **ArrayExpress** at *EBI* [42] are the largest public deposits of such experiments. Both of them store the data in the standard format **MIAME** (*Minimum Information About a Microarray Experiment*) [43] and have exploration tools on line. They host not only many transcriptomics experiments, but also data about the *microRNA expression*, the *genomic hybridization*, **SNP** (*Single Nucleotide Polymorphism*), **ChIP** (*chromatin immunoprecipitation*) [44], and profiles of peptides.

The *Allen Brain Atlas* contains the three-dimensional map (see an example in *Figure 1.4*), on genomic scale, of the expression of thousands of genes in all the areas of the brain both of an adult human [45] and of an adult mouse and during its development, until the cellular level [46].

1.3 Comparison of Sequences

To get an idea of the possible meaning of new sequences, both nucleic acid or proteins, it might be very useful to compare them to other sequences with that have been already studied.

Aligning is the most effective method for comparing two sequences. This is done through algorithms that, first automatically analyze the correspondence between nucleotides or amino acids of different sequences, and then, at-

⁸To more information consult the website: *www.affymetrix.com*.

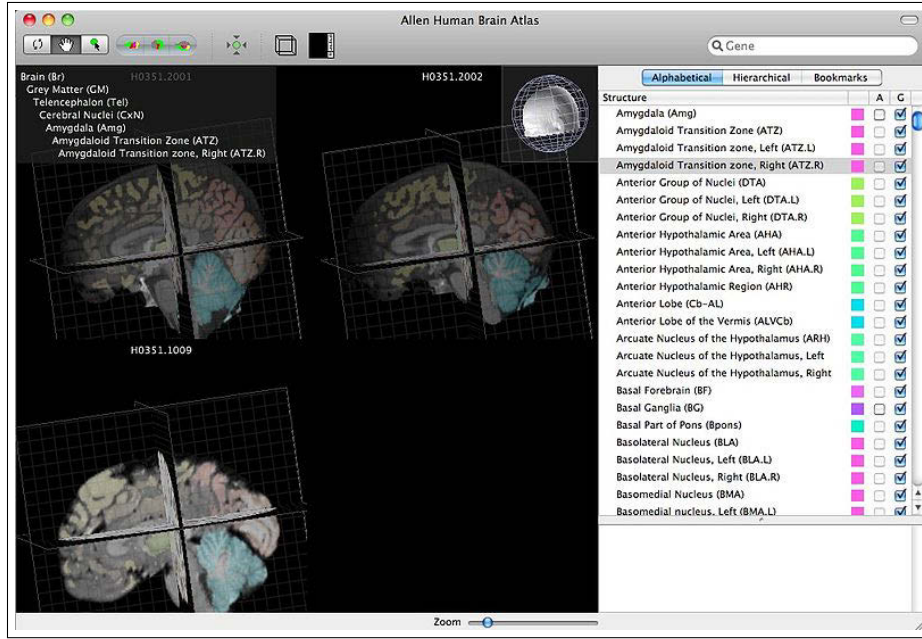


Figure 1.4: Allen Brain Atlas 3-D Brain Explorer Application.

tribute a score that reflects the degree of similarity. The softwares use graphical tools to view the alignments. These alignments can be global, if they include all the characters of each sequence (*Needleman–Wunsch* algorithm [6]), or local, if they include only the most similar regions (*Smith–Waterman* algorithm [8]).

The **BLAST** programs (*Basic Local Alignment Search Tool*), developed in 1990 at the NIH, are the most efficient tools for sequence comparison [11]. They offer a big variety of choices, depending on the type of sequence to examine, as also the purpose of the research and the database that queries. The BLAST programs, which highlight those regions of local alignment, divide both the sequence which queries the database (called “*query sequence*”) and the sequences contained in it in fragments called *words*; then, BLAST

starts searching for their matches. The initial search is made for a *word* of length W that has a score at least T^9 compared to the *query*. The *words* that are identified, called *hits*, are stretched in both directions, in an attempt to create an array with a score greater than a threshold value S .

1.3.1 Statistical significance

Considering that the databases contain a large quantity of sequences, there may be completely random cases of good similarity. By assigning to each alignment a statistical value, as the ***P-value*** or the ***E-value*** (*Expectation value*), it is possible to valuate how *significant* an alignment is. The parameter P is a number between 0 and 1 that indicates the probability that the alignment between the query sequence and a particular sequence of the database is the result of the case. A *P-value* of 0,05 indicates that there is a 5% chance that the alignment is meaningless. The *E-value* indicates the number of alignments having equal or better scores than the one observed, that might occur by chance.

Therefore, much smaller is P or E , the more significant is the alignment. P and E are related by $E = P \times S$, where S here is the size of the database.

Moreover, the *P-value* and *E-value* are not always enough to give a biological meaning to an alignment, and it is often an appropriate critical evaluation criteria with common sense. The low complexity regions, such as those with repeated sequences are a frequent problem, because the similarity

⁹The parameter T determines the *computational speed* and the *sensitivity* of the search, in particular more the parameter is high, then higher the similarity request, more research is fast, but increase the risk that you leave out similarity that are not *strong*, which may have biological significance.

based on that type of sequences is unreliable.

1.3.2 Functional Motifs

The main function of a database is to identify, among a huge number of sequences of genes and proteins, some characteristics that indicate a specific function. It was confirmed that genes or proteins that play a similar function have a similarity in some regions of their sequence. Thus, genes and proteins belonging to the same functional family should contain in their sequence a recurring motif that characterizes the family and distinguishes them from the others. One of the most useful things that can be obtained from the comparison of sequences is the identification of short areas that indicate a particular structure or function.

Thanks to their biological significance, these regions show high conservation in their sequences. The presence of these “*signatures*” is extremely useful to assign a new sequence to a specific family of genes or proteins, to be able to make assumptions about its function. In the computer language such signatures are called *motifs*, and can be described as short text strings, called *patterns*, or as *numeric arrays*. The patterns are located in a small region of high homology, while the profiles also consider long sequences. Patterns and profiles can be found in databases as **PROSITE** [47] or **JASPAR** [48].

PROSITE is a database of proteins’ domains, families and functional sites, integrated with computer tools to identify sequence motifs . It contains specific signatures for more than 1,500 families or protein domains and extensive documentation on their structure and function. It is possible

to quickly identify to which known protein family a given protein sequence belongs to, thanks to those sequence motifs, which represent transcription factors of DNA binding preferential sites, taken from the scientific literature.

The database JASPAR deals with the promoters, the DNA sequences regulating the expression of genes . They are located immediately before the gene transcription's starting point and tie a variety of regulatory proteins, called transcription factors. The particular combination of factors related to the promoter determines whether the gene will be turned on, or off. JASPAR 174 contains sequence motifs, which represent transcription factors of DNA binding preferential sites, taken from the scientific literature. They can be used for scanning genomic sequences.

1.4 Data Mining

The data represents a resource of great intrinsic wealth, to data only partially exploited. Technological progress has made the digitalization and storage of huge amounts of heterogeneous data possible. This exponential growth has given rise systems able to analyze in a *semi-automatic* way these data in order to *classify, synthesize, extrapolate trends, identify anomalies*, and so on. *Data mining*, also known as **KDD** (the analysis step of the *Knowledge Discovery in the Databases* process), is one of the most interesting areas of research in the community of databases. It consists of an automatic extraction of patterns representing knowledge implicitly present in large databank systems (*databases, data warehouses, web, etcetera*). This area collects scientific contributions by researchers from different fields, such as statistics,

artificial intelligence, machine learning and visualization.

The data mining finds wide application in Bioinformatics, for example in the classification and analysis of biological data as, for example, sequences, networks and expression profiles. In particular, the frontier of research in Bioinformatics disposing of technologies such as *deep sequencing* (eg *RNASeq*) in the coming years will be the core of a strong innovation that will focus on development of new algorithms and methods of learning.

1.4.1 The data mining process

The main purpose of data mining is not to give an explanation of a phenomenon but to discover the knowledge and to predict. This means to identify hidden structures in the data that make it possible to extract useful information and to make accurate predictions on the evolution of a phenomenon. This process typically follows several steps, and according to *CRISP-DM* (*Cross Industry Standard Process for Data Mining*) it can be define in six phases:

1. ***Problem definition***: the first phase consists in the understanding of the area problem, where the goal of the project is translated into a data mining problem definition. In this step data mining tools are not required.
2. ***Data exploration***: once finished the first phase, the data are collected, described and explored. Quality problems of the data are identified, and data analysis too are used in order to explore the data.

3. **Data preparation:** the data model are built for the modelling process. In particular, the data are cleansed and formatted so as to be able to apply some mining functions.
4. **Modelling:** various mining functions are selected and applied based on the type of data mining problem. The mining experts can repeated this phase several times, changing parameters from time to time until optimal values are achieved.
5. **Evaluation:** the model is evaluated. If mining experts valuate that the model does not satisfy their expectations, the modelling step is applied again and they rebuild a new model, by changing its parameters, in order to reach optimum values. It's clear that the modelling step and evaluation step are closely related.
6. **Deployment:** when the data mining results are obtained, they will be exported in a specif format or into database tables so as to be easily consulted.

1.4.2 Supervisioned Vs. Non-Supervised learning

It is important to distinguish between learning with or without supervision: in the second case, in fact, no *a priori* assumption on how to split the data is done, and the learning process occurs without specific knowledge of the content analyzed. In Bioinformatics, methods of unsupervised learning are used especially in the *data exploration* phase, to find in the data some not easily visible data structure.

Unsupervised learning allows to perform operations of data segmentation, that is, to identify instances exhibiting an inner regularity which is able to characterize them. Then, it can be used to partition the data in *clusters* (subsets) so that instances in each subset share some common features.

The supervised learning is usually fast and accurate and it can be applied to cases with a particular classification already known in a *training set*. The aim is to create a model that predicts this classification in new data.

1.4.3 Cluster Analysis

The expression “*cluster analysis*” indicates a number of unsupervised learning algorithms that distribute objects into *groups* according to *similarity* criteria. The number of groups may be determined automatically or chosen by the user. The similarity between objects is mathematically evaluated through a *distance measure*: less the objects are far from each other, the more similar and more easily part of the same group they will be. There are several measures of distance, such as the *Euclidean distance*, which is simply the geometric distance in the multidimensional space of the data, or *Pearson’s correlation coefficient* (technically called the *Pearson Product Moment Correlation* or *PPMC*) that shows the linear relationship between two variables.

Cluster analysis is applicable to a large variety of problems. In Bioinformatics, for example, it is very common for the examination of gene expression data on a large scale, obtained through microarrays. The most natural way to organize this data is to put in the same group those genes that have a

similar expression, because likely they will have good chance to participate in the same biological process. This does not imply that there is a direct interaction between the genes, since they can be co-expressed genes, separated by one or more intermediaries. It is better to use the correlation coefficient as the measure of distance between a pair of genes, which is more in line with the intuitive idea of co-expressed genes.

Hierarchical clustering for gene expression data

The most common approach for gene expression data is the hierarchical grouping, or tree grouping, which represents the relations between the genes by a sort of *tree*, where the proximity of the branches reflects the degree of similarity. First, the algorithms of hierarchical grouping consider each object as unconnected, and then, step by step, the objects are closer grouped together. Thus, gradually, larger and larger groups of objects more dissimilar are connected. Finally, all objects are linked together in a large tree (*dendrogram*). The number of groups, or clusters, is determined automatically by the algorithm.

K-means clustering in Bioinformatics

Sometimes it is more convenient to choose the number of groups to split up the objects in our choice and then, using the *K-means* technique, divide N objects in a k a number of groups, with $k < N$, on the basis of their attributes, and so that they are as distinct as possible.

The attributes of the objects are represented as vectors and each cluster is identified by a midpoint called *centroid*. The algorithm follows an iterative

procedure. Initially, the algorithm creates k groups, whose elements are randomly selected or in an empirical manner, and calculates the centroid of each group. Then it moves the objects between the groups with the aim to minimize the variability within them and to maximize it between one group and another. Thus, it creates a new subdivision, associating each point to the group whose centroid is closest to, then the algorithm recalculates the centroids for the new groups and so on, until it finds a stable solution.

The gene expression profiles of people with a particular disease may have their own *signature* which can be a powerful tool for accurate diagnosis and prognosis, as well as the choice of the best cure. However, it is necessary to improve the bioinformatics methods to recognize the *signatures* in a secure manner.

1.4.4 Hidden Markov models

A common way to recognize patterns is to use probabilistic models such as **HMM** (*Hidden Markov Models*) [49]. In Bioinformatics, such models are widely used to identify homologies or to predict both the coding regions in the genome sequence and the mode of folding proteins. They take their name from the *Markov chain*, a sequence of states in which the transition from a present state to a future one occurs with a probability that depends only, or nearly so, from the present state, and not from the process or its past. This means that the present state of the system allows to predict the future behavior, while the previous history has little influence.

The most common example is the flip of a coin: here the result is heads

or tails with equal probability, independently of previous flips. The theory of *Markov processes* are often used to predict the succession of weather, to estimate macroeconomic dynamic, or to give an Internet rank. For example, Google uses PageRank, an algorithm that assigns a numerical score to the web pages, in order to measure their relative importance. The algorithm is based especially on the concept of popularity, that is the frequency a page is visited. HMMs are more complicated, because in this case the states of the system we want to analyze are not directly visible, but it is possible to observe only the events related with a certain probability.

The main aim of HMM is not only to determine hidden states from the observable events but also to identify the parameters of the model, that is the transition probabilities from one state to the next. Once the model is drawn, this may be used for further analysis and predictions on new events.

1.4.5 Networks

Progressively it becomes more and more necessary to integrate the biomolecular information to the higher level of the biological function of cells, tissues and whole organisms. Complex networks of biological elements interacting with each other (such as genes, metabolites, proteins) regulate the operation of living cells. These huge networks are organized into subnetworks and each of them takes care of a particular aspect of the function of cell, such as *cell cycle*, the *signal transmission* and so on. These subnets, consisting of many elements interacting together to implement an activity of the cell, are called *functional modules*.

The reconstruction of the architecture of networks and modules, which in the past required a long collection of many experimental data, is now faster. This is possible thanks to the technologies that allow to quickly detect the expression of the genes and proteins and their changes in various conditions on a genomic range: *microarrays*, the *deep sequencing*, *proteomic technologies*, *SNP* (Single Nucleotide Polymorphism) analysis, *comparative analysis of genomes*, the *ChIP on chip*, the *epigenomics*. Bioinformatics and Systems Biology focus their research on the inference of the structure and the control mechanisms for various types of networks.

Commonly, the networks are inferred in a supervised way starting from very safe interactions, derived from data on proteins or gene expression. The networks are represented as *graphs*, in which the *nodes* are the genes or proteins and the *arcs* are the interactions. *Cytoscape* [50], *CellDesigner* [51] and *MIM* (*Molecular Interaction Maps*) [52] are only a few example of the tools available for drawing and view the diagrams.

Computational models are fundamental to understand the way a network regulates a biological function. A good network model allows to simulate cell behavior under a variety of stimuli and to facilitate the design of new drugs.

1.4.6 The regulatory sequences of the genes

It is possible to combine the information inside the genes in various ways to implement different activities. After studying the genomes, it has been clear that their length and the number of gene they contain are much less important than the way the genes are regulated and combined. For example, the

grain has a genome larger than the one of a man and contains approximately the same number of genes, but it is impossible to say that it is more evolved.

Deciphering the control mechanisms of the gene expression is essential to analyze the behavior of networks. This, together with the large amount of gene expression data on large scale, has motivated the research for methods for the analysis of DNA sequences that regulate the expression of genes. The algorithms for the identification of regulatory regions of genes have been unreliable due to a too high number of false positives that tends to make the vast majority of predictions futile. That's the reason why the researchers are developing new methods to make the predictions faster and less uncertain, although it remains necessary to verify in the laboratory.

The DNA sequence controlling the gene expression, the *promoter*, is located near the gene, usually at the extremity 5'. The promoter binds together a series of regulative proteins that allows, or not, the access to the gene of the machineries that produces the mRNA.

1.4.7 Prediction of functional binding sites

After the identification of a promoter, it is important to understand which *transcription factors* (**TF**) bind to it, to regulate it. Usually, the transcription factors prefer specific sequences, which can be captured in the form of sequence motifs. The sequence motifs, then, may help to predict the possible binding sites for a given transcription factor in the genomic sequences.

The motifs of binding to transcription factors are collected in the **TRANS-FAC** [53] and **JASPAR** [48] databases, which also give the chance to iden-

tify the sites capable of binding transcription factors in any DNA sequence. The only problem is that the identified sites are able to bind transcription factors in vitro, but it is not guaranteed that it happens within the cell. The reason is that the structure of chromatin nearby the promoter strongly influences the ability of a transcription factor to bind its target sequence. Moreover, in models based on motifs of sequence, it is usually assumed that the binding of a transcription factor to a promoter is not influenced by adjacent sequences and the proximity of other proteins. But this is wrong, because the combinatorial interactions between various factors linked to multiple sites are essential for the gene expression. The result is that only a small part of the binding sites in vitro are also in vivo, so it is impossible for JASPAR and TRANSFAC to distinguish those sites with a functional role from those without. The relationship between false and true positives can be so high that it can frustrate any assumption.

To improve the predictions of the binding sites, the sequence motifs can be combined with phylogenetic footprints, as in the algorithm *Consite* [54]. Some algorithms capture also the cooperative interactions among transcription factors, binding to groups of sites within a promoter. These methods allow to reduce the number of false positives of an order of magnitude, that, however, is still not enough to improve the performance of the prediction. The creation of bioinformatics algorithms is important to better represent the mechanisms that regulate the transcription of genes. For example, it is possible to identify regions containing significative combinations of transcription factors, biologically related. There are various methods, such as *MSCAN* [55], *MCAST* (*Motif Cluster Alignment and Search Tool*) [56]

and *ModuleScanner* [57], which use a variety of statistics and data mining techniques, as the *Bayesian networks*. The task of identifying precisely the functional binding sites is facilitated by the use of technologies as *ChIP on chip* and *ChIPSeq*, which reveal the genomic sites actually linked to a factor of transcription within a cell.

The problem of the abundant presence of false positives also affects the numerous programs, trying to predict *microRNA target genes*, such as *TargetScan* [58], *Diana-microT* [59], *PicTar* [60] and others. These programs seek regions at the untranslated 3' of mRNA with a complementarity sequence with miRNAs. There are various sequences potentially capable of binding a single microRNA, considering, not only that generally the sequence complementarity between microRNA and mRNA target is not absolute, but also the brevity of the sequence of microRNAs. The programs use empirical rules to give a score to the various alignments, and use of phylogenetic prints and also the presence or absence of more binding sites within the mRNA. However, even if they can provide useful guidance, their results are not satisfactory.

1.4.8 The model of networks

The goal of genetics is to explain the relationship between genes and the behavior of a cell or an organism. This connection is based on complex regulatory networks, having a modular structure. This means that the network is formed by a set of sub-networks of various forms, and each of them has a function which is distinct but also simpler than the one of the network as a

whole. The modular structure facilitates the modeling because it allows to consider separately the individual modules, which, although quite complex, are less complex than the global network. The events that take place in these networks can be thought of as logical elementary functions, bringing the cell from one state to another.

The networks modeling reproduces on a computer the implementation of these logic functions. The abundance of gene expression data, now available, makes it possible to decode complex gene networks through the reverse engineering. It is used to identify the interactions between the genes, and thus discover the way a biological network works, through the analysis of experimental data connected to its components (usually they are the data of expression of the mRNA).

Network and model databases

The analysis of the structure and the behavior of the genetic networks requires not only new theories and algorithms, but also databases capable of storing and displaying information interactions. ***COXPRESdb*** (*CO-exPRESSed gene database*) provides reports of coexpressed genes in mammals, obtained from expression profiles measured by microarrays [61]. It allows to create not only networks of coexpressed genes in the same tissue, but also genes with the same *GO annotation* and genes expressed in a similar way in humans and mice. The networks are displayed using the coefficients of correlation as criterion of proximity and are shown through *Google Maps API*.

The protein interaction data are collected in various databases, including

MINT (*The molecular InterAction database*) [62]. **GeneNetwork**¹⁰ collects gene interactions known in humans, obtained from both other databases, such as **HPRD**¹¹ (*Human Protein Reference Database*) [63], **BIND** (the *Biomolecular Interaction Network Database*) [64], *Reactome* [65], *KEGG* [24], and *GO*) [37, 38] and new experimental data. In addition, it generates predictions about possible new interactions.

JWS Online (*Java Web Simulation*) [66], **BioModels** [67] and **DO-QCS** (*Database Of Quantitative Cellular Signaling*) [68] are examples of databases of the models published on scientific journals.

The cellular signaling pathways not only have a great scientific interest, but are also considered a possible therapeutic target for many diseases.

1.4.9 Where you can find the algorithms?

The statistical and mathematical techniques useful for the exploration of biological data can be found using various commercial packages. Among them, **MATLAB**®¹² (*MATrix LABoratory*) has a section dedicated to Bioinformatics, and allows the users to analyze and view genomic and proteomic data, and to build models of biological systems.

Another possibility for a good analysis of biological data is to use *open sources*, such as **R**. It is an open software environment for free access, where it is possible to implement a variety of statistical and graphical techniques, such as linear and nonlinear modeling, statistical tests, time series analysis, classification and clustering algorithms, and so on. The basic version

¹⁰Website: <http://www.genenetwork.org/>

¹¹Website: <http://www.hprd.org>

¹²For more information visit the website: <http://www.mathworks.it/>

can be easily expanded through specialized software that can be obtained through **CRAN**¹³ (*Comprehensive R Archive Network*). **Bioconductor**¹⁴ is a project associated with R and focused on Bioinformatics applications, which provides tools for the analysis of genomic data.

1.5 New era of DNA sequencing

DNA sequencing is a method that used to line up the nucleotides that make up the DNA molecule, so it can be properly read and analyzed. The DNA sequence contains all the inherited genetic information that is the basis for the development of all living organisms. Within this sequence genes of every living organism are encoded, as well as instructions on how to express them in time and in space (regulation of gene expression). Determining the sequence is therefore useful in the research of why and how organisms live. There are portions of DNA whose functions we already know. Once sequenced, the DNA fragment analysis can compare the sequences already stored in the database cataloged online, even if a substantial part of the human genome remains unknown.

The knowledge of the genome is therefore useful in any field of biology and the advent of methods for DNA sequencing has significantly accelerated the research. In medicine, for example, the sequencing is used to identify and diagnose genetic diseases and to develop new treatments. In a similar manner, the genome of the pathogenic agents may lead to the development of medicines against contagious diseases. The speed of the process of se-

¹³For more information visit the website: <http://cran.r-project.org/>

¹⁴Website: <http://www.bioconductor.org/>

quencing today is a great help to the large-scale sequencing of the human genome. Similarly, the sequencing of the genome of various plant and animal organisms, as well as of many microorganisms has been completed .

In these last years the DNA sequencing methods are constantly evolving. On the one hand the researchers want to improve the speed of execution, trying to lower the cost, on the other hand they attempt to get more accuracy. The determination of DNA sequences is also useful in different application fields and DNA sequencing may be used to determine the sequence of individual genes, larger genetic regions, full chromosome or entire genomes.

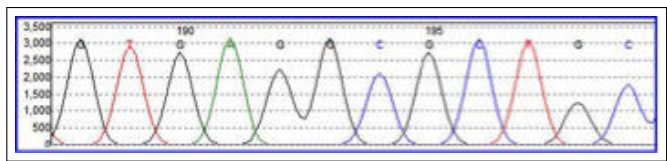


Figure 1.5: Electropherogram of a small portion of the DNA sequence.

1.5.1 Basic Methods

Several strategies have been devised to obtain the nucleotide sequence of the DNA. The first methods, including one developed by Allan Maxam and Walter Gilbert in 1973 [69], were quite complicated. A turning point came in 1975 with the first publication of a enzymatic strategy still widespread, developed by Frederick Sanger and coworkers (the so-called chain terminator methods, or the Sanger method, as seen in *Figure 1.6*) [70, 71]. This strategy soon became the method of choice, thanks to its relative ease and consistency. The Sanger method used fewer toxic chemicals and lower amounts of radioactivity

and for this reason it was the most widely method used in the first generation of DNA sequencers.

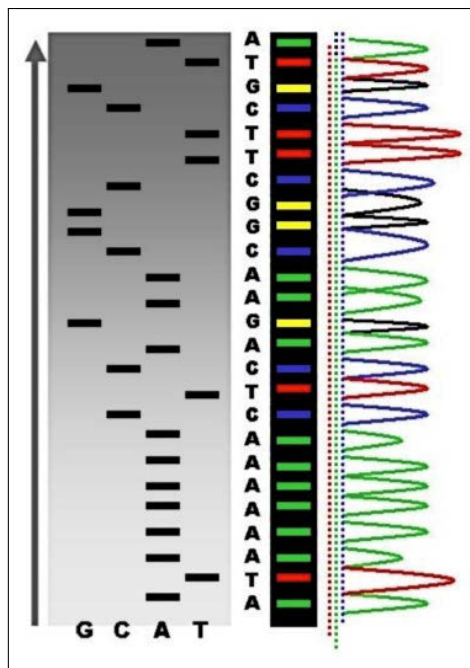


Figure 1.6: An example of the results of automated chain- termination DNA sequencing.

Another strategy, initially very popular, was developed by Maxam and Gilbert in 1977 and is known under the name of “*the method of Maxam and Gilbert*” [72]. This method allowed purified samples of double-stranded DNA to be used without cloning, even if the use of radioactive labeling and its technical complexity did not allow a real launch, unlike the Sanger method.

Later in 1980, Walter Gilbert and Frederick Sanger shared half of the chemistry prize “*for their contributions concerning the determination of base sequences in nucleic acids*¹⁵”.

¹⁵The Noble Prize in Chemistry 1980:
http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1980/.

1.5.2 Second generation HT-NGS

More recently, due to the increasing demand for low cost sequencing, new methods have been developed. They are characterized by the ability to sequence many DNA fragments simultaneously (although with lower efficiency in terms of number of bases sequenced per fragment) opening a new era of sequencing. These methods, that parallelize the sequencing process, are able to produce hundreds of millions of bases of raw sequence (Roche2) and they can generate up to billions of bases in a single run (Illumina, SOLiD). Among the most important method belonging to the second generation of HT-NGS there are [73, 74]:

- *Massively parallel signature sequencing (MPSS)*: this method was developed in the 1990s and it was the first of the next-generation sequencing technologies, but it was so complex to use.
- *Polony sequencing*: developed in the laboratory of George M. Church at Harvard, it was among the first next-generation sequencing system. This method was used to sequence a full genome in 2005 [75].
- *454 pyrosequencing*: a parallelized version of pyrosequencing was developed by 454 Life Sciences [76].
- *Illumina (Solexa) sequencing*: in this method DNA molecules and primers are first attached on a slide and amplified with polymerase so that local clonal DNA colonies, later coined "DNA clusters", are formed. To determine the sequence, four types of reversible terminator bases (RT-bases) are added and non-incorporated nucleotides are

washed away. A camera takes images of the fluorescently labeled nucleotides, then the dye along with the terminal 3' blocker is chemically removed from the DNA, allowing the next cycle. Unlike pyrosequencing, the DNA chains are extended one nucleotide at a time and image acquisition can be performed at a delayed moment, allowing for very large arrays of DNA colonies to be captured by sequential images taken from a single camera [77].

- *SOLiD sequencing*: in this method before sequencing, the DNA is amplified by emulsion PCR. The resulting beads, each containing single copies of DNA molecule, are deposited on a glass slide [78].

In the *Figure 1.7* it can be seen the technological features of the principal methods of the second generation sequencing [79]:

In the following figure is shown a comparison of the principal DNA sequencers of the first and second generation [80, 81]. If we want more accuracy the cost of sequencing will increase (as shown in *Table 1.1*); furthermore if we need a greater amount of sequenced DNA, we will lose in accuracy despite the price of sequencing is lowered.

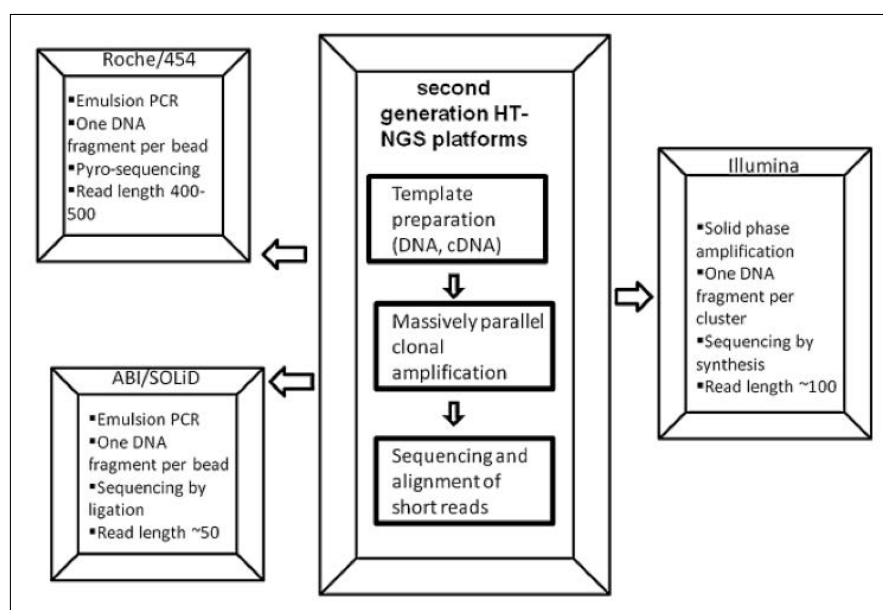


Figure 1.7: Advanced technological features of three leading second generation HT-NGS platforms.

Method	Single molecule real time sequencing	Ion semiconductor	Pyrosequencing (454)	Sequencing by synthesis (Illumina)	Sequencing by ligation (SOLiD sequencing)	Chain termination (Sanger sequencing)
Read length	2,900 bp average	200 bp	700 bp	50 to 250 bp	50 + 35 or 50 + 50	400 to 900 bp
Accuracy	87% to 90%	98%	99.9%	98%	99.9%	99.9%
Read per run	35 – 75 thousand	up to 5 million	1 million	up to 3 billion	1.2 to 1.4 billion	N/A
Time per run	30 minutes to 2 hours	2 hours	24 hours	1 to 10 days	1 to 2 weeks	20 minutes to 3 hours
Cost per 1 million bases	\$2	\$1	\$10	\$0.05 to \$0.15	\$0.13	2,400\$
Advantages	Longest read length. Fast.	Less expensive equipment. Fast	Long read size. Fast.	Potential for high sequence yield, depending upon sequencer model and desired application	Low cost per base	Long individual reads. Useful for many application
Disadvantages	Low yield at high	Homopolymer errors	Runs are expensive. Homopolymer errors.	Equipment can be very expensive	Slower than other methods.	More expensive and impractical for larger sequencing projects.

Table 1.1: Comparison of next-generation sequencing methods.

1.5.3 Third generation HT-NGS

“Although the PCR amplification has revolutionized DNA analysis, but in some instances it may introduce base sequence errors or favor of certain sequences over others, thus changing the relative frequency and abundance of various DNA fragments that existed before amplification. To overcome this, the ultimate miniaturization into the nanoscale and the minimal use of biochemicals, would be achievable if the sequence could be determined directly from a single DNA molecule, without the need for PCR amplification and its potential for distortion of abundance levels. This sequencing from a single DNA molecule is now called as the third generation of HT-NGS technology” [79]. Here is a list of the principle methods belonging to the third generation of HT-NGS we have, even if we do not go into the details of each individual method:

- *HeliscopeTM single molecule sequencer;*
- *Single molecule real time (SMRTtm) sequencer;*
- *Single molecule real time (RNAP) sequencer;*
- *Nanopore DNA sequencer;*
- *Real time single molecule DNA sequencer platform developed by Visi-Gen Biotechnologies;*
- *Multiplex polony technology;*
- *The Ion Torrent sequencing technology;*

In the *Figure 1.8* we can see the technological features of the principal methods of the third generation sequencing [79]:

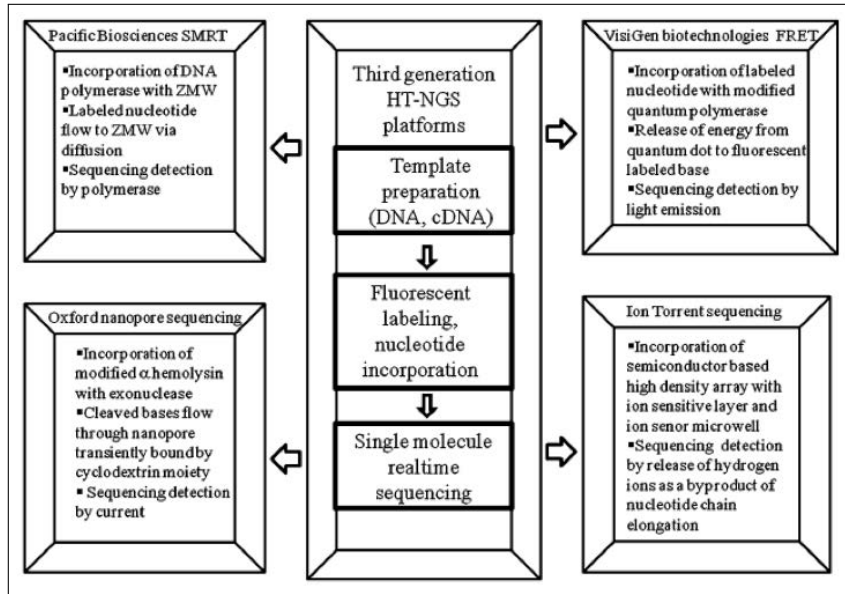


Figure 1.8: Advanced technological features of three leading third generation HT-NGS platforms.

1.5.4 Application of sequencing technologies on human genome research

As said above, there are many applications of the sequencing technologies especially in the field of research and medical care. Recently, a large quantitative of studies done by the use of the HT-NGS have emerged, particularly:

- *Epigenetics*;
- *ChiP-Seq*;
- *Genome wide structural variation in human population*;
- *Detection of inherited disorders*;
- *Complex human disease*;
- *Cancer research*;
- *RNA sequencing*;
- *Personal genomics*;
- *Sequencing of mitochondrial genome*.

For this reason, it comes the need to improve more and more speed, accuracy and price of the sequencing technologies. This is the challenge many laboratories are trying to overcome.

1.5.5 Alignment tools

Unlike *Sanger sequencing*, the new generation technologies of DNA sequencing produce a very large quantity of small (from 50 to 300 nucleotides) DNA fragments (from hundreds of thousands to billions sequences). One of the problems once obtained these data, is not only their storage, but also the mapping of each of them in the reference genome. In the *Figure 1.9* we can see an example of how the reads are mapped in the genome:

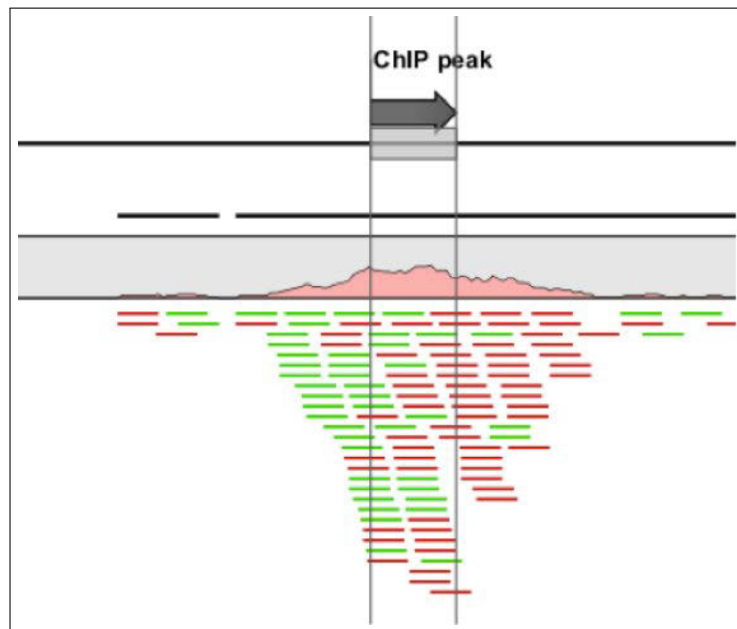


Figure 1.9: Alignment of the reads in the genomic reference.

There is a large number of tools for aligning reads to a genome:

- *Bowtie* (<http://bowtie-bio.sourceforge.net/>);
- *BWA* (<http://maq.sourceforge.net/>);
- *Eland* (<http://bioinfo.cgrb.oregonstate.edu/docs/solexa/>);

- **MAQ** (<http://sourceforge.net/projects/maq/>).

Each of them uses different scoring methods, including quality (Q) score of the reads. The most popular tool is Bowtie [82]. It uses the *Burrows-Wheeler transformation* in order to index the genome, which increases the speed of the alignment. Once it builds the index for the genome, the latter is used to identify potential matches to read. Furthermore, it allows mismatches and the user can control the number of matches in the first part of read.

Because the sequenced read are small, the mapping of them could be misleading, in particular when the read fall into a repetitive region. A typical approach to overcome this kind of problem is to ignore reads mapped into repetitive regions. Another problem consists in the fact that a single read can be aligned in several regions of the reference genome, and in this case just one mapped region is typically randomly chosen.

Chapter 2

Biostatistics algorithms and Markovian models

*Those who ignore Statistics are
condemned to reinvent it.*

Bradley Efron

Stanford University

2.1 What is *Biostatistics*?

The word *Biostatistics* comes from the juxtaposition of two words: *bios*, the Greek word for *life*, and *statistics*, a field of mathematics dealing with the collection, analysis, interpretation of masses of numerical data¹. Thus, the main object of Biostatistics is to use procedures and analysis of statistics in studying and practicing biology. Biostatistics is that branch of statistics with the specific aim to read and understand scientific data generated in the biol-

¹The Merriam-Webster's Collegiate Dictionary

ogy, public health and other health sciences (as, for example, the *biomedical sciences*). It can be considered as a broad discipline focusing on the application of statistical theory to real and daily problems, the practice of designing and conducting biomedical experiments and clinical trials, the study of related computational algorithms and display of data, and the development of mathematical statistical theory.

2.2 Steps to get to reliable results

According to the previous definition of Statistics, there are several steps to follow to get to trustworthy results:

- ***Collection of data***: this is the real first stage in the investigation and it is necessary for the data to be reliable and accurate.
- ***Organization of data***: first of all the data need to be edited for *completeness, inconsistencies, homogeneity, accuracy, reliability*. They are, then, classified, which means that they are arranged following a certain common characteristic. Finally, they are often tabulated, that is presented in rows and columns to be clearer.
- ***Analysis of data***: they are done through the observation and the application of statistical techniques (i.e. measures of central tendency, variation, and so on), as well as the creation and the use of mathematical models to mine the data in order to discovery some unknown information.

- **Interpretation:** this last step refers to the conclusion the scientists get to. While, obviously, wrong interpretation lead to unreliable conclusions and decisions, correct interpretation will take to good decisions.

The characteristic of these fields is that patients, mice, cells, etc. show a clear variation in their response to stimuli. This may be due either to the different treatment or to chance, measurement error, or even to other characteristics of the single subjects. Biostatistics tries not only to distinguish between correlation and causation, but also to make inferences from known samples about the populations² from which they were drawn. That is why to drive biology experiments, data is gathered and analyzed before, during, and after a biology experiment, with the purpose to get to logical solutions.

As Biostatistics represents the link between theory and practice, some experiment can also be entirely mathematical and expressed only in numerical terms. Furthermore, there are two types of data: *qualitative* (they are non numeric and in a written form, as, for example, the description of events, the transcription of an interview, written documents etc.), and *quantitative* (they are numerical).

2.3 Application fields

There are several fields where Biostatistics can be applied, such as biology, clinical medicine, public health policy, physiology and anatomy, epidemiology, genetics, health economics, proteomics, genomics. Moreover, it is

²In Biostatistics the word “*population*” is used with the meaning of “*set of measurements*”.

usually used in large-scale efforts, such as drug testing and environmental model-building, as it happens for the trials for new pharmaceuticals, where Biostatistics tracks and gives an interpretation about data. In Medicine, Biostatisticians are important also in the evaluation of the spread of a disease, analyzing the information given by the scientists (such as the samples of people who have contracted a disease, life history, and social conditions of others who live in the same area), in order to see why some people got a disease and others did not. In genetic research, Biostatistics helps in finding a cure for deadly diseases, and causes for genetic condition. Thanks to the combination of Biostatistics and probability theory, it is possible to determine with a given set of data the likelihood of a disease to hit populations, drugs to cure it, and the reaction of the population to those drugs.

Of course, even if the importance of Biostatistics is not questioned, it is honest to identify some limitations to this science. First of all if the samples used in a statistical test are not adequately representative of the population, the results can have little relevance to the data it come from. Moreover, unlike physical sciences, the laws of Statistics are not perfect, but related to probability, which means that its conclusions are true only on an average.

2.4 Deterministic and Stochastic models

As seen above, after the *organization* phase, the data are ready for the *modelling* phase. In this step, it is possible both to extract from the data some unknown information, able to describe the system considered, and to study the effects of different components so as to make predictions about the be-

haviour.

A mathematical model can be composed of a set both of *variables* and *relationships*. The variables are abstractions of interest of the system, that can be quantified, while relationships can be represented by operators (such as *functions*, *algebraic operators*, and so on) which may act with or without variables. Based on variables and relationships it is decided whether to apply a ***deterministic*** or a ***stochastic*** model. In a deterministic model there are not stochastic elements and the set of variable states is uniquely determined by parameters in the model and by sets of previous states of these variables. Therefore, deterministic models perform the same way for a given set of initial conditions. Contrariwise, a stochastic model has one or more stochastic elements and then variable states are not described by unique values, but rather by probability distributions. Examples of stochastic models are *Poisson* processes, *Gaussian* processes, *Markov* processes, *hidden Markov* processes, and so on.

2.5 Markov models

As said above, *Markov* models are stochastic models that can be listed as following:

- When the system is *autonomous*:
 - *Markov chains*: they are the simplest Markov models, in which states of the system are modelled by random variables that change over time. Furthermore, the distribution of the random variables

depend only on the distribution of the *previous* states.

- *Hidden Markov models (HMMs)*: they are Markov chains where states are partially observables. In fact, there is a sequence of observable states that is emitted by a sequence of internal hidden states³. In general, the transition from one hidden state to another one has the form of a *first-order* Markov chain.
- When the system is *controlled*:
 - *Markov decision processes (MDPs)*: they are similar to Markov chains where transitions of the states depend on the current state and the control of a decision maker.
 - *Partially observable Markov decision processes (POMDPs)*: they are *Markov decision processes* where, similarly to the HMMs, the states are partially observed.

To follow will be given a formalization of Markov chains and hidden Markov models

2.6 Markov chains

A discrete-time stochastic process $\mathbf{X} = \{X_n : n \in I\}$, where $I = \{0, 1, 2, \dots\}$, with finite or countable *state space*⁴ $X_n \in \{0, 1, 2, \dots\}$ is a *Markov chain* if it has the *Markov property* (*Property 1*), also known as “*memoryless*” prop-

³Hidden states can not be observed directly.

⁴In this formalization the *state space* is a countable set.

erty⁵ (In Figure 2.1 is shown an example of a *two-state Markov chain*).

Property 1. (*Markov property*)

For any $z, s_0, \dots, s_{n-1} \in S$ and any $n \geq 1$,

$$Pr(X_n = z \mid X_{n-1} = s_{n-1}, \dots, X_0 = s_0) = Pr(X_n = z \mid X_{n-1} = s_{n-1}). \quad (2.1)$$

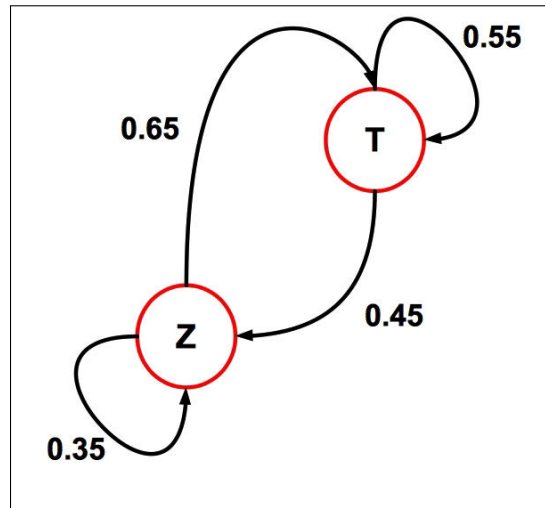


Figure 2.1: A simple two-state Markov chain.

It has just been defined a *first order Markov chain* in which the probability that an event occurs at the time n is conditionally dependent on the event that occurred at the previous instant.

In general, considering an integer $k > 1$ it is possible to define a *Markov chain of order k* as following:

⁵In words, the *Markov* property says given the present state of the stochastic process, the past is *conditionally independent* of the future.

Definition 1. (K^{th} order Markov chain)

$$\begin{aligned} & Pr(X_n = s_n \mid X_{n-1} = s_{n-1}, X_{n-2} = s_{n-2}, \dots, X_0 = s_0) \\ &= Pr(X_n = s_n \mid X_{n-1} = s_{n-1}, X_{n-2} = s_{n-2}, \dots, X_{n-k} = s_{n-k}) \text{ for } n > k. \end{aligned} \tag{2.2}$$

In the case of K^{th} order Markov chain the probability that an event occurs only depends on the values of the previously k states.

Finally, considering a sequence of n states $\Pi = (\pi_1, \pi_2, \dots, \pi_n)$ of *first order Markov chain*, we have:

- the value of π_i is only conditionally dependent on π_{i-1} (as seen in *Property 1*),
- in particular for each i , we have that:

$$Pr(\pi_i \mid \pi_{i-1}) = a_{\pi_{i-1}, \pi_i}, \tag{2.3}$$

where a_{π_{i-1}, π_i} is defined as *transition probability*. Then, given the sequence of states Π , the probability of the sequence, $Pr(\Pi) = Pr(\pi_1, \dots, \pi_n)$, can be calculated as follows:

$$\begin{aligned}
Pr(\pi_1, \dots, \pi_{n-1}, \pi_n) &= Pr(\pi_n \mid \pi_1, \dots, \pi_{n-1}) Pr(\pi_{n-1} \mid \pi_1, \dots, \pi_{n-2}) \cdot \dots \cdot Pr(p_2 \mid p_1) \\
&= Pr(\pi_n \mid \pi_{n-1}) Pr(\pi_{n-1} \mid \pi_{n-2}) \cdot \dots \cdot Pr(p_2 \mid p_1) \\
&= Pr(\pi_1) \prod_{i=2}^n a_{\pi_{i-1}, \pi_i}.
\end{aligned} \tag{2.4}$$

For nucleotide sequences, the Markov chain model has four states, $\{A, C, G, T\}$, and to define the transition probabilities of $4 \times 4 = 16$. We can also add an extra begin state to the model by defining $x_0 = B$. Then the probability of the first letter in the sequence is

$$Pr(x_1 = z) = a_{Bs}. \tag{2.5}$$

2.7 Hidden Markov models

In the scientific literature there are several methods of classification, e.g. simple methods as *decisional trees*, *k-Nearest Neighbor*, *Bayesian networks*, or more complicated as *Support Vector Machines*. Each of them requires a priori knowledge that is not always possible to obtain, as also the fact that often there is the need to define a larger number of variables. The stochastic approach of the *hidden Markov* models, unlike other classification methods, gives the possibility to efficiently and effectively carry out this task.

Once, the topology of the model and its variables are defined, it is possible to solve three kinds of problems:

- **Evaluation problem:** Given the observation sequence O and a model θ we can find $Pr(O | \theta)$;
- **Decoding problem:** Given the observation O and the model θ we can find the corresponding state sequence ! which is optimal in some meaningful sense;
- **Training problem:** Find the model parameter θ in order to maximize $Pr(O | \theta)$.

2.7.1 Formalization of Hidden Markov models

An HMM is characterized by the following parameters:

- M , the number of observation symbols per state. If it is considered an experiment with coin tosses, the alphabet of observations consists in two possible symbols, H and T ; instead, if there is an experiment of rolls of dice, there will be six distinct symbols (1, 2, 3, 4, 5 and 6); while, in the case of a DNA sequence, four possible symbols, A , T , G and C , will be taken into consideration.
- N , the number of states in the model. In the case of the experiment of the *occasionally dishonest casino* there will be two possible hidden states, when the croupier use a fair die (the state F) and when he use a loaded die (the state L). In the case of *CpG islands*⁶, there will be two different states “+” (in a CpG island) and “-” (outside of a CpG island).

⁶*CpG islands* are genomic regions with a high frequency of *CG* sites

- The *transition probabilities* that we can be indicated as follows

$$a_{kl} = Pr(\pi_i = l \mid \pi_{i-1} = k), 1 \leq k, l \leq N. \quad (2.6)$$

- The *emission probabilities* in state k ,

$$e_k(b_j) = Pr(x_i = b_j \mid \pi_i = k), 1 \leq j \leq M \text{ and } 1 \leq k \leq N. \quad (2.7)$$

- The *initial state* distribution is

$$a_{0k} = Pr(\pi_1 = k \mid \pi_0), 1 \leq k \leq N. \quad (2.8)$$

In this case the transition probability from the *begin state* (this state is an *silent state* where $Pr(\pi_0 = 1)$) to state k can be thought as the probability of starting in state k .

2.7.2 The *forward* algorithm

The *forward* algorithm is able to solve the *evaluation problem* and then calculate the probability of the observation sequence of length n , $\mathbf{x} = x_1, x_2, \dots, x_n$, given the model: the *transition probabilities*, the *emission probabilities* and the *initial state distribution*.

Because many different state paths can give rise to the same sequence x , we must add the probabilities for all possible paths to obtain the full probability of x . Then,

$$\begin{aligned}
Pr(x) &= \sum_n Pr(x, \pi) \\
&= \sum_{\pi_0} \sum_{\pi_1} \cdots \sum_{\pi_n} Pr(x_1, \dots, x_n, \pi_1, \dots, \pi_n, \pi_0) \\
&= \sum_{\pi_n} \cdots \sum_{\pi_1} \sum_{\pi_0} Pr(x_n | \pi_n) Pr(\pi_n | \pi_{n-1}) \cdots \\
&\quad Pr(x_2 | \pi_2) Pr(\pi_2 | \pi_1) Pr(x_1 | \pi_1) Pr(\pi_1 | \pi_0) Pr(\pi_0),
\end{aligned} \tag{2.9}$$

reordering the terms, it will be,

$$\begin{aligned}
Pr(x) &= \sum_{\pi_n} Pr(x_n | \pi_n) \sum_{\pi_{n-1}} Pr(\pi_n | \pi_{n-1}) Pr(x_{n-1} | \pi_{n-1}) \cdots \\
&\quad \sum_{\pi_1} Pr(\pi_2 | \pi_1) Pr(x_1 | \pi_1) \sum_{\pi_0} Pr(\pi_1 | \pi_0) Pr(\pi_0),
\end{aligned} \tag{2.10}$$

where $Pr(\pi_0 = 1)$ then,

$$Pr(\pi_1 | \pi_0) Pr(\pi_0) = Pr(\pi_1, \pi_0), \tag{2.11}$$

and,

$$\sum_{\pi_0} Pr(\pi_1, \pi_0) = Pr(\pi_1) = a_{0k}. \tag{2.12}$$

Finally,

$$Pr(x) = \sum_{\pi_n} Pr(x_n|\pi_n) \sum_{\pi_{n-1}} Pr(\pi_n|\pi_{n-1}) Pr(x_{n-1}|\pi_{n-1}) \cdots \sum_{\pi_1} Pr(\pi_2|\pi_1) Pr(x_1|\pi_1) Pr(\pi_1), \quad (2.13)$$

then,

$$\begin{aligned} Pr(x) &= \sum_{\pi_n} Pr(x_n|\pi_n) \sum_{\pi_{n-1}} Pr(\pi_n|\pi_{n-1}) Pr(x_{n-1}|\pi_{n-1}) \cdots \\ &\quad \sum_{\pi_2} Pr(\pi_3|\pi_2) Pr(x_2|\pi_2) \sum_{\pi_1} Pr(\pi_2|\pi_1) Pr(\pi_1, x_1), \end{aligned} \quad (2.14)$$

where,

$$Pr(\pi_2 | \pi_1) Pr(x_1, \pi_1) = Pr(\pi_2 | \pi_1, x_1) Pr(\pi_1, x_1) = Pr(\pi_2, \pi_1, x_1), \quad (2.15)$$

and then,

$$\sum_{\pi_1} Pr(\pi_2, \pi_1, x_1) = Pr(\pi_2, x_1). \quad (2.16)$$

Substituting *Expression 2.16* in *2.14* the result will be,

$$Pr(x) = \sum_{\pi_n} Pr(x_n|\pi_n) \sum_{\pi_{n-1}} Pr(\pi_n|\pi_{n-1}) Pr(x_{n-1}|\pi_{n-1}) \cdots \sum_{\pi_2} Pr(\pi_3|\pi_2) Pr(x_2|\pi_2) Pr(\pi_2, x_1). \quad (2.17)$$

And so on for the other terms, until obtaining the following expression:

$$Pr(x) = \sum_{\pi_n} Pr(x_n|\pi_n) \sum_{\pi_{n-1}} Pr(\pi_n|\pi_{n-1})Pr(\pi_{n-1}, x_1, \dots, x_{n-1}). \quad (2.18)$$

The number of all possible paths π increase exponentially with the length of the sequence, so *brute force procedure* to calculate $Pr(x)$ considering all paths is not practical, then it is possible to use a *dynamic programming procedure* in order to calculate the probability of the observation sequence, $P(x)$. This procedure is called *forward* algorithm. Considering the scheme in *Figure 2.2*:

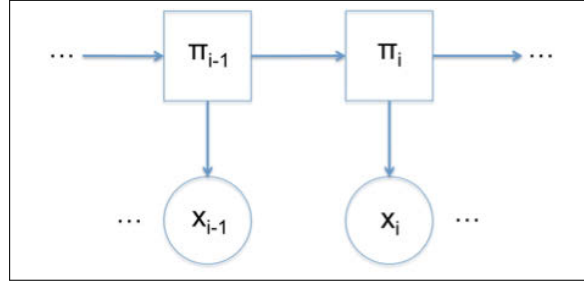


Figure 2.2: (HMM) Scheme of transition from the state π_{i-1} to the state π_i .

The *forward* algorithm computes the probability $Pr(\pi_i = l, x_{1:i})$, where $x_{1:i} = (x_1, \dots, x_i)$:

$$f_l(i) = Pr(\pi_i = l, x_{1:i}) = \sum_{\pi_{i-1}} Pr(\pi_i = l, \pi_{i-1}, x_{1:i}), \text{ for } 1 \leq l \leq N. \quad (2.19)$$

In order to use known parameters of the model, for $1 \leq l \leq N$, the expression can be write in the following way,

$$\begin{aligned}
f_l(i) &= Pr(\pi_i = l, x_{1:i}) \\
&= \sum_{\pi_{i-1}} Pr(x_i \mid \pi_i = l, \pi_{i-1}, x_{1:i-1}) Pr(\pi_i = l \mid \pi_{i-1}, x_{1:i-1}) Pr(\pi_{i-1}, x_{1:i-1}).
\end{aligned} \tag{2.20}$$

In particular, x_i is conditional independent on x_{i-1} and $x_{1:i-1}$ given π_i , then,

$$Pr(x_i \mid \pi_i = l, \pi_{i-1}, x_{1:i-1}) = Pr(x_i \mid \pi_i = l), \text{ for } 1 \leq k, l \leq N, \tag{2.21}$$

where $Pr(x_i \mid \pi_i)$ is the *emission probability*, a known parameter of the model. Furthermore, π_i is conditional independent on $x_{1:i-1}$ given π_{i-1} , and then

$$Pr(\pi_i = l \mid \pi_{i-1} = k, x_{1:i-1}) = Pr(\pi_i = l \mid \pi_{i-1} = k), \text{ for } 1 \leq k, l \leq N, \tag{2.22}$$

where $Pr(\pi_i \mid \pi_{i-1})$ is the *transition probability*, a known parameter of the model.

Substituting the *Expressions* 2.21 and 2.22 the result will be,

$$Pr(\pi_i = l, x_{1:i}) = \sum_{\pi_{i-1}} Pr(x_i \mid \pi_i = l) Pr(\pi_i = l \mid \pi_{i-1}) Pr(\pi_{i-1}, x_{1:i-1}) \tag{2.23}$$

where $Pr(\pi_{i-1}, x_{1:i-1}) = f_k(i-1)$. Then, the *recursion equation* can be

written as follows,

$$\begin{aligned}
 f_l(i) &= \sum_{\pi_{i-1}} Pr(x_i \mid \pi_i = l) Pr(\pi_i = l \mid \pi_{i-1}) f_k(i-1) \\
 &= Pr(x_i \mid \pi_i = l) \sum_{\pi_{i-1}} f_k(i-1), \text{ for } 1 \leq k, l \leq N \text{ and for } i = 2, \dots, n.
 \end{aligned} \tag{2.24}$$

For $i = 1$, the expression will become:

$$f_l(1) = Pr(\pi_1 = k, x_1) = Pr(x_1 \mid \pi_1 = k) Pr(\pi_1 = k), \text{ for } 1 \leq l \leq N. \tag{2.25}$$

where $Pr(\pi_1 = k)$ is the *initial probability* that is a known parameter of the model.

Finally, when all *forward variables* are computed, $f_l(i)$, it is possible to calculate:

$$Pr(x) = \sum_{k=1}^N f_k(n). \tag{2.26}$$

Pseudo-Code of the *forward* algorithm

```

input  :  $M, X = (x_1, \dots, x_n)$ 
output:  $p(x)$ 

Initialization;
for  $i = 1$  to  $k$  do
     $f_1(i) = P(p_1 = i) \cdot e_i(x_1);$ 
end

// Calculating of forward probabilities;
for  $t = 1$  to  $n - 1$  do
    for  $j = 1$  to  $k$  do
         $sum \leftarrow 0;$ 
        for  $i = 1$  to  $k$  do
             $sum \leftarrow sum + f_t(i) \cdot a(i, j);$ 
            // adding contributions of transition
            probabilities with  $f_t(i)$ 
        end
         $f_{t+1}(j) = sum \cdot e_j(x_{t+1});$ 
    end
end

// Calculating of  $p(x);$ 
 $p \leftarrow 0;$ 
for  $i = 1$  to  $k$  do
     $p \leftarrow p + f_{t+1}(i);$ 
end

```

Algorithm 1: Forward Algorithm

2.7.3 The *backward* algorithm

The *backward* algorithm for HMM is very similar to the *forward* algorithm, and it is able to solve the *evaluation problem*. For the *forward* algorithm is computed the $f_l(1), f_l(2), \dots, f_l(n)$, for $1 \leq l \leq N$, while for the *backward* algorithm the values will be computed from the *end* to the *start* of the sequence, in an opposite direction to one of the *forward* algorithm.

As previous algorithm, the *backward* algorithm calculates the probability of the observation sequence of length n , $\mathbf{x} = x_1, x_2, \dots, x_n$, given the model, so the *transition probabilities*, the *emission probabilities* and the *initial state distribution* are known. In order to explain this algorithm the *Figure 2.3* will be taken into consideration:

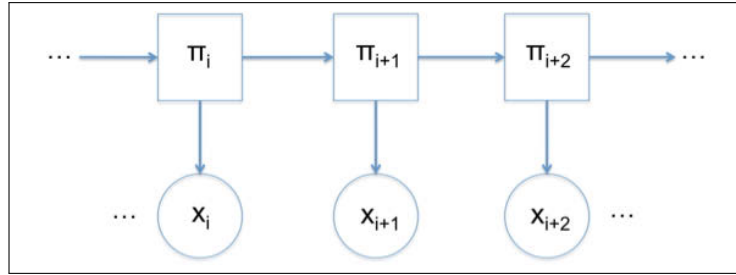


Figure 2.3: (HMM) Scheme of transition from the state π_i to the state π_{i+2} .

The backward algorithm computes the probability $Pr(x_{i+1:n} \mid \pi_i = k)$, where $x_{i+1:n} = (i+1, \dots, n)$. For $1 \leq k \leq N$.

$$\begin{aligned}
b_k(i) &= Pr(x_{i+1:n} \mid \pi_i = k) \\
&= \sum_{\pi_{i+1}} Pr(x_{i+1:n}, \pi_{i+1} \mid \pi_i = k) \\
&= \sum_{\pi_{i+1}} Pr(x_{i+2:n} \mid \pi_{i+1}, \pi_i = k, x_{i+1}) Pr(x_{i+1} \mid \pi_{i+1}, \pi_i = k) Pr(\pi_{i+1} \mid \pi_i = k).
\end{aligned} \tag{2.27}$$

Looking at the schema in *Figure 2.3*, it can say that $x_{i+2:n}$ is conditionally independent on π_i and x_{i+1} given π_{i+1} , then

$$Pr(x_{i+2:n} \mid \pi_{i+1} = l, \pi_i = k, x_{i+1}) = Pr(x_{i+2:n} \mid \pi_{i+1} = l), \text{ for } 1 \leq k, l \leq N, \tag{2.28}$$

where $Pr(x_{i+2:n} \mid \pi_{i+1}) = b_l(i+1)$. Furthermore, x_{i+1} is conditionally independent on $\pi_i = k$ given π_{i+1} , then

$$Pr(x_{i+1} \mid \pi_{i+1} = l, \pi_i = k) = Pr(x_{i+1} \mid \pi_{i+1} = l), \text{ for } 1 \leq k, l \leq N, \tag{2.29}$$

in which $Pr(x_{i+1} \mid \pi_{i+1} = l)$ is the *emission probability* that we know, and we also know the last term of $b_k(i)$, $Pr(\pi_{i+1} \mid \pi_i = k)$, in fact it is the *transition probability*.

Finally, for $1 \leq k \leq N$ and for $i = n-1, \dots, 1$, the *recursion equation* can be written as follows,

$$b_k(i) = \sum_{\pi_{i+1}} Pr(\pi_{i+1} \mid \pi_i = k) Pr(x_{i+1} \mid \pi_{i+1} = l) b_l(i+1). \quad (2.30)$$

For $i = n$, the expression will become:

$$b_k(n) = 1, \text{ for } 1 \leq k \leq N. \quad (2.31)$$

When all *backward variables* are computed, $b_k(i)$, it is possible to calculate:

$$Pr(x) = \sum_{l=1}^N Pr(\pi_1 = l) Pr(x_1 \mid \pi_1 = l) b_l(1). \quad (2.32)$$

Pseudo-Code of the *backward* algorithm

```

input  :  $M, X = (x_1, \dots, x_n)$ 
output:  $p(x)$ 

Initialization;
for  $i = 1$  to  $k$  do
    |    $b_n(i) = 1;$ 
end

// Calculating of backward probabilities;
for  $t = n - 1$  to  $1$  do
    |   for  $i = 1$  to  $k$  do
        |    $sum \leftarrow 0;$ 
        |   for  $j = 1$  to  $k$  do
            |    $sum \leftarrow sum + b_{t+1}(i) \cdot a(i, j) \cdot e_j(x_{t+1});$ 
            |   // adding contributions of transition
            |   probabilities with  $b_{t+1}(i);$ 
        |   end
        |    $b_t(i) = sum;$ 
    |   end
end

// Calculating of  $p(x);$ 
 $p \leftarrow 0;$ 
for  $i = 1$  to  $k$  do
    |    $p \leftarrow p + b_1(i);$ 
end

```

Algorithm 2: Backward Algorithm

2.7.4 The *Viterbi* algorithm

The *Viterbi* algorithm is able to solve the *decoding problem*, then it can be used to find the path of state, π^* , with the highest probability considering an observation sequence of length n , $\mathbf{x} = x_1, x_2, \dots, x_n$ and given the model, therefore the *transition probabilities*, the *emission probabilities* and the *initial state* distribution are known:

$$\begin{aligned}
 \pi^* &= \arg \max_{\pi} Pr(\pi, x) \\
 &= \arg \max_{\pi_{0:n}} Pr(\pi_0, \pi_{1:n}, x_{1:n}) \\
 &= \arg \max_{\pi_n} Pr(x_n | \pi_n) \arg \max_{\pi_{n-1}} Pr(\pi_n | \pi_{n-1}) Pr(x_{n-1} | \pi_{n-1}) \cdots \\
 &\quad \arg \max_{\pi_1} Pr(\pi_2 | \pi_1) Pr(x_1 | \pi_1) \arg \max_{\pi_0} Pr(\pi_1 | \pi_0) Pr(\pi_0),
 \end{aligned} \tag{2.33}$$

where $\arg \max_{\pi_0} Pr(\pi_1 | \pi_0) Pr(\pi_0)$ can be considered as the first most likely state, $\hat{\pi}_1$, of the sequence $\pi_{1:n}$.

$$\arg \max_{\pi_0} Pr(\pi_1 | \pi_0) Pr(\pi_0) = Pr(\pi_1). \tag{2.34}$$

Then,

$$\begin{aligned}
 \pi^* &= \arg \max_{\pi_n} Pr(x_n | \pi_n) \arg \max_{\pi_{n-1}} Pr(\pi_n | \pi_{n-1}) Pr(x_{n-1} | \pi_{n-1}) \cdots \\
 &\quad \arg \max_{\pi_1} Pr(\pi_2 | \pi_1) Pr(x_1 | \pi_1) Pr(\pi_1).
 \end{aligned} \tag{2.35}$$

At the same time, the probability of the most likely path π^* can be found,

$$\begin{aligned}
Pr(\pi^*, x) &= \max_{\pi_n} Pr(x_n | \pi_n) \max_{\pi_{n-1}} Pr(\pi_n | \pi_{n-1}) Pr(x_{n-1} | \pi_{n-1}) \cdots \\
&\quad \max_{\pi_{n-1}} Pr(\pi_2 | \pi_1) Pr(x_1 | \pi_1) Pr(\pi_1).
\end{aligned} \tag{2.36}$$

In order to explain this algorithm, the *Figure 2.4* will be taken into consideration:

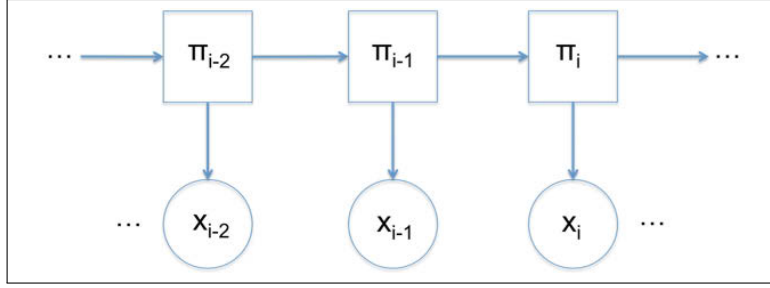


Figure 2.4: (HMM) Scheme of transition from the state π_{i-2} to the state π_i .

The aim is to find the most probable path, π^* , recursively by using the Markov properties of the model. Then the following expression can be used.

$$v_l(i) = \max_{\pi_{1:i-1}} Pr(\pi_{1:i}, x_{1:i}) \tag{2.37}$$

considering the *conditional independents*, it will be,

$$v_l(i) = \max_{\pi_{1:i-1}} Pr(x_i | \pi_i = l) Pr(\pi_i = l | \pi_{i-1}) Pr(\pi_{1:i-1}, x_{1:i-1}) \tag{2.38}$$

where here, it can be used the following property,

Property 2. *Given two non-negative functions $f(a) \geq 0, \forall a \in \mathbb{R}$ and $g(a, b) \geq 0$*

$0, \forall a, b \in \mathbb{R}$, then

$$\max_{a,b} f(a)g(a,b) = \max_a [f(a) \max_b g(a,b)] \quad (2.39)$$

Using $a = i - 1$ and $b = 1 : i - 2$ and assuming

$$f(a) = Pr(x_i \mid \pi_i = l) Pr(\pi_i = l \mid \pi_{i-1}), \quad (2.40)$$

and $g(a,b) = Pr(\pi_{1:i-1}, x_{1:i-1})$, then

$$v_l(i) = \max_{\pi_{i-1}} [Pr(x_i \mid \pi_i = l) Pr(\pi_i \mid \pi_{i-1}) \max_{1:i-2} Pr(\pi_{1:i-1}, x_{1:i-1})], \quad (2.41)$$

where,

$$v_k(i-1) = \max_{1:i-2} Pr(\pi_{1:i-1}, x_{1:i-1}). \quad (2.42)$$

Therefore, for $i = 2, \dots, n$,

$$\begin{aligned} v_l(i) &= \max_{\pi_{i-1}} Pr(x_i \mid \pi_i = l) Pr(\pi_i = l \mid \pi_{i-1}) v_k(i-1) \\ &= Pr(x_i \mid \pi_i = l) \max_{\pi_{i-1}} Pr(\pi_i = l \mid \pi_{i-1}) v_k(i-1). \end{aligned} \quad (2.43)$$

For $i = 1$, the expression will become:

$$v_l(1) = Pr(x_1 \mid \pi_1 = l) \max_{\pi_0} Pr(\pi_1 = l \mid \pi_0) v_k(0), \quad (2.44)$$

where setting $v_k(0) = 1$.

$$v_l(1) = Pr(x_1 \mid \pi_1 = l)Pr(\pi_1 = l \mid \pi_0). \quad (2.45)$$

Finally,

$$Pr(\pi^*, x) = \max_{\pi_{1:n}} v_k(n) = \max_{\pi_{1:n}} Pr(\pi_{1:n}, x_{1:n}), \quad (2.46)$$

Concerning the most likely path, it can be done the same considerations, and then it is possible to consider the following procedure:

$$ptr_i(l) = \arg \max_{\pi_{i-1}} Pr(\pi_i = l \mid \pi_{i-1})v_k(i-1), \text{ for } i = 2, \dots, n. \quad (2.47)$$

The final step will be:

$$\pi^* = \arg \max_{\pi_{i-1}} v_k(n). \quad (2.48)$$

When the last hidden state is found, it can be used a *traceback* procedure in order to have in output the most likely hidden state sequence, π^* , then

$$\pi_{i-1}^* = ptr_i(\pi^*), \text{ for } i = L, \dots, 1. \quad (2.49)$$

Pseudo-Code of the *Viterbi* algorithm

```

input  :  $M, X = (x_1, \dots, x_n)$ 
output:  $P = \{p_1, \dots, p_n\}$  tale che  $p(X, P)$ 

Initialization;
for  $i = 1$  to  $k$  do
     $V_1(i) = P(p_1 = i) \cdot e_i(x_1);$ 
     $\phi_1(i) \leftarrow 0;$ 
end
for  $t = 1$  to  $n - 1$  do
    for  $j = 1$  to  $k$  do
         $\max \leftarrow 0;$ 
         $\phi_1(j) \leftarrow 0;$ 
        for  $i = 1$  to  $k$  do
            if  $V_t(i) \cdot a(i, j) > \max$  then
                 $\max \leftarrow V_t(i) \cdot a(i, j);$ 
                 $\phi_1(j) \leftarrow i;$ 
            end
        end
         $V_{t+1}(j) = \max \cdot e_j(x_{t+1});$ 
    end
end

```

Algorithm 3: Viterbi Algorithm: first part.

```

// Calculating of  $p(X, P), P$ ;
 $P \leftarrow 0$ ;
for  $i = 1$  to  $k$  do
    if  $V_n(i) > P$  then
         $P = V_n(i)$ ;
         $\phi_n(n) = i$ ;
    end
end

```

Algorithm 4: Viterbi Algorithm: final part.

2.7.5 The *Baum-Welch* algorithm

Given a training set of observed sequences, $\mathbf{X} = x^1, \dots, x^J$, each of them of length n , and the goal is that to construct an HMM with the parameters, $\theta = \{a_{0k}, ek(\cdot), a_{kl}\}$, to maximize the likelihood of our data,

$$Pr(x^1, \dots, x^J \mid \theta). \quad (2.50)$$

When the paths of the hidden state are known for the training sequences the parameters of the model can be estimated using *Maximum Likelihood Estimation*. If the sequences are not labeled with the hidden states, as in this case, an iterative procedure in order to estimate the parameter values can be used. In particular the *Baum-Welch* algorithm will be used to do this. *Baum-Welch* algorithm is an iterative procedure used to estimate the model parameters when the state paths of the observed sequences are unknown. It belongs to the family of *Expectation Maximization* (**EM**) algorithms. Then,

it works by presuming initial parameter values, then estimating the likelihood of the data under current parameters. After that, these likelihoods are used to re-estimate the parameters, iteratively until a local maximum is reached.

Then after the initial parameter values are defined, *Baum-Welch* algorithm estimates the A_{kl} and $E_k(b_j)$ as the expected number of times each transition or emission is used, by using the *current* parameter values. After that, it uses these values to iteratively update the new values of *as* and *es*. This procedure is repeated until some stopping criterion is reached. Here, it will be considered the following criterion:

$$l(x^1, \dots, x^J \mid \hat{\theta}) - l(x^1, \dots, x^J \mid \theta) < \epsilon, \text{ for some small } \epsilon > 0, \quad (2.51)$$

where $l(x^1, \dots, x^J \mid \hat{\theta}) - l(x^1, \dots, x^J \mid \theta)$, and θ represents the *current* set of values of the parameters, while $\hat{\theta}$ is the *estimated* set of values of the parameters. In particular assuming that the sequences are independent, it can be written

$$\begin{aligned} l(x^1, \dots, x^J \mid \hat{\theta}) &= \log Pr(x^1, \dots, x^J \mid \hat{\theta}) \\ &= \sum_{j=1}^J \log Pr(x^j \mid \hat{\theta}) \\ l(x^1, \dots, x^J \mid \theta) &= \log Pr(x^1, \dots, x^J \mid \theta) \\ &= \sum_{j=1}^J \log Pr(x^j \mid \theta), \end{aligned} \quad (2.52)$$

then,

$$\left[\sum_{j=1}^J \log Pr(x^j | \hat{\theta}) \right] - \left[\sum_{j=1}^J \log Pr(x^j | \theta) \right] < \epsilon \quad (2.53)$$

Definition of variables

Before explaining the step of *Baum-Welch* algorithm, it is necessary to define some variables:

$$I_v^j = \begin{cases} 1 & \text{if } v \text{ is observed at time } t, \\ 0 & \text{otherwise.} \end{cases} \quad (2.54)$$

$I_v^j(t)$ is an indicator variable that is **1** when the t^{th} element of the j^{th} observed sequence is $v \in M$, where in the case of nucleotides there would be four possible symbols, $M = \{A, C, G, T\}$, and then four variables: $I_A^j, I_C^j, I_G^j, I_T^j$.

$$S_t^j(k) = \begin{cases} 1 & \text{if we are } k \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases} \quad (2.55)$$

$$S_t^j(k, l) = \begin{cases} 1 & \text{if we move from state } k \text{ to state } l \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases} \quad (2.56)$$

$S_t^j(k)$ is an indicator random variable that is **1** when the state at time t is k in the j^{th} observed sequence, while $S_t^j(k, l)$ is a random variable that is **1** when at time t we are in state k and we move to state l in the j^{th} observed sequence.

Expectation step

We define the probability of being in state i at time t , for the generic observed sequence $x^j = x_1^j, x_2^j, \dots, x_n^j$:

$$Pr(\pi_t = k \mid x^j, \theta) = \frac{Pr(x^j, \pi_t = k \mid \theta)}{Pr(x^j \mid \theta)}, \quad (2.57)$$

where the *Bayes' rule* has been applied,

$$\begin{aligned} Pr(\pi_t = k \mid x^j, \theta) &= \frac{Pr(x_1^j, \dots, x_t^j, x_{t+1}^j, \dots, x_n^j, \pi_t = k \mid \theta)}{Pr(x^j \mid \theta)} \\ &= \frac{Pr(x_1^j, \dots, x_t^j, \pi_t = k \mid \theta) Pr(x_{t+1}^j, \dots, x_n^j \mid \pi_t = k, \theta, x_1^j, \dots, x_t^j)}{Pr(x^j \mid \theta)}. \end{aligned}$$

Considering that x_{t+1}^j, \dots, x_n^j are conditional independent on x_1^j, \dots, x_t^j given π_t and θ , then,

$$\begin{aligned} Pr(\pi_t = k \mid x^j, \theta) &= \frac{Pr(x_1^j, \dots, x_t^j, \pi_t = k \mid \theta) Pr(x_{t+1}^j, \dots, x_n^j \mid \pi_t = k, \theta)}{Pr(x^j \mid \theta)} \\ &= \frac{Pr(x_{1:t}^j, \pi_t = k \mid \theta) Pr(x_{t+1:n}^j \mid \pi_t = k, \theta)}{Pr(x^j \mid \theta)}. \end{aligned}$$

Substituting the *forward* ([Expression 2.19](#)) and *backward* ([Expression 2.30](#)) variables, the expression will become:

$$Pr(\pi_t = k \mid x^j, \theta) = \frac{f_k^j(t) b_k^j(t)}{Pr(x^j \mid \theta)} \quad (2.58)$$

Concerning $Pr(x^j \mid \theta)$, it is possible to calculate through the *forward* or *backward* algorithm, furthermore $f_k^j(t)$ is the *forward* variable $f_k(t)$ defined above calculated for sequence j , while $b_k^j(t)$ is the equivalent *backward* variable.

The *expected number of times that we are in state k* and hence *the expected number of transitions away from the state k for x^j* is defined as:

$$\begin{aligned}
 E[\# \text{ of transitions from } k \text{ for } x^j] &= E\left[\sum_{t=1}^n S_t^j(k)\right] \\
 &= \sum_{t=1}^n E[S_t^j(k)] \\
 &= \sum_{t=1}^n Pr(\pi_t = k \mid x^j, \theta) \\
 &= \sum_{t=1}^n \frac{f_k^j(t) \cdot b_k^j(t)}{Pr(x^j \mid \theta)}, \text{ for } j = 1, \dots, J.
 \end{aligned} \tag{2.59}$$

Now, the probability of being in state k at time t and being in the state l at time $t + 1$, for the generic observed sequence $x^j = x_1^j, x_2^j, \dots, x_n^j$ can be defined as follows:

$$Pr(\pi_t = k, \pi_{t+1} = l \mid x^j, \theta) = \frac{Pr(x^j, \pi_t = k, \pi_{t+1} = l \mid \theta)}{Pr(x^j \mid \theta)}, \tag{2.60}$$

where the *Bayes' rule* has been applied,

$$\begin{aligned}
 Pr(\pi_t = k, \pi_{t+1} = l \mid x^j, \theta) &= \frac{Pr(x_1^j, \dots, x_t^j, x_{t+1}^j, \dots, x_n^j, \pi_t = k, \pi_{t+1} = l \mid \theta)}{Pr(x^j \mid \theta)} \\
 &= \frac{Pr(x_1^j, \dots, x_t^j, \pi_t = k, \pi_{t+1} = l \mid \theta) Pr(x_{t+1}^j, \dots, x_n^j \mid \pi_t = k, \pi_{t+1} = l, x_1^j, \dots, x_t^j, \theta)}{Pr(x^j \mid \theta)}
 \end{aligned}$$

considering that x_{t+1}^j, \dots, x_n^j are conditional independent on x_1^j, \dots, x_t^j and π_t given π_{t+1} , then

$$\begin{aligned}
&= \frac{Pr(x_{1:t}^j, \pi_t = k, \pi_{t+1} = l \mid \theta) Pr(x_{t+1:n}^j \mid \pi_{t+1} = l, \theta)}{Pr(x^j \mid \theta)} \\
&= \frac{Pr(\pi_{t+1}=l \mid x_{1:t}^j, \pi_t=k, \theta) Pr(x_{1:t}^j, \pi_t=k \mid \theta) Pr(x_{t+1}^j \mid \pi_{t+1}=l, \theta) Pr(x_{t+2:n}^j \mid \pi_{t+1}=l, \theta)}{Pr(x^j \mid \theta)}
\end{aligned}$$

where π_{t+1} is conditional independent on $x_{1:t}^j$ given π_t , then

$$= \frac{Pr(\pi_{t+1}=l \mid \pi_t=k, \theta) Pr(x_{1:t}^j, \pi_t=k \mid \theta) Pr(x_{t+1}^j \mid \pi_{t+1}=l, \theta) Pr(x_{t+2:n}^j \mid \pi_{t+1}=l, \theta)}{Pr(x^j \mid \theta)}. \quad (2.61)$$

Here, we have that

$$\begin{aligned}
Pr(\pi_{t+1} = l \mid \pi_t = k, \theta) &= a_{kl}, \\
Pr(x_{1:t}^j, \pi_t = k \mid \theta) &= f_k^j(t), \\
Pr(x_{t+1}^j \mid \pi_{t+1} = l, \theta) &= e_l(x_{t+1}^j), \\
Pr(x_{t+2:n}^j \mid \pi_{t+1} = l, \theta) &= b_l^j(t+1),
\end{aligned} \quad (2.62)$$

then,

$$Pr(\pi_t = k, \pi_{t+1} = l \mid x^j, \theta) = \frac{f_k^j(t) \cdot a_{kl} \cdot e_l(x_{t+1}^j) \cdot b_l^j(t+1)}{Pr(x^j \mid \theta)}. \quad (2.63)$$

Now, for $1 \leq k, l \leq N$ and $j = 1, \dots, J$, the *expected number of transitions from state k to state l for the observed sequence x^j* , it can be obtained as

follows:

$$\begin{aligned}
E [\# \text{ of transitions from } k \text{ to } l \text{ for } x^j] &= E \left[\sum_{t=1}^{n-1} S_t^j(k, l) \right] \\
&= \sum_{t=1}^{n-1} E [S_t^j(k, l)] \\
&= \sum_{t=1}^{n-1} Pr(\pi_t = k, \pi_{t+1} = l \mid x^j, \theta) \\
&= \sum_{t=1}^{n-1} \frac{f_k^j(t) \cdot a_{kl} \cdot e_l(x_{t+1}^j) \cdot b_l^j(t+1)}{Pr(x^j \mid \theta)} \quad (2.64)
\end{aligned}$$

Considering all the observed sequences of the training set \mathbf{X} , we will obtain:

$$\begin{aligned}
A_{kl} &= E [\# \text{ of transitions from } k \text{ to } l \text{ for } \mathbf{X}] \\
&= \sum_{j=1}^J \frac{1}{Pr(x^j \mid \theta)} \sum_{t=1}^{n-1} f_k^j(t) \cdot a_{kl} \cdot e_l(x_{t+1}^j) \cdot b_l^j(t+1). \quad (2.65)
\end{aligned}$$

Furthermore, considering the summation over all possible l' , it will be

$$\begin{aligned}
\sum_{\text{all } l'} A_{kl'} &= \sum_{\text{all } l'} \sum_{j=1}^J \frac{1}{Pr(x^j \mid \theta)} \sum_{t=1}^{n-1} f_k^j(t) \cdot a_{kl'} \cdot e_{l'}(x_{t+1}^j) \cdot b_{l'}^j(t+1) \\
&= \sum_{j=1}^J \frac{1}{Pr(x^j \mid \theta)} \sum_{t=1}^n f_k^j(t) \cdot b_k^j(t). \quad (2.66)
\end{aligned}$$

Equally, *the expected number of times that letter v appears in state k :*

$E_k(v) = E[\# \text{ of time in } k, \text{ when the observation was } v \text{ for } \mathbf{X}]$

$$= \sum_{j=1}^J \frac{1}{Pr(x^j | \theta)} \sum_{t=1}^n f_k^j(t) \cdot b_k^j(t) \cdot I_v^j(t), \text{ for } 1 \leq k \leq N, v \in M. \quad (2.67)$$

The inner sum is only over those positions t for which the symbol emitted is v .

Finally, the *expected number of times in state k at time $t = 1$* can be calculated, considering all training set of observed sequences:

$$I_k = E[\# \text{ number of times in state } k \text{ at time } t = 1 \text{ for } \mathbf{X}] \quad (2.68)$$

$$= \sum_{j=1}^J \frac{1}{Pr(x^j | \theta)} \sum_{t=1}^n f_k^j(1) \cdot b_k^j(1).$$

Maximization step (*MLE*)

Based on the probability estimates and the expectations calculated above, the *new model*, $\hat{\theta} = \{\hat{a}_{0k}, \hat{e}_k(\cdot), \hat{a}_{kl}\}$, can be constructed as follows:

- The *new initial state distribution*:

$$\hat{a}_{0k} = \frac{\text{expected number of times in state } k \text{ at time } (t = 1)}{\# \text{ of the observed sequences}} \quad (2.69)$$

$$= \frac{1}{J} \sum_{j=1}^J \frac{1}{Pr(x^j | \theta)} \sum_{t=1}^n f_k^j(1) \cdot b_k^j(1) = \frac{I_k}{J}.$$

- The *new emission probability distribution*:

$$\begin{aligned}
\hat{e}_k(v) &= \frac{\text{expected number of times in state } k \text{ and observing symbol } v}{\text{expected number of times in state } k} \\
&= \frac{\sum_{j=1}^J \frac{1}{Pr(x^j|\theta)} \sum_{t=1}^n f_k^j(t) \cdot b_k^j(t) \cdot I_v^j(t)}{\sum_{j=1}^J \frac{1}{Pr(x^j|\theta)} \sum_{t=1}^n f_k^j(t) \cdot b_k^j(t)} \\
&= \frac{E_k(v)}{\sum_{\text{all } v'} E_k(v')}, \tag{2.70}
\end{aligned}$$

where, $\hat{e}_k(v)$ is the *expected number of times that the output observations have been equal to v while in the state k relative to the expected total number of times in state k* .

- The *new transition probability distribution*:

$$\begin{aligned}
\hat{a}_{kl} &= \frac{\text{expected number of transitions from state } k \text{ to state } l}{\text{expected number of transitions from state } k} \\
&= \frac{\sum_{j=1}^J \frac{1}{Pr(x^j|\theta)} \sum_{t=1}^{n-1} f_k^j(t) \cdot a_{kl} \cdot e_l(x_{t+1}^j) \cdot b_k^j(t+1)}{\sum_{j=1}^J \frac{1}{Pr(x^j|\theta)} \sum_{t=1}^n f_k^j(t) \cdot b_k^j(t)} \\
&= \frac{A_{kl}}{\sum_{\text{all } l'} A_{kl'}}, \tag{2.71}
\end{aligned}$$

where, \hat{a}_{kl} is the *expected number of transitions from state k to state l relative to the expected total number of transitions away from state k* .

Sometimes when there are insufficient data, it may happen that a state k is never used in the training set of the sequences, then the estimation equations are undefined for that state, because both the numerator and denominator will have value zero. In that cases, in order to solve them it is preferable to add predetermined pseudocounts to the $e_k(v)$ and a_{kl} , then *Equations 2.70 and 2.71* will become:

$$\begin{aligned}\widehat{e}_k(v) &= \frac{E_k(v) + r_{kl}}{\sum_{\text{all } v'} E_k(v')}, \\ \widehat{a}_{kl} &= \frac{A_{kl} + r_k(b)}{\sum_{\text{all } l'} A_{kl'}}.\end{aligned}\tag{2.72}$$

The pseudocounts r_{kl} and $r_k(b)$ should reflect the prior biases about the probability values and for that they have a probabilistic interpretation as the parameters of *Bayesian Dirichlet prior distributions* (As will see in 2.8).

Pseudo-Code of the *Baum-Welch* algorithm**Initialization;**

Randomly initialization of emission and transition matrices (or according to the knowledge, so that the properties are observed);

Iteration;

Calculating of the *Forward* probabilities (f);

Calculating of the *Backward* probabilities (b);

Calculating of the matrices A , E using (f, b) for each X^i ;

Calculating of the new parameters a_{tl} , $e_t(b)$;

Calculating of the new log-likelihood $P(X^i, \dots, X^m)$;

EXPECTATION-MAXIMIZATION step;

Until the log-likelihood $P(X^i, \dots, X^m)$ does not change compared to the previous value.

Algorithm 5: Baum-Welch algorithm

2.8 The Dirichlet distribution

In Bayesian statistics it is necessary to use distributions over probability parameters as prior distribution. In those cases certainly the choice falls on the *Dirichlet* distribution. The *Dirichlet* distribution is a family of continuous multivariate probability distributions parametrized by a vector α :

$$\mathcal{D}(\theta \mid \alpha) = Z^{-1}(\alpha) \prod_{i=1}^K \theta_i^{\alpha_i - 1} \delta\left(\sum_{i=1}^K \theta_i - 1\right), \quad (2.73)$$

where $\alpha = \alpha_1, \dots, \alpha_K$, $\alpha > 0$, and the θ_i satisfy $0 \leq \theta_i \leq 1$ and sum to

1, being indicated by the delta function term $\delta(\sum_{i=1}^K \theta_i - 1)$.

The normalising factor Z in Equation 2.73 can be expressed in term of *gamma function*:

$$Z(\alpha) = \int \prod_{i=1}^{\alpha_i-1} \delta(\sum_{i=1}^K \theta_i - 1) d\theta = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(\sum \alpha_i)}. \quad (2.74)$$

When $K = 2$ (then for two variables) the *Dirichlet* distribution reduces to a *beta* distribution, and in particular, the normalising constant is the *beta* function.

A special case of the *Dirichlet* distribution is when all the elements of vector α have the same value. This kind of distribution is called *symmetric Dirichlet* distribution. When $\alpha = \mathbf{1}$ then *symmetric Dirichlet* distribution is equivalent to a uniform distribution.

2.8.1 Mixtures of Dirichlets

Sometime it is not very easy to express all the prior knowledge about a problem with a single *Dirichlet* distribution [83, 84]. In order to solve that it is mandatory to use several different *Dirichlet* distributions, then consider a *mixture of Dirichlet* distributions.

A *Dirichlet mixture* consists of m components, associated respectively with vectors $\alpha^1, \dots, \alpha^m$. A mixture prior with m components may be written as:

$$Pr(\theta \mid \alpha^1, \dots, \alpha^m) = \sum_k q_k \mathcal{D}(\theta \mid \alpha^k). \quad (2.75)$$

Here, q_k are called the *mixture coefficients*. In order for the mixture to be

2.9. ESTIMATORS FOR ESTIMATION PROBLEMS IN DISCRETE HIGH-DIMENSIONAL SPACES

a proper probability distribution, the *mixture coefficients* must be positive and sum to be one. Finally, it is possible to identify q_k as the *prior* probability $q_k = Pr(a^k)$ of each of the mixture coefficients.

The density of a *Dirichlet* mixture is defined to be a linear combination of those of its constituent components.

2.9 Estimators for estimation problems in discrete high-dimensional spaces

In the last few years we have seen a huge increase of biological dataset due to the advent of high-throughput data-acquisition technologies. If on the one hand the biological data are grown, on the other hand the parameter estimation is increasingly becoming a difficult problem to overcome.

The *Maximum-Likelihood* (**ML**) estimators are very useful to identify the most probable point in a space of the *unknowns* and definitely for several years they dominated statistical estimation and prediction. Two of the most import examples of estimators belonging to the family of *maximum likelihood estimators* are the minimum “*free-energy*” structure predictions and the *Viterbi decoding* hidden Markov models (as seen in 2.7.4).

They have three principal properties: *consistency*, *asymptotic normality*, and *asymptotic efficiency*. Although these properties only hold asymptotically, in high-D space they are not reached. In fact, in these cases ML estimators do not always give us a true estimation, because in the large sample space the most probable estimator has very low probability.

2.10 Centroid Estimation

In order to find an estimator which overcomes the limitations of ML estimators, centroid estimators [85, 86] are proposed. They have been proposed in studies, such as *sequence alignment* and *RNA secondary structure prediction*, that can be formalized in a *high-dimensional binary space* as follows [87]:

Problem 1. Prediction of secondary structures of RNA sequences

Given an RNA sequence x , predict its secondary structure as a point in $S(x)$, the space of all the possible secondary structures of x .

Problem 2. Pairwise alignment of two biological sequences

Given a pair of biological sequences (DNA, RNA, protein) x and x' , predict their alignment as a point in $A(x, x')$, the space of all the possible alignments of x and x' .

Here, the predictive space $S(x) \subset \{0, 1\}^{\frac{|x|(|x|-1)}{2}}$ (in Figure 2.5 is shown an example of a point in predictive space), while $A(x, x') \subset \{0, 1\}^{|x||x'|}$.

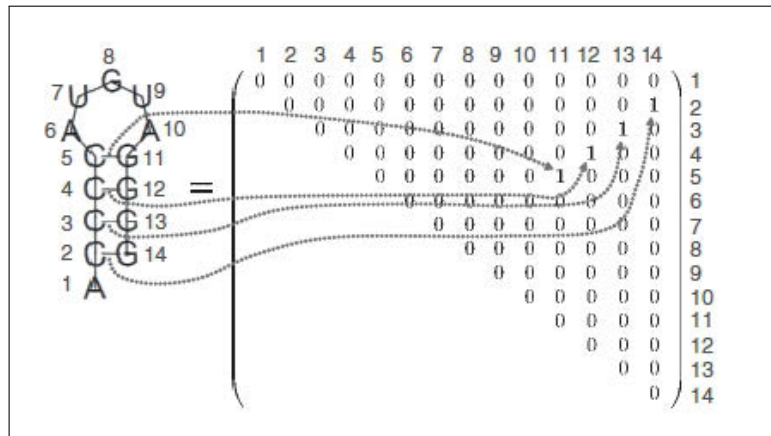


Figure 2.5: (A binary matrix representation of a secondary structure of an RNA sequence [88])

Furthermore, these two problems are particular examples of a more general problem [87]:

Problem 3. *Estimation problem on a binary space*

Given a data set D and a predictive space Y (a set of all candidates of a prediction), which is a subset of n -dimensional binary vectors $\{0, 1\}^n$, that is, $Y \subset \{0, 1\}^n$, predict a point y in the predictive space Y .

To find an estimator that is drawn in a high probability region, it is used a loss function, that increases with the distance from other members of the ensemble. In particular, the *Hamming* distance is used as *loss function*:

$$H(z, y) = \sum_{i=1}^n I(z_i \neq y_i) = n - \sum_{i=1}^n I(z_i = y_i), \quad (2.76)$$

where I is the *indicator function*,

$$I(a) = \begin{cases} 1 & \text{if } a \text{ is true,} \\ 0 & \text{otherwise} \end{cases} \quad (2.77)$$

and n is the dimension of both z and y .

The *posterior* risk of the *Hamming* loss function for some estimator $\hat{\theta}$, considering the data D , is

$$\begin{aligned}
\rho_H(\hat{\theta}(D)) &= E_{\theta|S}[H(\theta, \hat{\theta}(D))] \\
&= \sum_{\theta \in \Theta} H(\theta, \hat{\theta}(D)) Pr(\theta | D) \\
&= \sum_{\theta \in \Theta} \sum_{i=1}^n I(\theta_i \neq \hat{\theta}_i(D)) Pr(\theta | D) \\
&= n - \sum_{\theta \in \Theta} \sum_{i=1}^n I(\theta_i = \hat{\theta}_i(D)) Pr(\theta | D) \\
&= n - \sum_{i=1}^n Pr(\theta_i = \hat{\theta}_i(D) | D)
\end{aligned} \tag{2.78}$$

In order to minimize the risk, it is possible to apply the *posterior marginal sum* maximizer, choosing:

$$\hat{\theta}_c(D) = \arg \max_{\hat{\theta} \in \Theta} \sum_{i=1}^n Pr(\hat{\theta}_i(D) = \hat{\theta}_i(D) | D), \tag{2.79}$$

where $\hat{\theta}_c$ is defined as *centroid estimator*.

In the *Figure 2.6*, it is shown an example of centroid estimation in two different clusters that are in *posterior space*. In particular, the *green* dots indicate the cluster centroids, the *red* dot represents the ensemble centroid, while the *blue* dot is the minimum free-energy structure.

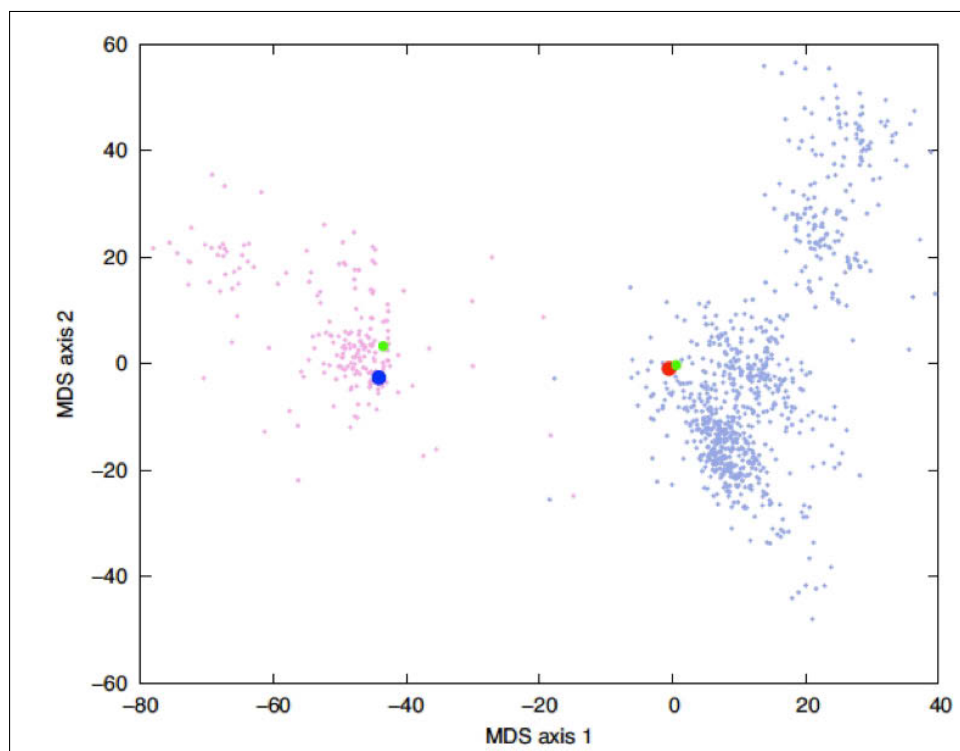


Figure 2.6: Multidimensional scaled distribution derived from 1,000 representative samples from Sfold from the secondary structure of *Dermocarpa sp. Ribonuclease P RNA* [86].

2.11 Gamma Centroid Estimator

The aim of the estimation is not always to find the exact solution with a very small probability (as in the case of ML estimator) or to find the solution with the minimum *Hamming* loss function (as seen in 2.10), but rather to find the *most accurate estimator*. For this reason, the concept of *maximum expected accuracy* (**MEA**) has been adopted in several bioinformatic problems, as in the case of CONTRAfold [89] for the secondary structure prediction. Unfortunately, the theoretical analysis has been shown that this kind of estimators

sometimes are not robust with respect to accuracy measures, and then the γ -centroid estimator has been proposed as solution for some specific problem.

The γ -centroid estimator represents a generalization for the *centroid estimator*, in particular it maximizes the expectation of $\gamma \cdot TP + TN$ ⁷.

2.11.1 Evaluation measures defined using TP , TN , FP and FN

There are different accuracy measures, but certainly standard and traditional ones are the *Sensitivity* (**SEN**), the *positive predictive value* (**PPV**), and the *Matthew's correlation coefficient* (**MCC**) that can be defined by using TP , FP , TN , FN :

$$\begin{aligned} SEN &= \frac{TP}{TP + FN}, \\ PPV &= \frac{TP}{TP + FP}, \\ MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \end{aligned} \tag{2.80}$$

in which, TP , FP , TN , FN are accuracy measures functions that are defined as follows:

⁷ TP is the number of the *true positives* and TN is the number of the *true negatives*

$$\begin{aligned}
TP &= TP(\theta, y) = \sum_i I(y_i = 1)I(\theta_i = 1), \\
FP &= TP(\theta, y) = \sum_i I(y_i = 1)I(\theta_i = 0), \\
TN &= TP(\theta, y) = \sum_i I(y_i = 0)I(\theta_i = 1), \\
TN &= TP(\theta, y) = \sum_i I(y_i = 0)I(\theta_i = 1).
\end{aligned} \tag{2.81}$$

Here, considering the *Problem 1*, then $\theta, y \in S(x)$, where θ is *correct (reference)* structure, while y is *predicted* secondary structure. For this problem when the sensitivity is equal to 1, the predicted structure contains all the correct base pairs (it is not excluded that it can also contain some base pairs) and when PPV is equal to 1, the predicted secondary structure will contain only the correct base pairs. Then, if both the sensitivity and PPV are equal to 1 the perfect prediction will be achieved.

2.11.2 Formalization of γ -centroid estimators

In order to design an estimator that optimizes the expected numbers of TP , TN , FP and FN with respect to the entire distribution $Pr(\theta \mid x)$, we can consider a *gain function* which yields *positive scores* for the number of true predictions (TP and TN) and negative scores for those of *false predictions* (FP and FN). Let $\overline{G}(\theta, y)$ be a linear combination for the accuracy measure functions (see in *Equations 2.81*):

$$\overline{G}(\theta, y) = \alpha_1 TP + \alpha_2 TN - \alpha_3 FP - \alpha_4 FN \quad (2.82)$$

where $\alpha_k > 0$ ($k = 1, 2, 3, 4$).

By using the identity $I(y_i = 0) + I(y_i = 1) = 1$ (by considering the *Equations 2.81*), it can be written that:

$$\overline{G}(\theta, y) = C_0 G_\gamma^{(c)}(\theta, y) + C_\theta, \quad (2.83)$$

where $G_\gamma^{(c)}(\theta, y) = \gamma TP + TN$ is the *gain function* of the γ -centroid estimators, $C_0, \gamma > 0$ are constants, and C_θ is a function of θ independent of y . Then the aim is that to design an estimator that predicts \hat{Y} that maximizes the expectation value of $G_\gamma^{(c)}(\theta, y)$ (as said above) with respect to $Pr(\theta | x)$,

$$\hat{Y} = \arg \max_{y \in S(x)} E_{\theta|x}[G_\gamma^{(c)}(\theta, y)] \quad (2.84)$$

$$E_{\theta|x}[G_\gamma^{(c)}(\theta, y)] = \sum_{\theta \in S(x)} G_\gamma^{(c)}(\theta, y) Pr(\theta | x). \quad (2.85)$$

The γ -centroid parameter is able to adjust the sensitivity and PPV of the prediction \hat{Y} . In particular, estimators with larger γ values produce *better* sensitivities (and *smaller* PPVs), while those with smaller γ values produce better PPVs (smaller sensitivities). Furthermore, the γ -centroid estimator is equivalent to the *centroid estimator* when $\gamma = 1$.

The *Equation 2.85* can be rewritten as follows:

$$E_{\theta|x}[G_{\gamma}^{(c)}(\theta, y)] = \sum_{i=1}^n [(\gamma + 1)p_i - 1]I(y_i = 1) + \sum_{i=1}^n (1 - p_i) \quad (2.86)$$

where,

$$p_i = Pr(\theta_i = 1 \mid D) = \sum_{\theta \in \Theta} I(\theta_i = 1) Pr(\theta \mid D). \quad (2.87)$$

Here, p_i is the marginalized probability of the distribution for the i -th dimension of the predictive space. The second term in 2.86 is a constant which does not depend on y , while the γ -centroid estimator maximizes the first term. Furthermore, considering the *Theorem 3* in [87], the γ -centroid estimator is equivalent to the estimator that maximizes the sum of marginalized probabilities p_i that are greater than $\frac{1}{(\gamma+1)}$ in the prediction.

Finally, taking into account the *Corollary 1* in [87], we have that the γ -centroid estimator for $\gamma \in [0, 1]$ contains its estimator $\hat{y} = \{\hat{y}_i\}$:

$$\hat{y} = \begin{cases} 1 & \text{if } p_i > \frac{1}{(\gamma+1)}, \\ 0 & \text{if } p_i \leq \frac{1}{(\gamma+1)} \end{cases} \text{ for } i = 1, 2, \dots, n. \quad (2.88)$$

Chapter 3

MicroRNA Biogenesis and RNA Editing Phenomenon

*We will have drugs based on
microRNA, and a lot of novel
diagnostic and prognostic
markers will be developed. It will
be a revolution*

Carlo M. Croce

Ohio State University

3.1 microRNA Biogenesis

MICRORNAS (miRNAs) are a large class of small non-coding RNAs of about 21-25 nucleotides. They negatively regulate the gene expression at the post-transcriptional level, inducing the *degradation* of specific messenger

RNA (mRNA), or preventing the translation into protein [90, 91, 92]. In particular, miRNAs recognise specific mRNA target in order to determine the *degradation* or the *repression* of the translation.

On a functional point of view, not only there are numerous miRNAs capable of recognizing more than one target, but also many of these targets can be regulated by different miRNAs. MiRNAs can be considered as small control elements of more complex regulatory biological *pathways*, that are the basis of several fundamental functions. Those functions can be related both to the cell processes (such as *cell cycle regulation*, *cell proliferation*, and *apoptosis*), and the ones relating to the entire organism (including embryonic development and its immune response, metabolism and cariogenesis, development and function of the nervous and immune system) [93, 94].

Even if, thanks to their temporal and spatial *expression patterns*, miRNAs perform their task in a physiological way, when their expression is altered, they may be involved in several complex diseases, including numerous cancers. Moreover, the miRNA may have specific expression profiles according to stage development, tissues and various pathologies. This implies that each tissue is characterized from a specific set of miRNAs, whose expression profile is distinctive of that tissue [95].

Recently, numerous studies have demonstrated the involvement of miRNAs in different cardiovascular [96] and neurological [97] diseases, but also in obesity, diabetes [98], and especially in cancer [99].

3.1.1 Organization of microRNA in Human Genome

As seen in 1.2.1, all miRNAs discovered, either through an experimental approach or a computational analysis, are stored in miRBase database (miRBase 2013), a register of miRNAs, noting all their features [32, 33]. Moreover, miRBase expresses information about the genomics of miRNAs, that is their organization on human chromosomes. In fact, the genes that encode miRNAs are distributed on all chromosomes, except for the Y chromosome, and genes for different miRNAs are mostly located closer. This creates real *clusters*, often related to each other also on a functional point of view, as it happens in the cluster of miRNAs *hsa-let-7a-1*, *hsa-let-7f-1*, and *hsa-let-7d* in chromosome 9. According to their localization, miRNAs in the genome can be distinguished in:

- *Intergenic MiRNAs*;
- *Intronic miRNAs in encoding transcripts*;
- *Intronic miRNAs in non-coding transcripts*;
- *Esonic MiRNAs in non-coding transcripts*.

3.1.2 MicroRNA Biogenesis

MicroRNAs are in plants, in eukaryotes and in some viruses, and are encoded by different types of genes. Those molecules, that are active in the regulation of their target mRNAs, are defined “*mature*” miRNAs. These small RNAs are between 19 and 22 nucleotides of length and are formed by processing bigger ribonucleic sequences, encoded by the genome itself. MiRNAs are encoded by

genes located in genome, individually or in clusters [100]. Generally, about 70% of the genes for miRNAs are in the intergenic regions, while the residual 30% is located in intronic sequences of specific genes, called “*guest*”. This means that miRNAs can be located in transcriptional and independent units, even if a large number of them is generated from transcripts containing either cluster of miRNAs or intronic sequences of the *host* gene.

The process to get to the formation of mature miRNAs is complex; it has its origins in the nucleus and finishes at the cytoplasm level (see *Figure 3.1* [101]).

This molecule undergoes two consecutive reactions, catalysed by two endonucleases, ***Drosha*** and ***Dicer***. Both of them act within protein complexes, containing domains capable of binding together double-stranded RNA molecules (*dsRNA binding domains* - *dsRBDs*). The proteins Drosha and Dicer both possess the specific conserved catalytic domains ***RNase type III*** (*RIIIda*, *RIIIDb*), which act generating 3' extremity, protruding for two nucleotides.

The first reaction occurs in the nucleus thanks to Drosha, which forms a complex with the ***Pasha*** protein (*DGCR8*). The union of Drosha and Pasha cleaves the *pri-miR*, and generates a molecule with a hairpin structure of about 70 nucleotides, called *pre-miR*. After that, the pre-miR is moved from the nucleus to the cytoplasm by the ***Exp5***. Once inside the cytoplasm, the hairpin precursor is cut by Dicer, forming a small duplex RNA molecule of a variable size between 21 and 25 nucleotides. This molecule contains both the filament of the *mature miRNA* and its complementary strand [101].

In mammals, the Dicer interacts with the proteins ***Ago1-Ago4***. When

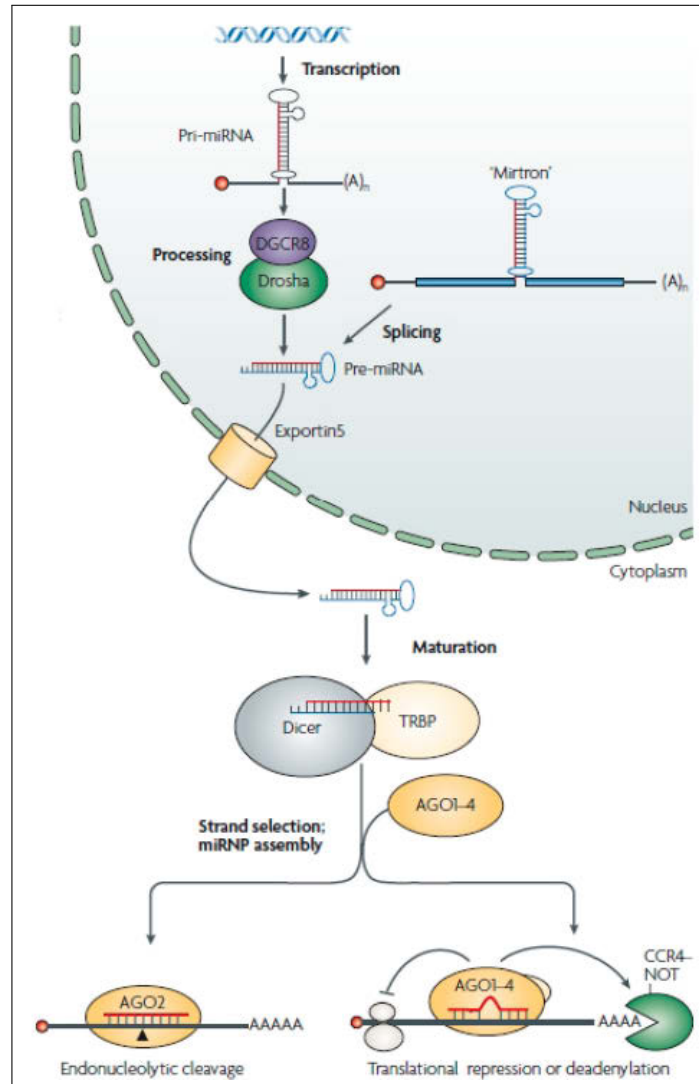


Figure 3.1: Model for biogenesis and activity of transcriptional repression of microRNAs.

combined with the mature miRNA, they form the **miRISC** complex (*microRNA Induced Silencing Complex*), which has the aim to drive the miRNA to the recognition of target messenger. The Ago proteins belong to the *Argonaute* family, which are in all the eukaryotes and also are equipped for specific motifs: **PAZ** and **MID** are the domains for the anchoring of the target RNA at 3' and 5', while **PIWI** is the domain for cutting.

In the miRISC complex, there are not only Ago proteins, but also proteins belonging to the **GW182** family (*TNRC6A*, *TNRC6B*, *TNRC6C* in mammals). They have an important role in transcriptional repression mediated by miRNA [101] and act as cofactors of the Ago. As soon as the duplex is formed by the action of Dicer, thermodynamically, the two filaments have a different stability at the extremity 5'. Although the mature miRNA can be localized either in a filament or in the other, it is almost always originates from the filament with a more unstable 5' extremity, while the other filament is degraded [102]. It can rarely happen that the two 5' extremities have a similar stability, so each of the them can create a mature miRNA with biological activity with equal probability [90].

The regulation of the biogenesis of miRNAs is very important but not studied in a complete way. However, there is a significant trend: a surprising number of miRNA genes are formed under the control of many target that it regulate. For example, the transcription of *miR-7* gene in *Drosophila* is repressed by a transcription factor called *Yan*, whose translation is in turn repressed by *miR-7*, resulting in a negative *feedback-loop* [103]. Another example is in *C. Elegans*: here the miRNA *let-7* inhibits the translation of *lin-28* which in turn inhibits the transcription of *let-7* [104].

3.1.3 Post-Transcriptional Regulation Mediated by mi-croRNAs

Once the formation of *miRISC*, which contains the *miR-mature*, is completed, it pairs with the mRNA target. Then, miRNAs act as the adapters for the miRISC complexes to recognize certain target mRNA. miRNA's binding sites in animal mRNAs are located in the 3' UTR and are usually expressed in multiple copies¹ [105]. However, it has been observed in vitro that the recognition *miRNA-mRNA* might also take place either in the coding regions or in the 5' UTR of mRNA, even if these pairing sites would not have enough silencing capacity and could play only a marginal role [106]. The majority of animal miRNAs bind together their target with imperfect complementarity, forming bulges and loops, although a key feature of the target recognition involves the pairing of nucleotides 2-8 of miRNA, representing the *seed* region. On the other side, in most plants miRNAs bind with almost perfect complementarity to specific sites in the coding region [100].

The level of complementarity between miRNA and target is a key factor in the regulatory mechanism. While the perfect complementarity allows the cutting of the mRNA's filament catalyzed by Ago, the central mismatch of the duplex *miRNA/mRNA* excludes the cut and promotes the repression of the translation.

Several studies on miRNAs of the animals indicate that the repression of the translation is not followed by the destabilization of the mRNA. However, for some interactions miRNA-target there is a significant reduction of the

¹This is a necessary condition to have an efficient repression of translation

concentration of the mRNAs due to a increase of the degradation [107, 108] (see *Figure 3.2* [101]). At the moment, it is not clear why this degradation happens only in some of the targets and not in others. An hypothesis can be that it is related to the number, the type and the location of the mismatch of the duplex miRNA/mRNA, playing an important role in determining the degradation or the arrest of the translation [109].

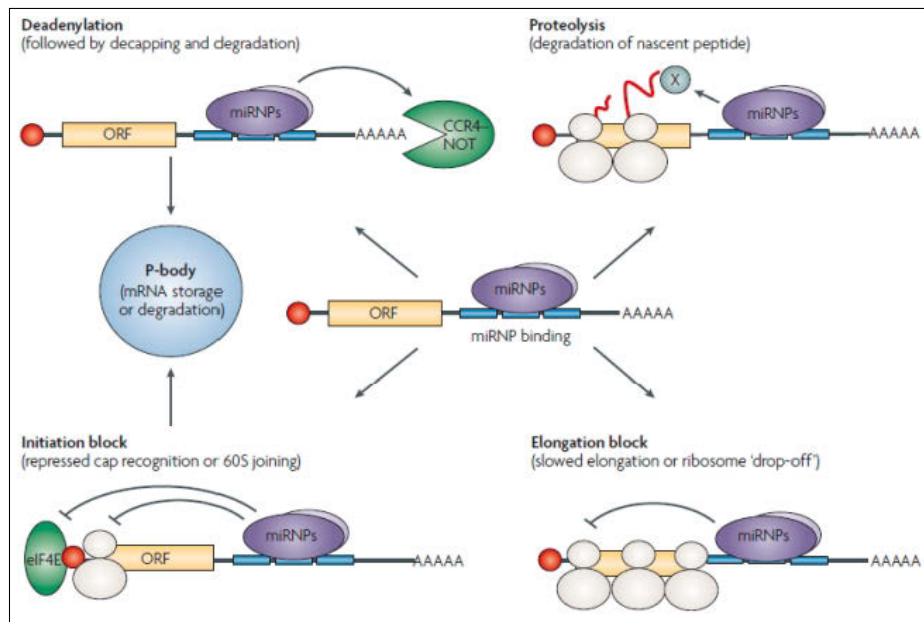


Figure 3.2: Representation of some of the possible mechanisms of action of the RISC complex induced by miRNA. The mRNA target can be deadenylated and degraded, or its translation may be inhibited in several ways represented here.

3.1.4 Regulation of microRNA Expression

The regulation of miRNA expression is fundamental to the role that these molecules play. They are regulated in various levels during their biogenesis:

- *Regulation of transcription;*
- *Adjusting the processing;*
- *Editing;*
- *microRNA decay.*

The regulation of the transcription is the same as many coding genes, thanks to the presence of the same regulatory elements (*TATA box sequences*, *CpC Islands*, *inization elements*) at the level of the promoters of miRNAs.

Many *transcription factors* (**TF**) regulate the expression of *tissue-specific* or *development-specific*, such as *MYC/MYCN*, stimulating the expression of oncogenic cluster *miR-17-5p* in lymphoma cells [110], or *REST* that inhibits the expression of *miR124* in non-neuronal cells or neuronal progenitors through histonic de-acetylation and methylation of the promoter [111]. In turn, miRNAs can regulate the expression of TFs, creating in this way circuits of positive or negative adjustment. Here, the total control of either miRNA's quantity or of TFs determines the final physiological effect.

The regulation of miRNAs processing occurs in several levels: Drosha, Dicer and their accessory proteins. For example, some helicases of the rat and the *SMAD* proteins act at the Drosha level, controlling the production of pri-miR [112], while at the pre-miR level, the levels of Dicer are controlled and stabilized by *TRBP*, its cofactor [113].

As it will be explained more deeply in 3.2.4, the editing phenomenon of both the pri-miR and the pre-miR by ADAR proteins² would alter the

²The ADAR catalyses the conversion of adenosine to inosine.

secondary structure (and therefore its stability), while the editing of the seed of mature miRNAs by other proteins would alter the recognition of the target [112].

Finally, the regulation of the stability and degradation of the mature miRNA can control the final quantity inside the cell and then the biological effect. It was observed that miRNAs are generally more stable than second class messengers and have a half-life, ranging from a few hours to many days. A complete control of the decay may have a fundamental role in the development mechanisms and in the switch on-off response type, as, for example, in the development of the mice's retina, where the levels of miR-204 and miR-211 decrease rapidly in neurons but not in the glia [112].

3.1.5 Bioinformatics Prediction of microRNAs' Molecular Targets

The discovery of miRNAs introduced a new paradigm in the gene regulation systems. A primary point in understanding the functional role of miRNAs and the complex molecular networks at the base of the gene regulation is the identification of genes regulated by miRNAs themselves. miRNA sequences are very short and are characterized by a pairing imperfection with the molecular target, so this creates a complexity in the identification of the mRNA targets of the miRNAs. In recent years, the basic principles of this interaction have been extrapolated from experimental studies, in order to develop numerous mathematical algorithms for the prediction in silico of the hypothetical target mRNAs. They include:

Tool	Website	References
TargetScan	http://www.targetscan.org/	[58]
PicTar	http://pictar.mdc-berlin.de/	[60]
DIANA-microT	http://www.diana.pcbi.upenn.edu/cgi-bin/micro_t.cgi	[59]
miRanda	http://www.microrna.org/	[114, 115]
MirTarget2	http://mirdb.org/	[116]
RNAhybrid	http://bibiserv.techfak.unizbielefeld.de/rnahybrid/	[117]

Table 3.1: List of some of the most important predictors of miRNA targets.

- Imperfect complementarity between the miRNA and the 3' UTR of the target and the strong bond between 6-8 nucleotides of the *seed* region at the 5' than at the 3'.
- Evolutionary conservation among the species of the target sequences at the 3' of the target.
- Thermodynamic stability of the duplex *miRNA-mRNA*.
- Cooperativity between multiple sites in close proximity.
- Multiplicity and cooperativity of the *miRNA-target* interaction.
- Loss of the secondary structure of the mRNA target at the binding site for the miRNA.

As seen in 1.4.7, there are several softwares for the prediction of miRNA targets available online. In the Table 3.1 are listed some of most known predictors of miRNA targets. Some of the results obtained with these algorithms have been experimentally validated, and this has allowed to significantly improve the performance for the in silico prediction of miRNA targets.

3.1.6 Circulating microRNAs

MicroRNAs are also located in *extra-cellular* human body fluids such as *serum*, *plasma*, *saliva* and *urine*, and this is often associated with various pathological conditions including cancer. The circulating microRNAs have been found within vesicles called *exosomes*. However, most of them are in plasma and human serum, complexed to the protein *Argonaute2* (**Ago2**), rather than within vesicles. MicroRNAs circulate in the bloodstream in an extra-cellular highly-stable form, which means that they could be used as biomarkers for non-invasive diseases such as cancer [118, 119], cardiovascular diseases [120] and pediatric *Crohn's disease* [121].

In the dominant model for the stability of circulating miRNAs, miRNAs are released by the cells in vesicles constituted by membranes, protecting them from the *Rnase* activity in the blood. The vesicles acting as *carriers* of circulating miRNAs include exosomes, which are vesicles of 50-90-nm hailing from multi-vesicular bodies and released by exocytosis [122]. However, it has been shown that a significant portion of miRNA circulating in human plasma and serum is associated to Ago2 [123, 124].

Ago2 is the essential component of the complex *miRISC*. It not only binds directly miRNAs, but also mediates the repression of the mRNA [125, 126]. Although it has been speculated that the miRNAs that are found in exosomes are involved in intercellular communication [127, 128], many extra-cellular miRNAs could be derived from dead cells that remain in the *extra-cell* space due to both the high stability of the protein Ago2 and the Ago2-miRNA complex [124]. These recent results suggest that the analysis of miRNAs as

biomarkers should include all types of circulating miRNAs found in biological fluids.

3.2 RNA Editing Phenomenon

3.2.1 The birth of RNA Editing

In 1986, a process that became known as RNA editing was discovered by the research group of Benne Rob [129]. They found a post-transcriptional process in which mitochondrial messenger RNAs were altered by the insertion and deletion of uridine³. This phenomenon is explained by the fact that the mitochondrial genomes of protozoa encoded a small number of proteins and many of those genes which either showed disruption of **ORF** (*Open Reading Frames*) or even did not have the start codon of trascription⁴.

According to what was known by the scientific community about nucleotide modification in RNA and alternative splicing of mRNA, it was impossible to explain that mitochondrial mRNA contained insertions of one or more non-genomically encoded uridines, without any flanking consensus sequence at the site of insertion.

Even if a the beginning of 1990s the editing was described in diffent species, the real interest for this phenomenon started around 1994. In fact, from this year on, several international conferences were organized by Harold

³The *uridine* is a nucleoside that consists from the pyrimidine base of *uracil* to which is attached a *ring of ribose*. If the uracil is attached to a deoxyribose ring, you get a molecule of deoxyuridine.

⁴**AUG** codon encodes the *methionine*.

Smith and Steve Hajduk⁵, by Glenn Bjork, Ted Maden and Henri Grosjean⁶, and by Paul Sloof and Rob Benne⁷. In 1993 Rob Benne wrote the first text dedicated to the theme of RNA editing [130]. In 1997, the inaugural *Gordon Research Conference* was dedicated to the modification and RNA editing [131]. This rapid growth has shown how the mechanisms of RNA and DNA editing are important for the biological phenomena present in the cell.

The Biological Phenomenon of RNA Editing

RNA editing is a process in which the nucleotide sequence of RNA is altered from the genomic code. The editing is related to the *insertion/deletion* of nucleotides, or the *base modification*. Its peculiarity is that the result of RNA editing is a change in the diversity and/or abundance of proteins expressed in the proteomes of organisms, in particular in their tissues or organelles.

The coordination of the activities of the editing is fundamental to other cellular pathways involving RNA, as, for example, *transcription*, *processing* and *translation*. There are different factors involved in the recognition of the RNA substrate and in the catalysis⁸, such as the single enzymes involved in both the substrate identification and the catalytic activity, the macromolecular complexes containing proteins and small RNA molecules as guides for the recognition of the substrate, and there are also multiple proteins to coordinate the activities of editing. When the editosome edits the base of

⁵1994, Albany Conference, Rensselaerville, NY, USA.

⁶1994, EMBO Workshop, Aussois, France.

⁷1996, EMBO Workshop, Maastricht, The Netherlands.

⁸The *catalysis* is a chemical phenomenon through which the speed of a chemical reaction undergoes changes intervention of substance (or mixture of substances), said *catalyst*, which is not consumed by the proceeding of reaction.

the nucleotides, such as in *A-to-I* and *C-to-U*, the editing factor acts in multiple sites.

In the last years it was discovered that A-to-I RNA editing can regulate the production of *RNA interference* (*RNAi*) and thus it maybe an important cellular mechanism in the modulation of the abundance of individual sequences within the transcriptome.

3.2.2 RNA Editing in Different Organisms

This section will explain briefly how RNA editing occurs in plants, animals and viruses, for their proper functioning in particular biological processes.

RNA Editing in Plants

Even if it happens rarely, the conversions C-to-U and U-to-C are the only types of RNA editing happening in mitochondria and plastids of the plants. In particular, RNA editing sites are found in majority in the coding regions of the mRNA, introns, and other non-translated regions [132].

Even if the exact mechanism is yet not known, considering that there are too many editing sites that needed to be changed in these organelles for a deaminase, some studies have suggested the involvement of *gRNA* and *editosome complex*.

The importance of RNA editing is seen also in the normal functioning of both the *translation* and the *respiration activity* in the plants [133]. RNA editing may be able to reactivate the functionality of the *tRNAs* [134, 135], as it corrects the *base-pairing* of these molecules [136]. Moreover, it has been

connected to the production of RNA- edited proteins, embedded within a polypeptides complex of the respiration pathway. Furthermore, it is very probable that the polypeptides synthesized by unedited RNAs would not work properly and would prevent the activity not only of mitochondria, but also of plastids.

RNA Editing in Animals

The process of *polyadenylation* (***polyA***) that occurs in the mitochondria of the animals was the first observation of the phenomenon of RNA editing. The polyadenylation is responsible for the derivation of the final part 3' in several mRNAs in animals. This mRNA downstream region is essential to complete some transcripts and ensures the correct translation of proteins in the mitochondria of the animals.

RNA Editing in Viruses

The RNA editing in viruses, such as *measles*, *mumps*, or *parainfluenza*, is used to give stability and to generate various proteins [137].

3.2.3 Editing by Deamination

The deamination is the deletion of an *amino group* from a molecule, resulting in a production of a molecule of *ammonia*⁹. The enzymes catalyzing this reaction are called *deaminase*. Moreover, for the reaction to occur, it requires a molecule of water. This is the reason why it is also called *oxidative deam-*

⁹Ammonia is a compound of nitrogen with the chemical formula NH_3 . It occurs as a colorless, toxic gas with the characteristic odor.

ination, because it oxidizes the carbon where the amino group was linked to, and replaces it with a *carbonyl group*. In humans, the deamination takes place primarily in the brain and in the liver, but it can also occur in the kidneys. This process allows the removal of the potentially harmful atoms of nitrogen that are in the amino acids. The process of deamination can occur both in the bases of *deoxyribonucleotides* (DNA molecule) and in those of *ribonucleotides* (RNA molecule).

The Deamination Process in the DNA

The process of deamination occurs spontaneously. When happening in DNA molecules, it can lead to genetic mutations, unless the damage is repaired. Below are described the main examples of deaminations affecting the DNA.

Deamination of Cytosine The process of spontaneous oxidative deamination of cytosine causes the formation of uracil (see *Figure 3.3*).

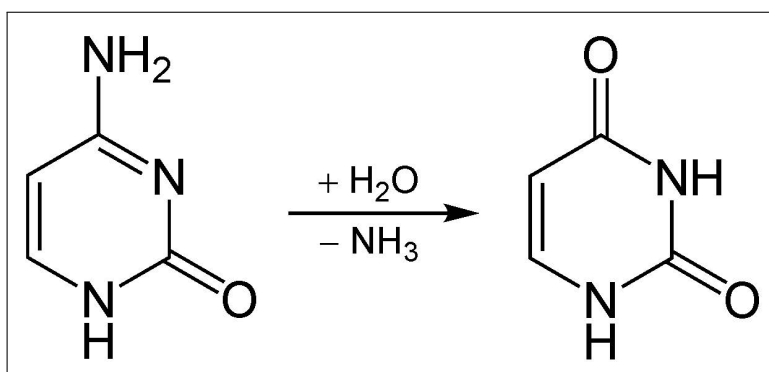


Figure 3.3: Spontaneous oxidative deamination of cytosine.

It can be induced *in vitro*, in order to distinguish in the double helix of the DNA the strand with *not-methylated* cytosine and the one with normal

cytosine. In the other case, if the process occurs in *vivo*, an uracil will be inserted in place of cytosine. This mismatch can be recognized and repaired by the DNA repair systems; if the error is not repaired within the next DNA replication, the new molecules of synthesized DNA will contain a mutation that will no longer be repairable.

Deamination of 5-methylcytosine It is possible to find *5-methylcytosine* mainly in prokaryotes. It is formed as the result of the methylation of a cytosine through an enzyme called *methyl transferase*. The deamination of this base causes the formation of thymine. In general, the DNA repair systems are not able to correct this reaction, since it does not recognize the thymine as incorrect, and so the mutation persists. This defect in the repair mechanisms contributes to the formation of rare *CpG* sites in eukaryotic genomes¹⁰. But there are also those rare enzymes that are able to both recognize the mismatch deriving from this phenomenon (**T-G**) and to replace the thymine with the cytosine.

Deamination of guanine The result of the deamination of guanine is the formation of the *xanthine* molecule¹¹. Instead of with cytosine, xanthine pairs with thymine. This process creates a mutation of post-replicative transition, in which the base pair that, at the beginning, was a G-C is now transformed into an A-T pair base. This kind of mutations can be corrected by the action

¹⁰They are regions of DNA in which along the linear sequence of bases occurring successions of cytosine followed by guanine, alternately. The notation is used for *CpG* to distinguish this linear sequence from the coupling of the base of cytosine with guanine

¹¹*Xanthine* is a purine base. In nature it exists as methyl derivative on the various nitrogen atoms

of the *alkyl adenine glycosylase* enzyme.

Deamination of adenine The result of the deamination of adenine is the formation of the *hypoxanthine* molecule¹². Instead of with thymine, the hypoxanthine is coupled selectively with cytosine. As it happens in the previous case, this process creates a mutation of post-replicative transition, so that the initial A-T base pair is transformed into the G-C base pair.

The Deamination Process in the RNA

The RNA editing by deamination is the process of deamination of the RNA. In the following sections, the main examples of RNA editing produced by deamination of the ribonucleotides base will be taken into consideration. Moreover, in the *Figure 3.4* the main effects resulting from the RNA editing process are shown.

C-to-U Editing The editing produced by the *cytidine deaminase* enzyme deaminates a base of cytidine and transforms it into a base of uridine. The *apolipoprotein B gene (APO B35)* in humans is an example of *C-to-U* editing. There are two isoforms in the human body: the *APO B100*, in the liver, and the *APO B48* synthesized exclusively in the small intestine. While the sequence of the *B100* apolipoprotein is **CAA**, when it is edited in the intestine it becomes **UAA**, which is a **STOP** codon. This phenomenon, however, does not occur in the liver. This concept is expressed in the *Figure 3.5*.

¹²The hypoxanthine is a purine derivative that occurs in nature. It is occasionally found as a component of nucleic acids

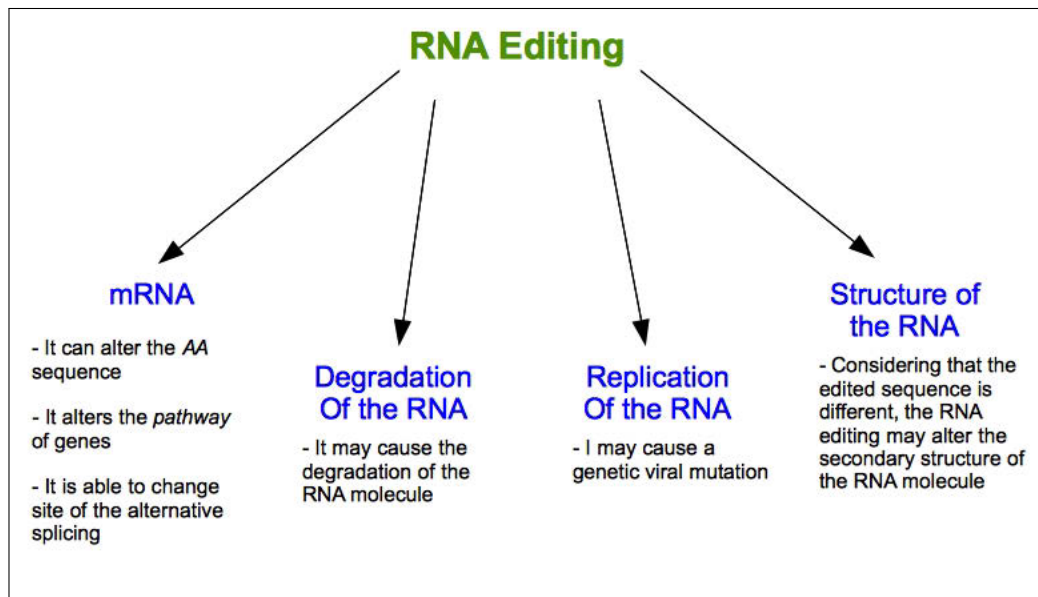


Figure 3.4: Possible effects caused by RNA editing.

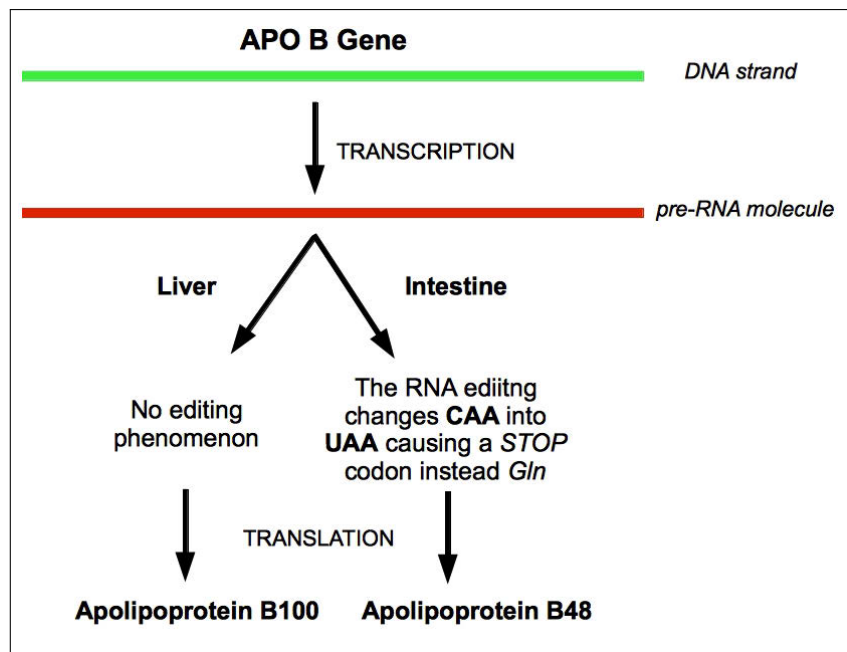


Figure 3.5: Example of C-to-U RNA editing in the Apo B gene of Human.

A-to-I editing The A-to-I RNA editing is the most studied editing phenomenon in eukaryotes and is induced by the *ADAR* family¹³. These enzymes modify a specific site of adenosine to inosine (that's why this particular type of RNA editing takes the name of *A-to-I*) in the pre-mRNAs. It seems that the editing produced by ADAR occurs in all metazoans, and this is essential for the development of mammals. The A-to-I RNA editing occurs in regions of double strand RNA (dsRNA). In the Figure 3.6 it is shown the action of ADAR in a double-strand region.

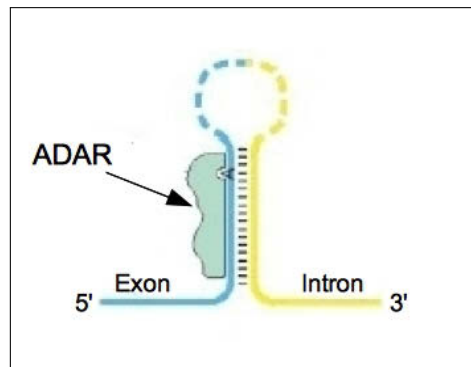


Figure 3.6: Example of action of the ADAR in a double-strand region.

The A-to-I editing can be either *specific* (if a single adenosine is edited within the dsRNA) or *promiscuous* (if the adenosines edited are up to 50%). The *specific* editing occurs within a short double-strand region (for example, those editing sites that are formed in a mRNA in which the bases of the intronic sequences pair in a complementary way to the bases of the exon sequences), while the promiscuous editing occurs within large duplex regions (e.g. *pre-* or *pri-miRNAs*, duplexes deriving either from transgenes¹⁴ or from

¹³*ADAR* is an acronym for adenosine deaminases acting on RNA

¹⁴The *transgene* is a gene that is introduced into an organism, and this gene is alien to the entire genome of the host organism

viral expressions, and, finally, duplexes resulting by the pairing of repetitive regions).

The consequences resulting from the A-to-I editing may be different. This can be related to the fact that the *inosine* (*I*) has the same behavior of *guanosine* (*G*) not only in the process of translation, but also in the formation of the RNA secondary structure (see both *Figure 3.7* and *3.8*).

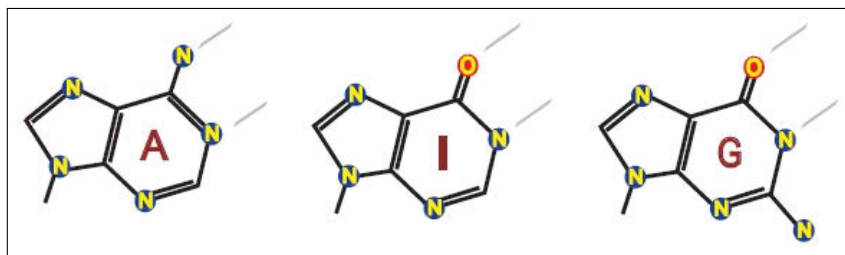


Figure 3.7: Molecular structures of adenine and inosine.

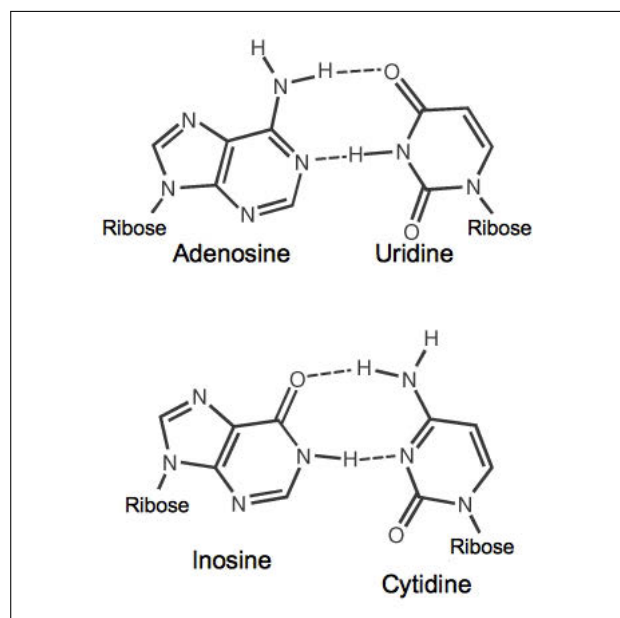


Figure 3.8: Inosine behavior, similar to the Guanosine one.

Among the effects there are, for example:

- *alteration of coding capacity,*
- *alteration of the set of miRNA and siRNA targets,*
- *formation of heterocromatina,*
- *inhibition of the process of miRNA and siRNA,*
- *splicing alteration.*

Those can be illustrated in *Figure 3.9*.

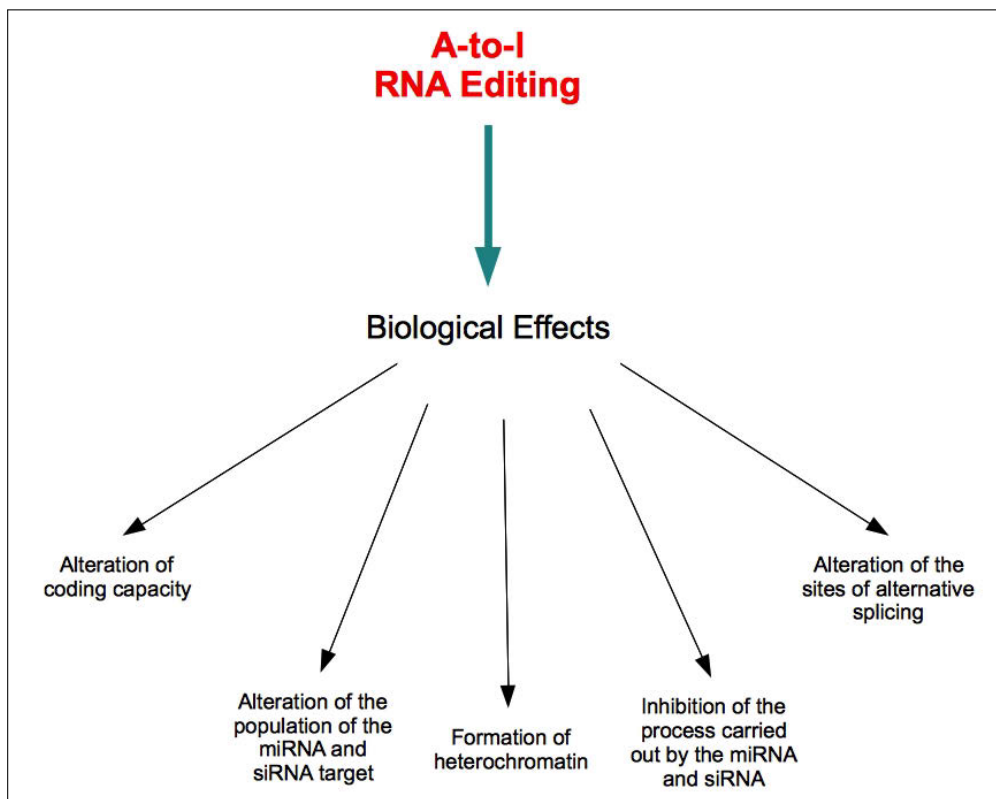


Figure 3.9: Main effect of the A-to-I RNA editing.

3.2.4 A-to-I RNA editing analysis

As said before, the A-to- I RNA editing can cause different effects on the stability of the structure of the RNA and its codification, but it is also able to affect the correct functioning of adjustment mechanisms, such as miRNAs and siRNAs. In this section, the history of RNA editing A-to-I, from its discovery to the present days, will be analyzed.

The origins of A-to-I RNA editing

In 1991 it was discovered an *A/G* discrepancy between the *cDNA* (*coding DNA*) and the genomic sequences of the *GluR-2 subunit*¹⁵ of mammals due to a modification of the base at the RNA level [138]. The modification of this nucleotide of adenosine converted a codon that codified *glutamine* in a codon encoding *arginine*.

Thanks to the discovery of the codification of adenosine caused by the RNA editing, several other cases in the transcripts of the nervous system were identified. In each of those cases, the change of a single nucleotide, causing the substitution of an amino acid, could be connected to the change in the function of the protein. The simple fact that the variants of both the edited and the non edited proteins were co-expressed in the same cells, let the scientists realize that the RNA editing was not only an important mechanism for the diversity of the genetic information, but that it also had the ability to increase the complexity of both the eukaryotic transcriptome and the proteome.

¹⁵It is a protein which in humans is encoded by the gene *GRIA2*. It is a neurotransmitter receptor in the brain of human and is activated in varied physiological processes

When the editing in the mRNA encoding GluR-2 was discovered, the processes involved were unknown and two were the explanations given to the *A-to-G* change observed in the cloned cDNA. The *first* was that it was considered as the result of a process of unknown modification of the adenosine that alters the purine into another purine base equivalent to the guanosine (such as the *hypoxanthine*); the *second* explication was that it was caused by a mechanism involving first the removal of either the base or the entire nucleotide, and then the introduction of guanosine.

As it was known in the past as well, the *adenosine deaminase enzyme* (**ADA**) converts the adenosine mononucleotides in *hypoxanthine* nucleotides (also called *inosine*) but it is also able to mediate the metabolism of both the eukaryotes and the prokaryotes nucleotides. The ADA has important therapeutic *marker* as the *ADA deficiency* which leads to various types of disorders of the immune system [139]. Moreover, the ADA modifies adenosine mononucleotides using a mechanism of *hydrolytic* deamination, as explicated in the *Figure 3.10*.

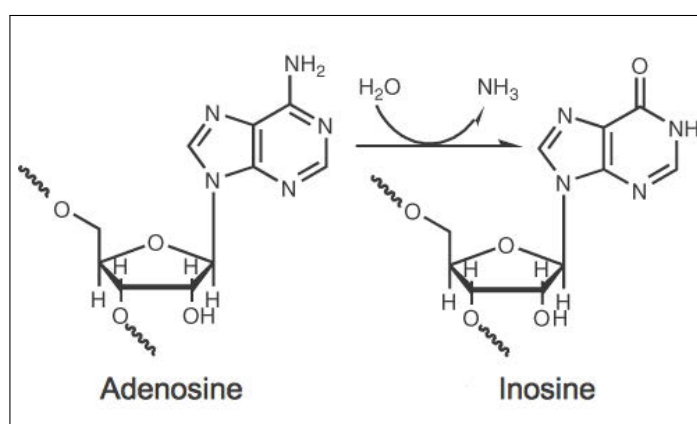


Figure 3.10: Transition from adenosine to inosine.

In addition to the modification of mononucleotides by the ADA, it was also known the modification of adenosine into inosine genomically encoded in *transfer RNA* (*tRNA*), which is a critical process for the generation of the genetic code.

Right before the discovery of the editing process changing the adenosine in the pre-mRNAs, it was discovered a new enzyme activity. It hit the adenosine incorporated in the double-strand RNA molecules (dsRNA) [140, 141] and through the analysis of the reaction products, it was verified that the actual molecular process was to modify adenosine into inosine. Thus, the double-strand structure of the RNA was an essential feature for the editing to occur, even if it was not observed any primary sequence neither upstream nor downstream the editing. This is the main difference between the inner working of the A-to-I editing and the C-to-U deamination processes, involving secondary structure elements as well as a motif of primary sequence, guiding the RNA modification system.

The protein responsible for all this process was initially called *dsRAD*, or *Drada*, and was later renamed as *ADAR1* (in *Figure 3.11* the tertiary structure of the protein ADAR1 is shown).

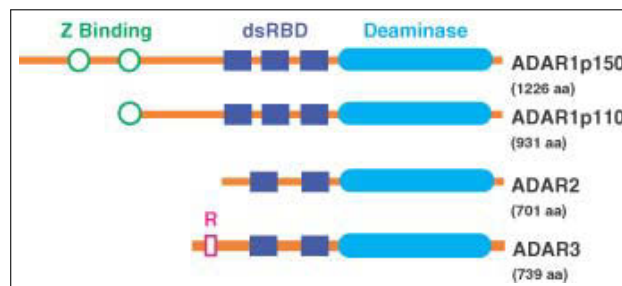


Figure 3.11: Comparison of ADAR proteins.

This same protein was previously studied in several laboratories as both a protein with a potential viral function [142] and as specific A-to-I editing activity in the dsRNA in mammalian cells [143, 144]. *ADAR2* and *ADAR3*, together with other similar forms in vertebrates [145], flies [146] and worms [147], were discovered after the cloning of the first ADAR (*ADAR1*). In Figure 3.12 is shown the various forms of the ADAR protein.

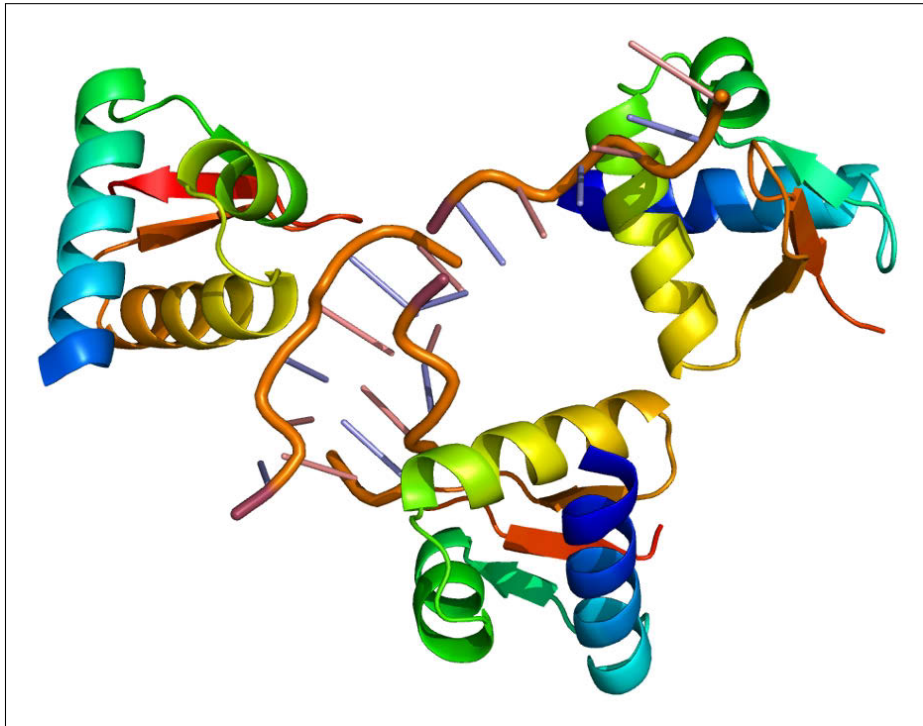


Figure 3.12: Tertiary structure of the *ADAR1* protein.

RNA Editing in miRNA molecules

The double-strand RNAs are the possible target not only for the ADAR, but also for any other protein binding to the dsRNA, such as the components of the process of RNA silencing. It is easy to realise the connection between the

processes of RNA editing and its silencing. There may be often a competition between the editing machinery and the silencing one for the double-strand molecules. Concerning the RNA, the result may be due to the set of enzymes that operate in the RNA molecule as first. According to another model, the RNA editing may be a nuclear event that induces the silencing at the level of chromatin [148].

Thanks to the discovery of some members of microRNA subject to A-to-I editing [149, 150, 151, 152], the relationship between editing and silencing is now more evident. Analyzing the characterization of the secondary structures of known targets of ADAR enzymes, it can be hypothesized that the molecules of miRNAs might undergo the RNA editing. As seen above, before their maturation miRNAs have a hairpin-shaped molecular structure. Pri-miRNAs (the pre-mRNA transcripts of miRNAs) have usually a few hundreds nucleotides length and they are first of all processed in pre-miRNA within the nucleus. The pre-miRNA, which is composed of approximately 70-90 nucleotides, is exported to the cytoplasm, where a second process generates a 20 to 22 nucleotides molecule, representing the functional miRNA.

The occurrence of A-to-I RNA editing in a molecule of miRNA has been described first in *miRNA22* of both the human and the mouse [149]. The observed editing events were localized both inside and outside the *seed* of the miRNA. Depending on where the editing occurs, there can be two different consequences: either stopping the function of the mature miRNA or allowing the miRNA to bind to those RNA molecules to which it could not bind before.

In a work by Glen M. Borchert and others, published in September 2009 in the journal *Human Genetics Molecular*, it has been analyzed the deamination

of adenosine that occurs in the human cDNA [153]. They hypothesized that there is a relationship between the events of A-to-I editing in the non-coding regions 3' (3' *UTR*) and the portion of bond *mRNA::miRNA*. They found meaningful correlations between the A-to-I editing and the complementarity modification of the miRNAs. In fact, over 3.000 on 12.723 evaluated editing sites were found complementary to the seed matches form of a subset of human miRNAs.

In addition, the group noted in 200 ESTs the editing sites within a motif of 13 nucleotides long. The deamination of this motif simultaneously creates the seed matches for three microRNAs, an impossible situation if the editing had not occurred. According to these results, one of the functions of the ADAR is to create regulatory sites for miRNAs. This means that many of them might be identified among the miRNA target sites only through the examination of expressed sequences.

It has been estimated that between 6% and 10% of all miRNA genes are subject to the A- to-I modification [152]. Thanks to the identification of editing events in miRNAs, it has been showed that not only the transcripts of miRNAs are subject to post- transcriptional modification, but also that the functions of miRNAs might not be fully deductible when their genomic sequence is analyzed.

Part II

Biological Databases

Chapter 4

Biological databases and their analysis: *miRandola*, *miR-EdiTar* and *VIRGO*

*One sometimes finds what one is
not looking for.*

Sir Alexander Flamming

Nobel Prize in Physiology

or Medicine (1945)

WITH the development of the techniques of massive sequencing, the genomic has gained billion sequences, so that the amount of data to analyze and manage is huge. Thus, the storage and analysis of biological data require enormous computing resources and clusters of thousands of processors. This opens the doors to the inception and the developmen of *biological databases*. They can be considered as libraries of life sciences information, collected from scientific experiments, published literature, high-throughput

experiment technology, and computational analyses.

Biological databases help researchers to understand and explain different biological phenomena and give the chance to create a permanent data platform, where data are easily available but not perishable.

I have been working on the creation of biological databases: *miRandola*, *miR-EdiTar*, and *Virgo*. All of them available on line.

- *miRandola* is an extracellular circulating MicroRNAs Database. It is connected to miRò, the miRNA knowledge base, allowing users to infer the potential biological functions of circulating miRNAs and their connections with phenotypes [154].
- *miR-EdiTar* is a database of predicted A-to-I edited miRNA binding sites. The database contains predicted miRNA binding sites that could be affected by A-to-I editing ("current" sites), and sites that could become miRNA binding sites as a result of A-to-I editing ("novel" sites). It has an experimental example of a miRNA binding site created by editing events. The goal is to facilitate the identification of miRNA binding sites potentially affected by A-to-I editing as a function of the number of base pair matchings, the degree of accessibility of the binding site and the stability of the interaction, and to aid the discovery of new potential miRNA binding sites that might be created by editing events [155].
- *VIRGO* is a web-based tool that maps A-to-G mismatches between genomic and EST sequences as candidate A-to-I editing sites. VIRGO is built on top of a knowledge-base integrating information of genes

from UCSC, EST of NCBI, SNPs, DARNED, and Next Generations Sequencing data. The tool is equipped with a user-friendly interface allowing users to analyze genomic sequences in order to identify candidate A-to-I editing sites [156].

4.1 *miRandola*: Extracellular Circulating MicroRNAs Database

Despite a lack in literature, a precise classification among the different forms of circulating miRNAs is required. Mirandola is a database of *extracellular/circulating* miRNA [154]. MiRNAs are classified into four categories, based on their extracellular form:

- *miRNA-Ago2*;
- *miRNA-exosome*;
- *miRN-HDL*;
- *circulating miRNA*.

The database provides the users with a variety of information including the associated diseases, the samples, the methods used to isolate the miRNAs, and the description of the experiment. The information about the targets of miRNAs and their records are provided through links to miRò, “*the miRNA knowledge base*” [157]. miRò integrates data from different sources to allow the identification of associations among genes, processes, functions, and diseases through validated and predicted targets of miRNAs.

4.1.1 Mirandola BackEnd

Mirandola is a database containing 119 articles, 2276 entries, and 590 unique mature miRNAs. It was created in MySQL and consists of 6 tables:

- ***mirnas***: this table collects information about miRNAs, such as *family*, *mature miRNA*, and type of miRNAs with the following fields, “mirna_ID” (*primary key*), “Mature_mirna”, “mirna_family”, “mirna_type”, “sample_ID”, “article_ID”, “experiment_ID”, “Mirbase_accession”, “Last_Version”, “AC_code” , “potential biomarker”;
- ***samples***: this table contains information about the samples where miRNAs were found. Its fields are “sample_ID” (primary key), “sample”, “sample_source”, “sample_name”;
- ***articles***: this table regards the details about the articles used to extract the information in miRandola. It has the following fields: “article_ID” (primary key), “author”, “journal”, “title”, “link”, “data”, “database”;
- ***experiments***: this table provides details about the experiment performed. It has the following fields: “experiment_ID” (*primary key*), “organism”, “method”, “description”, “database”, and “link”.

The table miRNAs is linked to the table samples, articles and experiments respectively via “sample_ID”, “article_ID” and “experiment_ID”. This allows to know the sample from which it was extracted each miRNA, with a trial description of the organism, the method used, and the sources of the infor-

4.1. MIRANDOLA: EXTRACELLULAR CIRCULATING MICRORNAS DATABASE122

mation, such as *paper*, *article title*, *author*, *date of publication* and a link find it. More details are visible in *Figure 4.1*.

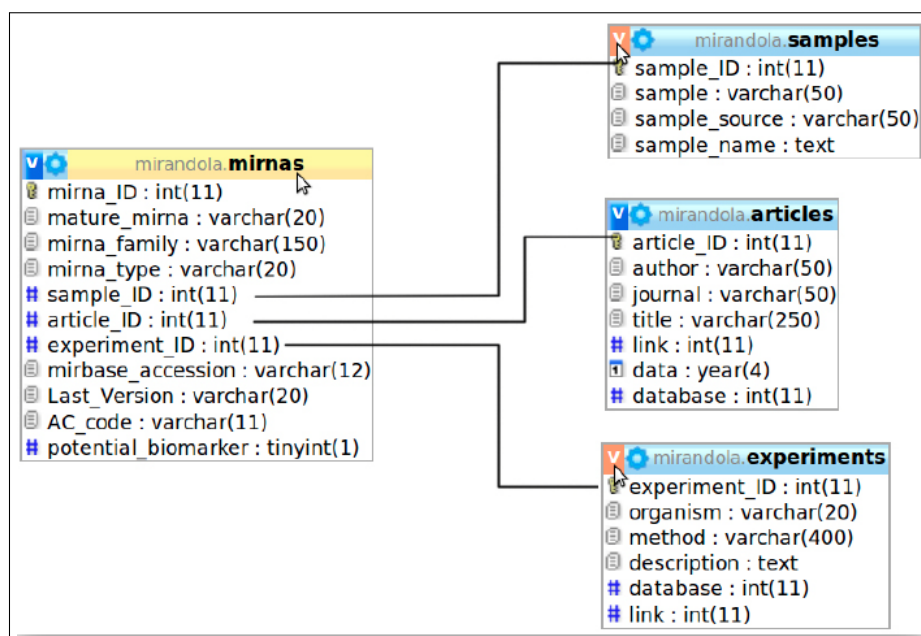


Figure 4.1: Tables of *miRandola* relative to *mirRNAs*, *samples*, *articles* and *experiments*.

- **mirna_converter**: this table converts any of the mature miRNAs from the miRBase version 12 to miRBase version 18. It has the following fields: “id” (primary key), “mirbase_accession”, “version_12”, “version_13”, “version_14”, “version_15”, “version_16”, “version_17”, “version_18”;
- **submission**: this table contains the information of those that want to enter their circulating miRNAs in *miRandola*. It has the following fields: “submission_ID” and “PubMed” (together they are the primary key), “mature_mirna”, “mirna_type”, “sample”, “sample_name”,

4.1. MIRANDOLA: EXTRACELLULAR CIRCULATING MICRORNAS DATABASE¹²³

“species”, “firstname”, “lastname”, “affiliation”, “email”.

mirandola.mirna_converter	mirandola.submission
mirbase_accession : varchar(12)	submission_ID : int(11)
version_12 : varchar(20)	mature_mirna : varchar(100)
version_13 : varchar(20)	mirna_type : varchar(100)
version_14 : varchar(20)	sample : varchar(100)
version_15 : varchar(20)	sample_name : text
version_16 : varchar(20)	species : varchar(200)
version_17 : varchar(20)	PubMed : int(11)
version_18 : varchar(20)	firstname : varchar(50)
id : int(11)	lastname : varchar(50)
	affiliation : varchar(100)
	email : varchar(100)

Figure 4.2: Tables of *mirna_converter* and *submission*.

miRandola’s frontend was realized by using HTML, PHP, CSS and JQuery.

The website is available at <http://atlas.dmi.unict.it/mirandola>.

4.1.2 Sections of Mirandola

There are several sections:

Homepage It contains a website description, with the database outline, the sources used, and small box “News”, showing the latest news¹.

Search Here it is possible searching for a miRNA by category²:

- *Mature miRNA*;
- *miRNA Family*;

¹Website: <http://atlas.dmi.unict.it/mirandola/index.html>

²Website: <http://atlas.dmi.unict.it/mirandola/browse.php>

4.1. MIRANDOLA: EXTRACELLULAR CIRCULATING MICRORNAS DATABASE124

MicroRNAs (miRNAs) are small (approximately 22 nt) noncoding RNAs that play an important role in the regulation of various biological processes through their interaction with cellular messenger RNAs. They are frequently dysregulated in cancer and have shown promise as tissue-based markers for cancer classification and prognostication. Extracellular miRNAs in serum, plasma, saliva, urine and other body fluids have recently been shown to be associated with various pathological conditions including cancer. miRNAs circulate in the bloodstream in a highly stable, extracellular form, thus they may be used as blood-based biomarkers for cancer and other diseases. Circulating miRNAs are protected by encapsulation in membrane-bound vesicles such as exosomes, but the majority of circulating miRNAs in human plasma and serum cofractionate with Argonaute2 (Ago2) protein, rather than with vesicles. In the present work, we performed a comprehensive classification of different extracellular circulating miRNA types. A direct link to the knowledge base miRb together with the inclusion of datamining facilities allow users to infer possible biological functions of the circulating miRNAs and their connection with the phenotype. To our knowledge miRandola is the first database that provides information about all kind of extracellular miRNAs and we believe that it will constitute a very important resource for researchers.

Useful Links

- miRb
- miRiam
- miRScope
- FerroLab
- Lipari School
- miR-EdiTar
- Virgo

References

If you use miRandola please cite the following paper:
Francesco Russo, Sebastiano Di Bella, Giovanni Nigita, Valentina Macca, Alessandro Laganà, Rosalba Giugno, Alfredo Pulvirenti, Alfredo Ferro,
miRandola: Extracellular Circulating microRNAs Database
PLOS ONE 7(10): e47786 doi:10.1371/journal.pone.0047786

Comments, questions? Contact [Francesco Russo](#) or [Sebastiano Di Bella](#)

Database Schema

```
graph TD
    Vesiclepedia[http://www.microvesicles.org/] --> miRandola
    PubMed[http://www.ncbi.nlm.nih.gov/pubmed/] --> miRandola
    miRb[http://fermiab.dmi.unict.it/mirb/index.html] --> miRandola
    ExoCarta[http://www.exocarta.org/] --> miRandola
    Submission[http://atlas.dmi.unict.it/mirandola/down\_up.php] --> miRandola
    miRBase[http://www.mirbase.org/] --> miRandola
```

News

Integration of DAVID web service

Figure 4.3: Homepage of *miRandola*.

4.1. MIRANDOLA: EXTRACELLULAR CIRCULATING MICRORNAS DATABASE125

- *miRNA type*;
- *Sample*;
- *Disease and Malignant Cell Lines*;
- *Potential Biomarker*.

For each category, the research can be done either by selecting an entry from select or typing in the relevant area the element's name (*Figure reffig-miRandola4*). For each research, both the number of results found and an information sheet to each miRNA are provided. Moreover, the results can be saved in three different formats, *pdf*, *txt*, and *csv*.



Figure 4.4: *Search* page in *miRandola*.

4.1. MIRANDOLA: EXTRACELLULAR CIRCULATING MICRORNAS DATABASE126

Results found for hsa-let-7a: #16		
1	<p>Each entry in miRandola provides links to miRd knowledge base</p> <p>Legend:</p> <ul style="list-style-type: none"> • miRNA associated Diseases (D) • miRNA associated Functions (F) • miRNA associated Processes (P) • miRNA associated Tissues (T) 	
	Mature miRNA from literature:	hsa-let-7a (D) (F) (P) (T)
	Mature miRNA from miRBase:	hsa-let-7a-5p 
	miRBase Accession:	MIMAT0000062
	miRNA family:	let-7
	Potential Biomarker:	unknown
	miRNA type:	exosome
	Sample:	serum
	Sample source:	ts
	Diseases and Malignant Cell Lines:	papillary adenocarcinoma of ovary
	First Author:	Taylor DL et al.
	Journal:	Gynecol Oncol.110(1):13-21.
	Title:	Microme signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer.
	PubMed ID:	18589210 
	Date of Publication:	2008
	Methods:	457microme array
	Experiment Description:	Microme signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer.
	Database:	

Figure 4.5: Example of results page.

4.1. MIRANDOLA: EXTRACELLULAR CIRCULATING MICRORNAS DATABASE¹²⁷

Advanced Search In this page, an advanced research can be done, after choosing between two categories³:

- *miRNA family*;
- *mature miRNAs*

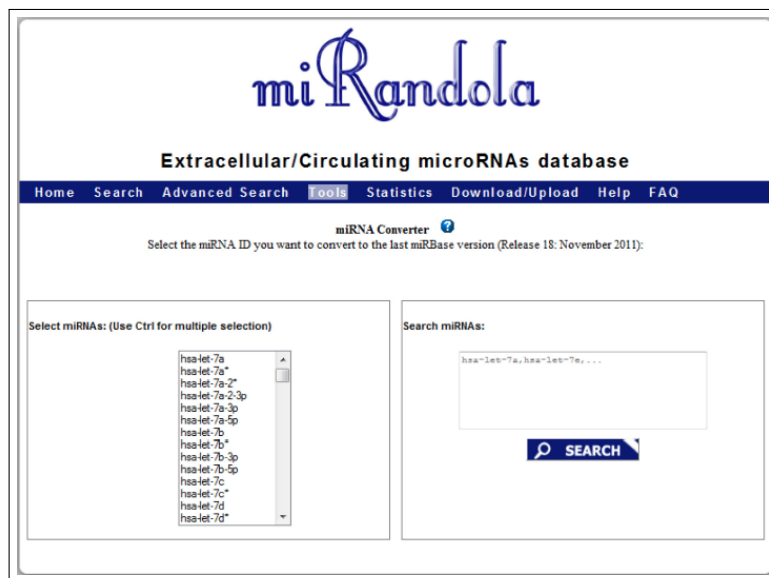
From a select it is possible to select the entry and then filter the research by miRNA *type* and *samples*.

Figure 4.6: *Advanced search* page in *miRandola*.

Tools In this page it is possible to convert the nomenclature of a mature miRNA into a miRBase version 18⁴. The conversion takes place either by using the select or typing in the textarea a list of miRNAs separated by “,”.

³Website: http://atlas.dmi.unict.it/mirandola/search_ad.php

⁴Website: <http://atlas.dmi.unict.it/mirandola/tools.php>

Figure 4.7: *Tools* page in *miRandola*.

Statistics Here there are some statistics about the database, such as the number of publications in miRandola, the number of entries, the distribution of miRNAs in the samples, a *wordle* of the most present miRNAs, the most frequent journals in miRandola, and the list of all the articles treated manually⁵.

Download / Upload This page allows not only to send a new record to insert in miRandola, but also to download the data already in the system⁶.

Help It provides a helpful tutorial in using miRandola⁷.

⁵Website: <http://atlas.dmi.unict.it/mirandola/statistics.php>

⁶Website: http://atlas.dmi.unict.it/mirandola/down_up.php

⁷Website: <http://atlas.dmi.unict.it/mirandola/help.html>

FAQ This page contains the answers to the frequently asked questions about miRandola⁸.

4.1.3 miRandola - miRò

The function of circulating miRNAs is still largely unknown. According to some reports, endogenous miRNAs can be carried by *high-density lipoprotein* (**HDL**): in this way they are able to enter inside receiving cells and contribute to the repression of their targets [158]. Moreover, exosomes appear to play an important role in the development of metastases. The role of miRNAs in targeting sites far from the primary organ where they were originated is, however, still unknown [159].

For all these reasons, and to help formulating hypotheses about the function of miRNAs, found into the extracellular space, connections between miRandola and miRò were made. miRò provides the user with information about functional annotations through validated and predicted targets of miRNAs.

Each entry in miRandola provides links to other diseases, processes and functions in which each miRNA is involved, and the tissues in which it is expressed.

⁸Website: <http://atlas.dmi.unict.it/mirandola/faq.html>

4.1. MIRANDOLA: EXTRACELLULAR CIRCULATING MICRORNAS DATABASE130

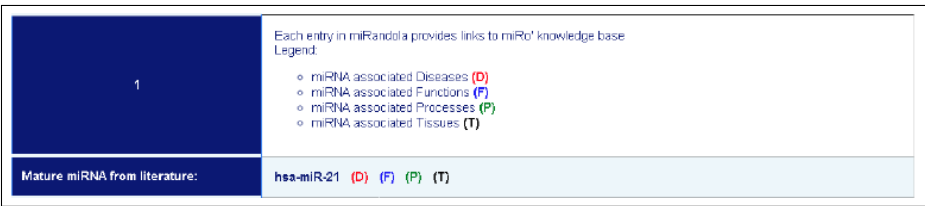


Figure 4.8: Link between *miRandola* and *miRò* (disease, functions, processes, tissues).

After clicking on the letter **D** (in red), the user will be directed to the page of *miRò* referred to all the pathologies associated to the specific miRNAs (e.g. *hsa-miR-21* in the Figure 4.9):

miRò
The miR-ontology Database

By miRNA: **hsa-miR-21**
Click on miRNA name to see details

hsa-miR-21 associated Broad Phenotypes (Diseases)
Click on gene name or sources to see details

Order by Broad Phenotypes (Diseases) Name
Order by Gene Name

Show page 1 Go

Broad Phenotypes (Diseases): 3198 entries

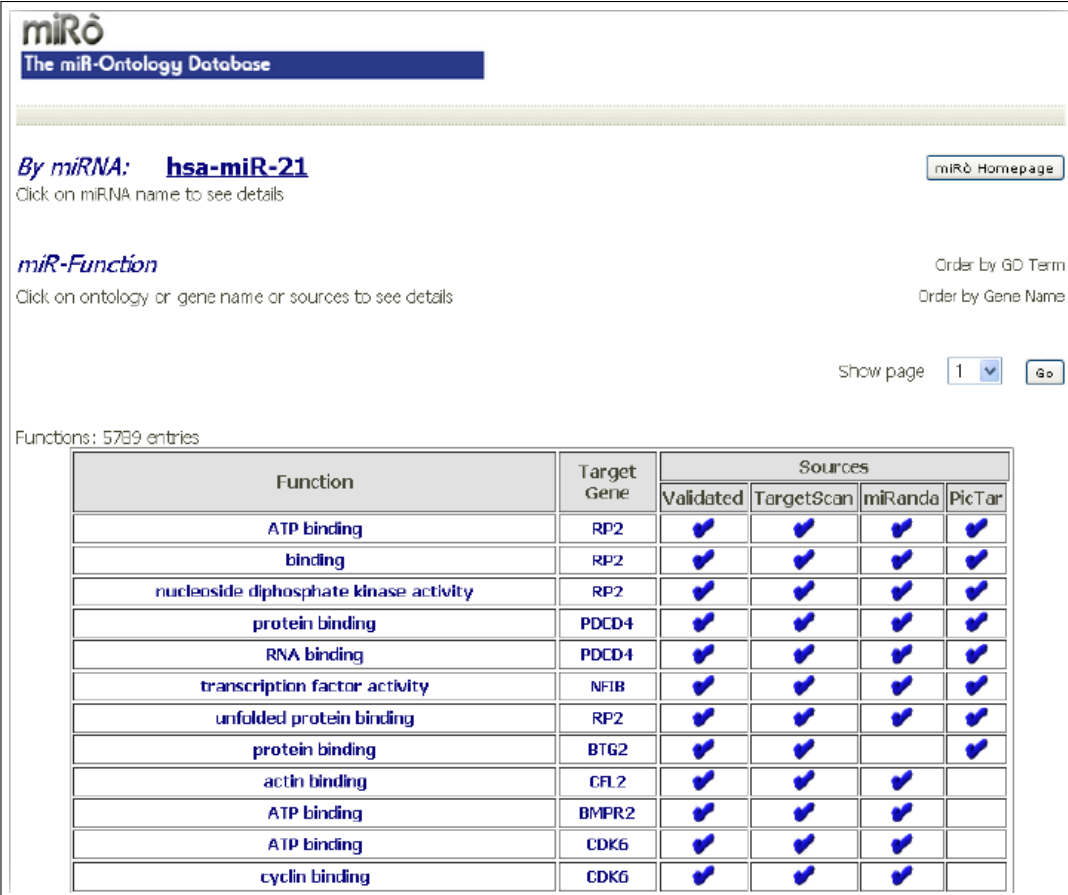
Broad Phenotypes (Diseases)	Target Gene	Sources			
		Validated	TargetScan	miRanda	PicTar
retinal dystrophy	RP2	✓	✓	✓	✓
retinitis pigmentosa	RP2	✓	✓	✓	✓
diabetes, type 1	SOC55	✓	✓	✓	
heart anomalies, congenital	BMPR2	✓	✓	✓	
hypertension	BMPR2	✓	✓	✓	
juvenile polyposis	BMPR2	✓	✓	✓	
pulmonary hypertension	BMPR2	✓	✓	✓	
vasoreactivity	BMPR2	✓	✓	✓	
clubfoot	APAF1	✓	✓		
melanoma	APAF1	✓	✓		
Alzheimer's Disease	FAS	✓			
arthritis lupus erythematosus	CDKN1A	✓			

Figure 4.9: Page of *miRò* relative to diseases of *has-miR-21*.

In Figure 4.9 it is possible to see the result returned by *miRò*. Each disease

4.1. MIRANDOLA: EXTRACELLULAR CIRCULATING MICRORNAS DATABASE131

is associated with one or more targets of miRNAs, validated or predicted by different prediction programs. If, instead, the user clicks on the letter **F** (in *blue*), he/she will be directed to the page concerning the biological functions associated with the miRNA (see the *Figure 4.10*).



The screenshot shows the miRò database interface for hsa-miR-21. It includes a header with the miRò logo and 'The miR-Ontology Database'. Below the header, there's a section for 'By miRNA: hsa-miR-21' with a link to 'miRò Homepage'. A 'miR-Function' section is also present, with a link to 'Click on ontology or gene name or sources to see details'. The main content is a table titled 'Functions: 5789 entries' showing various biological functions, their target genes, and the sources used for validation (Validated, TargetScan, miRanda, PicTar).

Function	Target Gene	Sources			
		Validated	TargetScan	miRanda	PicTar
ATP binding	RP2	✓	✓	✓	✓
binding	RP2	✓	✓	✓	✓
nucleoside diphosphate kinase activity	RP2	✓	✓	✓	✓
protein binding	PDCD4	✓	✓	✓	✓
RNA binding	PDCD4	✓	✓	✓	✓
transcription factor activity	NFIB	✓	✓	✓	✓
unfolded protein binding	RP2	✓	✓	✓	✓
protein binding	BTG2	✓	✓		✓
actin binding	CFL2	✓	✓	✓	
ATP binding	BMP2	✓	✓	✓	
ATP binding	CDK6	✓	✓	✓	
cyclin binding	CDK6	✓	✓	✓	

Figure 4.10: Page of *miRò* relative to functions of *has-miR-21*.

By clicking on the letter **P** (in *green*) all the biological processes associated with the miRNA will be shown (see the *Figure 4.11*).

Finally, by clicking on the letter **T** (in *black*), the user will be able to view the expression levels of miRNA in different tissues (see the *Figure 4.12*).

4.1. MIRANDOLA: EXTRACELLULAR CIRCULATING MICRORNAS DATABASE132

miRò

The miR-Ontology Database

By miRNA: **hsa-miR-21**

miRò Homepage

Click on miRNA name to see details

miR-Process

Order by GO Term

Order by Gene Name

Click on ontology or gene name or sources to see details


Show page 1 Go

Processes: 5403 entries

Process	Target Gene	Sources			
		Validated	TargetScan	miRanda	PicTar
apoptosis	PDCD4	✓	✓	✓	✓
beta-tubulin folding	RP2	✓	✓	✓	✓
cell aging	PDCD4	✓	✓	✓	✓
CTP biosynthetic process	RP2	✓	✓	✓	✓
DNA replication	NFIB	✓	✓	✓	✓
forebrain development	NFIB	✓	✓	✓	✓
GTP biosynthetic process	RP2	✓	✓	✓	✓
hindbrain development	NFIB	✓	✓	✓	✓
lung development	NFIB	✓	✓	✓	✓
negative regulation of cell cycle	PDCD4	✓	✓	✓	✓
negative regulation of cell proliferation	NFIB	✓	✓	✓	✓
negative regulation of JNK activity	PDCD4	✓	✓	✓	✓

Figure 4.11: Page of *miRò* relative to processes of *has- miR-21*.

4.1. MIRANDOLA: EXTRACELLULAR CIRCULATING MICRORNAS DATABASE133



The miR-ontology Database

By miRNA: [hsa-miR-21](#)

Click on miRNA name to see details

hsa-miR-21 associated Tissues

Click on gene name or sources to see details

miRò Homepage

Order by Tissue Name

Order by Expression Value

Show page

Tissues: 161 entries

Tissue	Library	Expression
	hsa_USSC-d7	0.67% - 59/8820
Adeno-CA breast, epithelial like	hsa_Breast-adenocarcinoma-MCF7	0.66% - 58/8820
AML	hsa_AML1-d29	0.12% - 11/8820
AML	hsa_AML-THP1	0.15% - 13/8820
AML	hsa_AML3-d0	0.39% - 34/8820
AML	hsa_AML1-d0	0.26% - 23/8820
AML	hsa_AML2-d0	0.17% - 15/8820
AML	hsa_AML-HL60	0.01% - 1/8820
AML	hsa_AML3-d29	0.1% - 9/8820
Astroblastoma	hsa_Astroblastoma-DD040800	1.52% - 134/8820
B-ALL	hsa_B-ALL8-d43	0.05% - 4/8820
B-ALL	hsa_B-ALL2-d0	0.1% - 9/8820

Figure 4.12: Page of *miRò* relative to tissues of *has- miR-21*.

4.2 miR-EdiTar: A database of predicted A-to-I edited miRNA target sites

Alterations of A-to-I editing have been associated to several human diseases, such as infections, neurological diseases and cancer [160, 161, 161]. Moreover, A-to-I editing can influence micro RNA (miRNA)-mediated gene regulation [162]. Several cases of A-to-I editing of miRNA precursors have been reported [163, 164]. This phenomenon can suppress processing by Drosha and Dicer, while the presence of inosines in the mature sequences can alter the recognition of their target sites [?]. A-to-I editing is most abundant in the 3' untranslated regions (UTRs) of the human transcriptome [165, 166]. This could affect the existing miRNA binding sites as well as generate novel binding sites [167].

The importance of RNA editing in miRNA activity suggests the need for computational tools to predict and analyze the effects of RNA editing on miRNA-mediated regulation.

4.2.1 The construction of miR-EdiTar

miR-EdiTar is a database of predicted A-to-I edited miRNA binding sites [155]. The database contains predicted miRNA binding sites that could be affected by A-to-I editing and sites that could become miRNA binding sites as a result of A-to-I editing⁹.

⁹miR-EdiTar is freely available online at <http://microrna.osumc.edu/mireditar>

4.2. MIR-EDITAR: A DATABASE OF PREDICTED A-TO-I EDITED MIRNA TARGET SITES

Prediction of A-to-I edited miRNA binding sites

The first step of the research is the collection of data. In fact, 1,139 human 3' UTR sequences with a total of 10,571 A-to-I editing sites were gathered together from the first release of *DAtabase of RNa Editing* (**DARNED**) [168]. The computational method **miRiam** [169] was used to predict miRNA-target interactions involving the edited sites and exploits binding rules inferred based on experimentally validated miRNA/target pairs and the structural accessibility of the target sites. This last feature is estimated based on the local pairing probability computed by RNAplfold of the Vienna RNA package [170] with the parameters $W = 80$ (sliding window length), $L = 40$ (interactions outside the span size of 40 are not allowed) and $u = 4$ (the stretch of consecutive bases for which the probability of being unpaired is computed), as recommended in [171]. In particular, the accessibility is computed as the average probability of stretches of 4 nucleotides to be unpaired in the predicted binding site. The score of the duplex structure and its free energy are also computed. In particular, the latter is computed by using the tool **RNA duplex** from the *Vienna RNA package* [172]. For each affected binding site the accessibility, the duplex structure and the free energy are computed for both the unedited and the edited version of the duplex, in order to evaluate the effects of the editing events on the binding.

Predictions were performed on the complete set of 1,922 human miRNA sequences, retrieved from *miRBase Release 18* [33]. 9,532 out of 10,571 (90%) edited adenosines were predicted to fall in at least one miRNA binding site. 1,102 UTRs (96.75%) had at least one edited adenosine on an miRNA

4.2. MIR-EDITAR: A DATABASE OF PREDICTED A-TO-I EDITED MIRNA TARGET SITES

seed binding site, while 771 (67.7%) had all their edited adenosines on at least one miRNA seed binding site. On the miRNA side, 1,664 miRNAs (86.6%) had at least one seed binding site potentially affected by editing.

The duplexes were then classified into two categories, depending on whether the edited adenosines were located on an miRNA seed binding region or not. Seed matches were classified as *6mer*, *7mer-A1*, *7mer-m8* and *8mer*, as in citeBartel2009.

Furthermore, an important aim was to find all the novel miRNA binding sites potentially generated by A-to-I editing. By changing all the edited adenosines in guanosines in the set of the 1,139 human 3' UTR sequences and repeating the above analysis, 1,076 UTR sequences (94.45%) had at least one novel binding site created by editing events and 1,400 miRNAs (72.8%) had at least one target site potentially created by editing.

The *table 4.1* below summarizes the descriptive statistics:

4.2.2 miR-EdiTAr contents

miR-EdiTAr contains a collection of predicted human miRNA binding sites in A-to-I edited 3' UTR sequences. The database contains two kinds of sites:

- “current” sites, that are those sites predicted to be miRNA binding sites but that could be affected by A-to-I editing;
- “novel” sites are those sites not predicted to be miRNA binding sites but that could become miRNA binding sites as a result of A-to-I editing.

The web site can be searched by miRNA and/or by target. Given an miRNA, the list of its predicted targets is shown in a box. When a target is

4.2. MIR-EDITOR: A DATABASE OF PREDICTED A-TO-I EDITED MIRNA TARGET SITES

<i>Edited sites</i>	
Edited sites on the 3' UTRs	10,571
Edited sites on a predicted miRNA binding site	9,532
<i>Targets</i>	
3' UTR sequences affected by editing	1,139
3' UTR sequences with at least one edited base in a predicted miRNA binding site	1,102
3' UTR sequences with all their edited bases in a predicted miRNA binding site	771
3' UTR sequences with at least one novel predicted binding site created by editing	1076
<i>miRNAs</i>	
miRNAs with predicted sites affected by editing	1,664
miRNAs with predicted novel sites created by editing	1,400

Table 4.1: Overall Descriptive Statistics.

selected, the corresponding interaction details are displayed on a table and available for download in comma separated value (CSV) format. The binding sites are grouped into two categories based on their type (current sites or novel sites). Several data elements are provided, such as the position of the binding site on the UTR, the seed type, the free energy of the duplex, the structural accessibility degree, the interaction score and the duplex structure. The edited bases are highlighted in bold characters and the corresponding alignment pipes are replaced with an X, indicating the potential disruption of the corresponding bond. In the case of current sites, an entry indicates whether the edited bases are located in the seed region. Moreover, the values of seed type, free energy, accessibility, interaction score and duplex structure are provided for both the edited and unedited forms of the site. Similar results can be obtained by choosing a target from the list and then selecting

4.2. MIR-EDITAR: A DATABASE OF PREDICTED A-TO-I EDITED MIRNA TARGET SITES

one of its predicted miRNAs.

Check boxes can be used to filter the results visualized. In particular, users can choose to filter the interactions based on the type of predicted site (current or novel), the fact that the seed region is edited or not, the type of seed match (*6mer*, *7mer-A1*, *7mer-m8* and *8mer*) and the energy of the duplex.

Finally, similarly in miRandola, miR-EdiTar is connected to miRò, a web environment that provides users with miRNA functional annotations inferred through their validated and predicted targets [157].

4.2.3 Database implementation and web interface

All the data are collected and maintained up-to-date in a *MySQL database* (*v5.1*) running on an *Apache server* (*v2.2.15*). The web application was implemented in *Ruby on Rails* (*v2.3.5*), a framework based on the *MVC* (*Model-View-Controller*) design pattern, allowing a fast development and management of the application. The queries that the database allows to perform were coded leveraging on the association mechanisms between models that the framework provides. The interface makes use of the *Ajax* technology to improve the usability through a fast client-side update of selections and results.

4.2.4 Utility and discussion

The modifications of predicted miRNA binding sites are classified into two categories, based on whether the editing events occur in the seed region or in

4.2. MIR-EDITOR: A DATABASE OF PREDICTED A-TO-I EDITED MIRNA TARGET SITES

another part of the duplex. The replacement of adenosines with inosines in the seed region can change A-U matches into G-U wobbles which are sometimes tolerated, especially in the presence of compensatory matches elsewhere in the duplex, but which have been reported to weaken the interaction or even abrogate binding [173]. This process is shown in *Figure 4.13*.

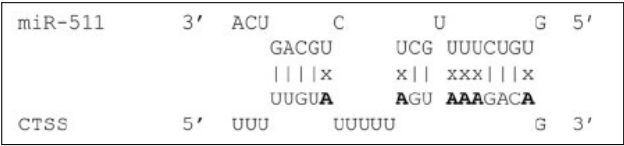


Figure 4.13: Predicted binding site for miR-511 in the 3' UTR of CTSS

Editing events that occur outside of the seed binding region could also influence targeting. They might either reduce the stability of the duplex, through the introduction of G-U wobbles and mismatches, or increase it by improving the seed match or by creating new matches outside the seed area, as specified in *Figure 4.14*.

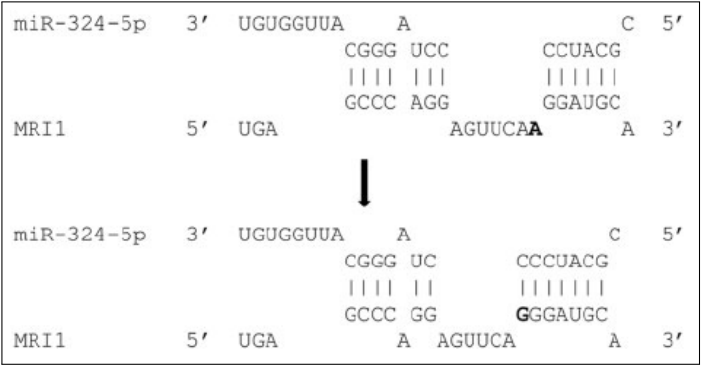


Figure 4.14: An edited adenosine in a potential binding site for *miR-324-5p* on the 3' UTR of *MRI1* may improve the seed match by adding an extra CG bond and changing the type from *7mer-A1* to *8mer*

The presence of inosines in miRNA binding sites could also alter their

4.2. MIR-EDITOR: A DATABASE OF PREDICTED A-TO-I EDITED MIRNA TARGET SITES

secondary structure and, as a consequence, increase or reduce the chances of binding. It has been demonstrated that single nucleotide polymorphisms (SNPs) can significantly change mRNA secondary structure [174, 175] and that changes in secondary structure can considerably affect the binding of miRNAs [176, 177]. Therefore, it is plausible that editing events may yield similar effects (see *Figure 4.15* and *4.16*).

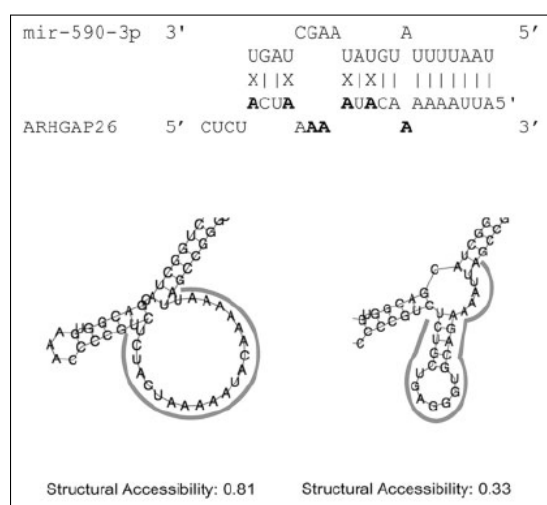


Figure 4.15: An example of variation of structural accessibility of predicted miRNA binding sites affected by A-to-I editing. The estimated structural accessibility of a predicted binding site for *miR-590-3p* in the 3' UTR of the gene *ARHGAP26* decreases by 40% due to editing events. The predicted interactions are shown along with the secondary structures of the unedited and edited versions of the binding sites

4.2. MIR-EDITOR: A DATABASE OF PREDICTED A-TO-I EDITED MIRNA TARGET SITES

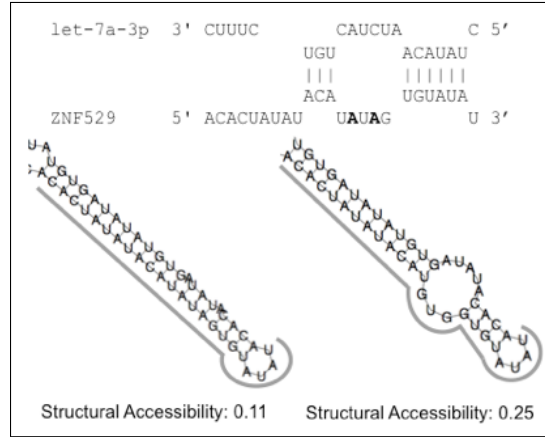


Figure 4.16: Example of variation of structural accessibility of predicted miRNA binding sites affected by A-to-I editing. Two edited adenosines in a non-seed area of the binding site for *let-7a-3p* on the 3' UTR of *ZNF529* increase the estimated degree of accessibility 2.15 times. The predicted interactions are shown along with the secondary structures of the un-edited and edited versions of the binding sites. Structural accessibility is computed as the average probability of stretches of 4 nucleotides to be unpaired in the predicted binding sites. Individual probabilities are calculated by the tool RNAPfold on 40 nt windows. Secondary structures of the targets are shown as computed by RNAfold on an 80 nt window encompassing the predicted binding site.

Other than affecting existing miRNA binding sites, A-to-I editing can generate novel miRNA/target interactions by either changing mature miRNA sequences or creating new sites on UTRs, as already reported by a few studies [153, 164]. As a proof of principle one of the predicted novel binding sites, the gene *MDM4*, was validated. It is an important negative regulator of the tumor suppressor *p53* (Markey 2011). The 3' UTR of *MDM4* presented a cluster of 4 edited adenosines generating a novel binding site for *miR-500a-3p*. A fragment of the *Wild Type* (**WT**) 3' UTR of *MDM4* gene containing the predicted binding site was cloned downstream of the luciferase gene on

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

a reporter construct. A mutant version of the plasmid (*MUT*) mimicking the editing events was generated by replacing the adenosines reported to be edited into guanosines (*Figure 4.17a,b*).

Moreover, *H460* cells (*non-small cell lung carcinoma*) was transfected with the luciferase reporter construct along with a precursor of *miR-500a-3p* or a scramble miRNA as negative control. There were not any significant difference in the luciferase activity between cells transfected with the WT plasmid along with either the scramble miRNA or *miR-500a-3p* precursor. On the contrary, a 32% reduction in the luciferase activity ($P < 0.01$) was observed in cells transfected with *MUT* and the *miR-500a-3p* precursor compared to cells transfected with *MUT* and the scramble miRNA (see *Figure 4.17c*). This data clearly confirms that the editing process can produce new binding sites for miRNAs on specific regions of the 3'UTR of a gene.

All these hypotheses and preliminary experiments suggest a new layer of dynamic regulation in miRNA-mediated gene expression control and encourage further investigations.

4.3 VIRGO: Visualization of A-to-I RNA editing sites in genomic sequences

4.3.1 Databases of RNA Editing sites

Few systems are available on the web. The first web-oriented database for annotated RNA editing sites was *dbRES*, but the last update goes back to 2007 and contains only a few dozen of human editing sites [178].

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

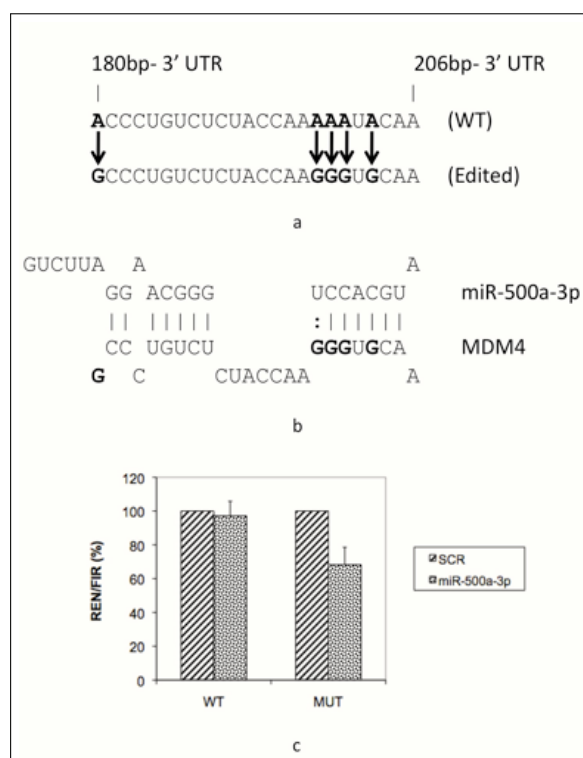


Figure 4.17: Experimental validation of a novel predicted site for *miR-500a-3p* created by editing in the 3' UTR of *MDM4*. **(a)** A 24 nt long fragment of the 3' UTR sequence of *MDM4* with 5 edited adenosines and the corresponding mutated version mimicking the editing events. **(b)** The predicted duplex of the miRNA/target interaction created by the editing events. **(c)** *Renilla luciferase* activity following co-transfection of a negative control miRNA (*SCR*) and *miR-500a-3p* along with the non-edited luciferase reporter construct (*WT*) and its mutated version (*MUT*) into *H460* cells. A 32% reduction in the luciferase activity ($P < 0.01$) is observed in the cells transfected with *MUT* and the *miR-500a-3p* precursor compared to the cells transfected with *MUT* and the negative control miRNA. No effect is observed in the cells transfected with *miR-500a-3p/SCR* and *WT*.

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

A few years later, Kiran and Baranov created **DARNED** [168], a database of human RNA editing sites providing a centralized access to published data. RNA editing locations are mapped on the reference human genome. DARNED is periodically updated and at the time of the writing of the thesis it contains more than 300,000 editing sites, but no statistical significance is provided [179].

In 2011, Picardi et al. presented **Expedit** [180]. It is a web application that maps data and, given individual sequence reads as input, executes a comparative analysis against DARNED editing sites. No statistical significance of results is given.

More recently, Ramaswami and Li have created **RADAR** a rigorously annotated database of A-to-I RNA editing in human, mouse and drosophila [181]. RADAR is the the largest dabatabe of human RNA editing which contains more than 1.4 million of A-to-I editing sites.

4.3.2 The creation of VIRGO

VIRGO¹⁰ (*Visualization of A-to-I RNA editing sites into GenOmic sequences*) is a knowledge-base equipped with a web-interface allowing users to map putative and known A-to-I editing sites into gene regions (including coding sequences, introns, and UTRs) [156]. In this work is considered as putative editing sites A-to-G mismatches between genomic and EST sequences, while known A-to-I editing sites are obtained from DARNED.

VIRGO borrows from literature the basic computational techniques that are used to identify A-to-G mismatches as putative editing sites. These

¹⁰Available on web in <http://atlas.dmi.unict.it/virgo/>

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

bioinformatics methods and resources (i.e. alignment between genomic and EST sequences, clustering, double strand RNA region identification, Next Generation Sequencing data) are then integrated into a workflow (see *Figure 4.18*) allowing users to facilitate the analysis of genomic sequences.

In particular, the VIRGO knowledge-base has been created by matching all the human genes regions obtained from UCSC (*hg19*) to the EST database using filters and NGS data. The filters allow the selection of candidate editing events in clusters [182], lying in repeated and double strand regions and not classified as SNPs. Moreover, VIRGO locally maps all the editing events stored in DARNED. This feature allows the visualization of all DARNED editing sites through the VIRGO web interface.

Finally, VIRGO uses the DARNED editing sites for which NGS information is available to compute the expected frequencies of A to G substitution that can happen in a mismatch aligned column. This knowledge is then used to compute p-values for all VIRGO editing events for which NGS information is available. The VIRGO web interface allows annotation of genomic sequences, provided by users, known editing sites and those sites passing the filters described above.

4.3.3 Construction and content

VIRGO is a knowledge base that integrates information retrieved from specialized biological databases. The core of the system has been developed in *C++*, while the front-end consists of a web interface developed in PHP.

The data integration process implemented in VIRGO consists of a se-

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

quence of steps carried out to identify putative A-to-I editing sites (see *Figure 4.18*). The database construction, which has been done offline, includes six steps. All filters are mandatory, therefore, a site that does not pass one of such steps is discarded. The last step is applied only when mismatches align with the NGS reads. The steps are described below.

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

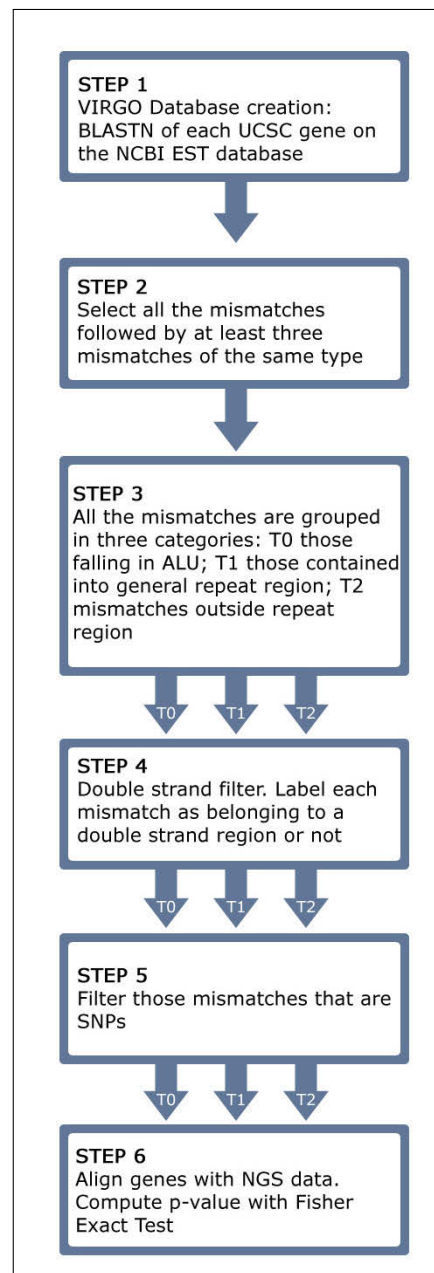


Figure 4.18: Sequence of steps to identify putative A-to-I editing sites.

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

Step 1: Blasting EST sequences in Human Genoma

In the first phase, the whole set of human genes from UCSC¹¹ has been downloaded. Then, using BLASTN, all the genes with NCBI EST database have been aligned. Although this step is very time consuming, it allows to identify all the potential A-to-I editing sites. VIRGO creates an initial dataset by selecting the A-G mismatches between the genes and the EST sequences.

Step 2: Clustering filter

According to [36], editing events usually happen in cluster. After binding the mRNA, ADAR creates bunches of close editing events. An edited sequence typically shows editing in many close-by sites. Therefore, it is very unlikely to observe isolated editing events inside a sequence. The clustering filter implements the methodology presented in [182] by selecting A-G mismatches that are followed by at least three mismatches of the same kind, without gaps or other types of mismatches (see *Figure 4.19* for an example).

Step 3: Partitioning of mismatches

VIRGO partitions the selected mismatches in three categories. To achieve that, the genes are falling in *ALU regions* (**T0**), in *repeat regions* (**T1**), and in *non repeat region* (**T2**) have been labelled.

¹¹It is available in <http://genome.ucsc.edu/buildGRCCh37/hg19>

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

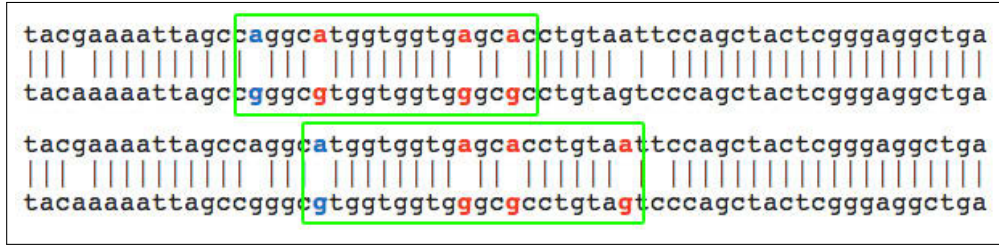


Figure 4.19: **Clustering filter**: The A-G mismatch in *blue* color is followed by three mismatches of the same type (in *red* color). Furthermore, no gaps are present. The three mismatches following the initial candidate editing site are included as putative editing events and are highlighted in the alignment with ESTs.

Step 4: Double strand filter

VIRGO verifies whether mismatches (from all the classes created above) occur into double-stranded regions. For this purpose a technique already used in [6,36] for the prediction of the double strand portion of a RNA secondary structure has been applied. It creates a short reverse complementary sequence centered on each mismatch by retrieving upstream and downstream flanking nucleotides. Then it searches for the constructed reverse complementary sequence into the gene where the mismatch has been found. In particular, when a mismatch occurs into an ALU repetitive region the length of the short complementary sequence is equal to the length of the ALU region. Otherwise, the length of the short sequence is equal to 251 nucleotides including the mismatch.

Next, VIRGO aligns the created sequence with a region with no more than 4001 nucleotides centered on the A-G mismatch. Since the length of the reverse complement in ALU and repeat regions is not constant we set the minimum length for the alignments to be 85% of the length of the sequence

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

(i.e. the alignment consists of at least 214 nucleotides over 251). Consequently, in the alignment we look for an identity of at least 85%. VIRGO annotates that mismatch as occurring into a double-strand region [6,36] (see *Figure 4.20* for an example).

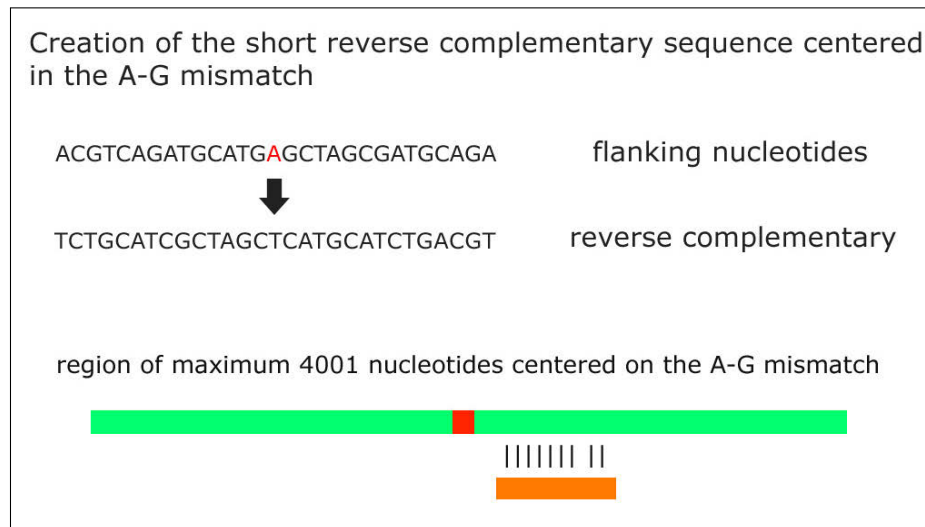


Figure 4.20: **Fourth Step** of VIRGO. The A-G mismatch in *blue* color is followed by three mismatches of the same type (in *red* color). Furthermore, no gaps are present. The three mismatches following the initial candidate editing site are included as putative editing events and are highlighted in the alignment with ESTs.

Step 5: Filterig of SNPs

VIRGO, uses the database *All SNPs(135)* contained in UCSC, to filter the mismatches that are already classified as SNPs.

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

Step 6: Calculating of the statistical significance

VIRGO performs an alignment of the genes with a subset of NGS data taken from the following experiments: *SRP002274* - *GSE19166*¹² and *SRP007465*¹³.

The subset of short reads is constructed as follows. Alignment of human genome with short reads has been performed by BOWTIE [82]. In order to reduce noise, only the best alignments with at most two mismatches by using *-a* and *-v* parameters was accepted. By specifying *-a*, VIRGO instructs BOWTIE to report all valid alignments, subjected to the alignment policy *-v 2* (at most two mismatches was allowed).

The selected short reads has been mapped on each VIRGO mismatch, selecting those mismatches occurring into at least five short reads. This alignment allows to compute, for some of the editing events, *p*-value and *adjusted p*-value yielding the confidence that the candidate mismatch is not a false positive.

The approach to compute the *p*-values of candidate sites uses the expected A/G frequencies in the aligned columns versus the observed one in connection to a Fisher exact test. To compute these expected frequencies we used all the DARNED editing sites having an alignment with some NGS reads (it has been set to five the minimum number of reads aligning the gene region). In order to calculate the *p*-value, for each selected mismatch the nucleotides present in the corresponding alignment columns has been considered. Only columns containing adenosine and guanosine are taken into account. For

¹²Available in <http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=SRP002274>

¹³Available in <http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=SRP007465>

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

each editing site reported in DARNED and aligned with the NGS reads the frequencies of A and G nucleotides in the column corresponding to the mismatch have been computed. Then, we take the average frequencies of A and G for all aligned DARNED editing sites. It has considered as observed frequencies those coming from a mismatch visualized by VIRGO which has an alignment with NGS reads. These frequencies (*expected/observed*) was then used through the *Fisher's Exact Test* to compute the putative site *p*-value (see *Figure 4.21* for an example). The significance of those mismatches for which it was not possible to compute the *p*-values was annotated as unknown.

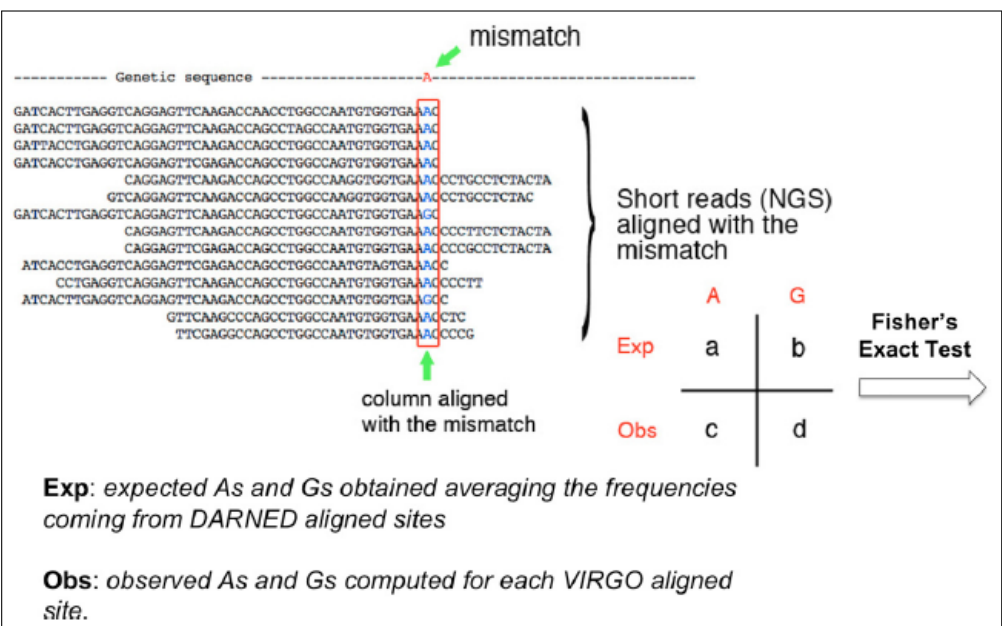


Figure 4.21: Example for the *p*-value computation.

Finally, *p*-values have been adjusted applying *FDR correction* for testing multiple hypotheses, with $\alpha = 0.01$. Each *p*-value is periodically updated by using new NGS experiments.

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

4.3.4 Utility and discussion

VIRGO aims to be an efficient and user-friendly system, providing an interface by which users can analyze and visualize their data, and export results into *xml* and *txt* files.

The central purpose of VIRGO is to provide users with a periodically updated system storing high quality candidate editing sites. This will allow users to quickly and easily identify whether their genomic sequences are subject to A-to-I RNA Editing or not.

The user can submit an input file containing headers of sequences in a specific BED-like format. Once the analysis starts, a temporary page containing a link to the results page is generated (see *Figure 4.22*).

The left part of the results page shows the sequences that have been analyzed. Each sequence is partitioned into segments of 80 nucleotides each. All known mismatches (obtained from DARNED) are identified by blue marks placed on top of them (see number 1 in *Figure 4.22*). In *Figure 4.23* it is shown, through a Venn diagram, the number of common sites shared by VIRGO and DARNED.

Only a small portion of VIRGO editing sites overlaps with those present in DARNED, for several reasons. First of all, RNA editing is a dynamic event; this means that the presence of edited adenosines can have, in principle, a strong variability. For example, a sequenced transcript can have an edited adenosine in a specific position in an experiment, which is absent in the same sequenced transcript in a second experiment. This conjecture is supported by the fact that most of the data included in DARNED come from experiments

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

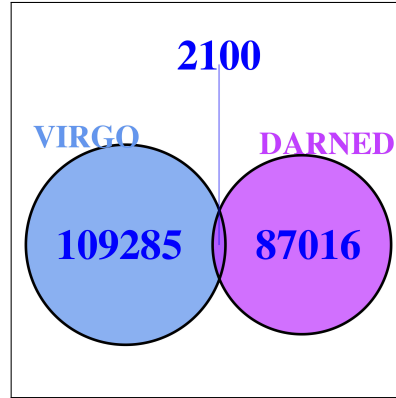


Figure 4.23: Venn diagram concerning the number of editing sites in common between VIRGO and DARNED.

in which authors synthesized their own ETSs or NGS transcripts. Within this context, tools as Virgo are useful to help to investigate. A second reason relies on the fact that the second phase (clustering filter) of VIRGO hides those candidate editing events that do not happen in clusters. However, since editing is rarely an-all-or-nothing mechanism, this dataset, based on the actual EST sequence reads, gives an accurate measure for the editing events occurring in vivo.

The sites identified by VIRGO are marked with different colors (*yellow, orange, red, purple*) according to the *Number of Aligned ESTs (NAEs)*. The colors with respect to the NAEs are:

- *yellow*: $1 \leq NAE \leq 5$;
- *orange*: $5 < NAE \leq 10$;
- *red*: $10 < NAE \leq 20$;
- *fuchsia*: $NAE \leq 20$.

4.3. VIRGO: VISUALIZATION OF A-TO-I RNA EDITING SITES IN GENOMIC SEQUENCES

They are placed at the bottom of sequences (see *number 2* in *Figure 4.22*). By clicking on a *blue marker*, VIRGO shows the following information: *chromosome, genomic position, strand, p-value, tissue/organ* (if known), if it is a *SNP* and the *PUBMED resources*.

Markers relative to newly predicted sites will give information on chromosome, genomic position, strand, and p-value. When a mismatch occurs inside a repeat region, its start/end genomic position, strand, chromosome, name, class and family will be given. It is given the list of EST sequences in which the mismatch occurs. For each EST sequence, VIRGO shows the EST name, tissue and organ (if known), the alignment between the input gene and EST sequence, and the NCBI information. The list of isoforms where the mismatch occurs is also provided. For each isoform, information such as the *refSeq ID, chromosome, strand, starting* and *ending* genomic position, among others, are provided (see *number 3, 4 and 5* in *Figure 4.22*).

Finally, the results of the analysis will be stored into the server for 5 days and then removed.

Part III

HMMs and their Application to miRNA Targeting

Chapter 5

Profile HMM for microRNA Target and Design

*Nature shows us only the tail of
the lion. But there is no doubt
in my mind that the lion belongs
with it even if he cannot reveal
himself to the eye all at once
because of his huge dimension.*

Albert Einstein

Nobel Prize in Physics (1921)

5.1 Introduction to profile HMM for microRNA target

Since its definition, profile HMM has been implemented in several bioinformatics tools [183] in which there is the need of multiple alignment of

5.1. INTRODUCTION TO PROFILE HMM FOR MICRORNA TARGET159

sequences.

Profile HMM are particularly powerful since they allow to define family of conserved sequences (i.e. RFAM [184], PFAM [185]). In the application of profile HMM to microRNA targeting the problem is different. In this case we have two sequences as input, the microRNA and the mRNA.

Therefore the parameters of the machine are conditioned with respect to the nucleotides present in the miRNA. The architecture of the profile HMM is presented in *Figure 5.1*:

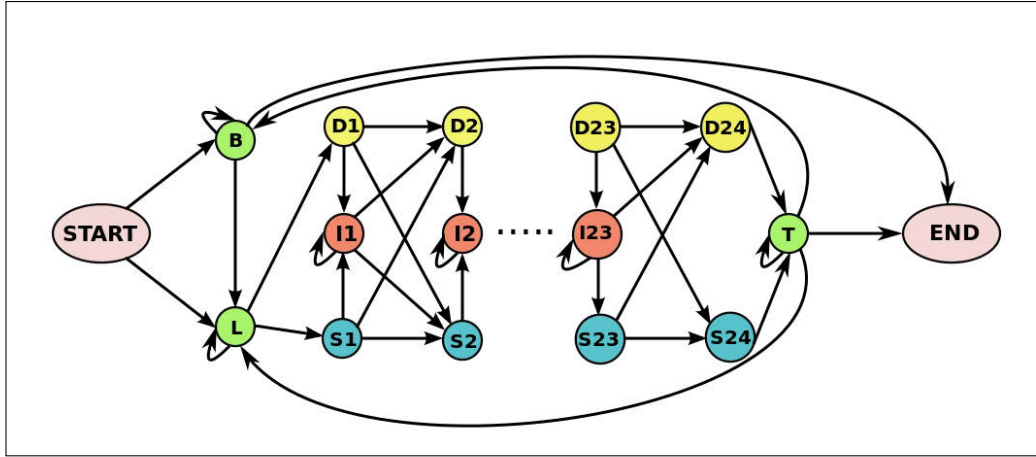


Figure 5.1: Transition structure of the profile HMM for the microRNA targeting.

Notice that in this model we have a portion of the model that is positional, that is the one corresponding to the profile, and a portion that is not positional, in which the nucleotide of mRNA sequence is considered background. In the *Figure 5.1* the nodes in *yellow* indicate the states of *deletions*, the ones in *red* to indicate the states of *insertions*, while the ones in *blue* to indicate the states of matching between a nucleotide belonging to the mRNA target

5.1. INTRODUCTION TO PROFILE HMM FOR MICRORNA TARGET160

<i>Start</i>
<i>B: Background</i>
<i>L: Leader</i>
<i>T: Trailer</i>
<i>I_i: insertion of microRNA target site position $i = 1, 2, \dots, 23$</i>
<i>D_i: delete of microRNA target site position $i = 1, 2, \dots, 24$</i>
<i>S_i: match of microRNA target site position $i = 1, 2, \dots, 24$</i>
<i>End</i>

Table 5.1: States of the profile HMM for MicroRNA targeting. Notice that the description is done assuming that the length of miRNAs is 23 nucleotides.

and a microRNA that acts as the regulator of gene expression.

All states in the schema above are shown in [Table 5.1](#).

In the first phase were designed and implemented *Forward* and *Backward* algorithms, which were subsequently integrated in the *Baum-Welch* algorithm for the parameter estimation.

In order to be guaranteed the reliability of the results, was used as training set the ***MiRecoord*** database [186], which contains experimental validated profiles of alignment between miRNAs and mRNAs. In addition, the decoding algorithm to has been implemented, in which given a set of observable states (in our case represented by the sequence of the microRNA and mRNA) the aim is that to find the most likely hidden states that determine the *pairwise alignment* between the two molecules.

We implemented few algorithms for the targeting through profile HMM. We started with Viterbi, however we noticed a very poor behavior in the recovery of experimentally validated binding sites. Then we implanted a posterior decoding based on γ -centroid and a stochastic backtrace.

Our experiments show that γ -centroid based decoding is the most reliable

in terms of quality.

5.2 Forward and Backward Algorithms

To implement the forward and backward algorithms we need matrices to store the probability of both parts of the machine.

We have 3 vectors concerning the non positional part of the machine:

- f_B represents the *background* probability at each sequence position.
- f_L represents the *leader* probability at each sequence position.
- f_T represents the *trailer* probability at each position.

Then, we have the positional part of the machine, composed by 3 matrices.

- f_{M_k} stores the *matching* probability position across all the microRNA length.
- f_{I_j} stores the *insertion* probability position across all the microRNA length but one.
- f_{D_j} stores the *deletion* probability position across all the microRNA length.

The same number of matrices are used for the backward algorithm.

5.2.1 Forward Algorithm

Let $X = x_1, \dots, x_T$ be the mRNA, and let $Y = Y_1, \dots, Y_N$ be microRNA.

The following equations refer to the forward algorithm in logspace.

Initialization ($i = 0$):

$$F^B(0) = 0, F^L(0) = -\infty, F^T(0) = -\infty,$$

$$F_j^M(0) = -\infty \forall j = 1 \dots N,$$

$$F_j^I(0) = -\infty \forall j = 1 \dots N - 1,$$

$$F_j^D(0) = -\infty \forall j = 1 \dots N$$

Recursion ($i = 1, \dots, T, j = 1, \dots, N$):

$$\begin{aligned}
F^B(i) &= \log(e_B(x_i)) && + \log [a_{B,B} \exp(F^B(i-1)) \\
&&& + a_{T,B} \exp(F^T(i-1))] \\
F^L(i) &= \log(e_L(x_i)) && + \log [a_{L,L} \exp(F^L(i-1)) \\
&&& + a_{B,L} \exp(F^B(i-1)) + a_{T,L} \exp(F^T(i-1))] \\
F^T(i) &= \log(e_T(x_i)) && + \log [a_{T,T} \exp(F^T(i-1)) \\
&&& + a_{M_N,T} \exp(F_N^M(i-1)) + a_{D_N,T} \exp(F_N^D(i-1))] \\
\\
F_1^M(i) &= \log(e_{M_1}(x_i|y_1)) && + \log [a_{L,M_1} \exp(F^L(i-1))] \\
F_1^D(i) &= \log [a_{L,D_1} \exp(F^L(i-1))] \\
F_1^I(i) &= \log(e_{I_1}(x_i)) && + \log [a_{M_1,I_1} \exp(F_1^M(i-1)) \\
&&& + a_{I_1,I_1} \exp(F_1^I(i-1)) + a_{D_1,I_1} \exp(F_1^D(i-1))] \\
\\
F_j^M(i) &= \log(e_{M_j}(x_i|y_j)) && + \log [a_{M_{j-1},M_j} \exp(F_{j-1}^M(i-1)) \\
&&& + a_{I_{j-1},M_j} \exp(F_{j-1}^I(i-1)) + a_{D_{j-1},M_j} \exp(F_{j-1}^D(i-1))] \\
F_j^I(i) &= \log(e_{I_j}(x_i)) && + \log [a_{M_j,I_j} \exp(F_j^M(i-1)) \\
&&& + a_{I_j,I_j} \exp(F_j^I(i-1)) + a_{D_j,I_j} \exp(F_j^D(i-1))] \\
F_j^D(i) &= && \log [a_{M_{j-1},D_j} \exp(F_{j-1}^M(i)) \\
&&& + a_{I_{j-1},D_j} \exp(F_{j-1}^I(i)) + a_{D_{j-1},D_j} \exp(F_{j-1}^D(i))]
\end{aligned}$$

Termination:

$$\log(P(X|Y)) = \log [\exp(F^B(T)) + \exp(F^T(T))]$$

5.2.2 Backward Algorithm

The backward algorithm is implemented in logspace.

Initialization ($i = 0$):

$$B^B(T) = 0, \quad B^L(T) = -\infty, \quad B^T(T) = 0,$$

$$B_j^M(T) = -\infty \quad \forall j = 1 \dots N,$$

$$B_j^I(T) = -\infty \quad \forall j = 1 \dots N - 1$$

$$B_j^D(T) = -\infty \quad \forall j = 1 \dots N$$

Recursion ($i = T - 1, \dots, 1, j = N, \dots, 1$):

$$B^B(i) = \log [e_B(x_{i+1})a_{B,B} \exp(B^B(i+1)) \\ + e_L(x_{i+1})a_{B,L} \exp(B^L(i+1))]$$

$$B^T(i) = \log [e_T(x_{i+1})a_{T,T} \exp(B^T(i+1)) \\ + e_L(x_{i+1})a_{T,L} \exp(B^L(i+1)) + e_B(x_{i+1})a_{T,B} \exp(B^B(i+1))]$$

$$B_N^M(i) = \log [e_T(x_{i+1})a_{M_N,T} \exp(B^T(i+1))]$$

$$B_N^D(i) = \log [a_{D_N,T} \exp(B^T(i+1))]$$

$$B_j^M(i) = \log [e_{M_{j+1}}(x_{i+1}|y_{j+1})a_{M_j,M_{j+1}} \exp(B_{j+1}^M(i+1)) \\ + e_{I_j}(x_{i+1})a_{M_j,I_j} \exp(B_j^I(i+1)) + a_{M_j,D_{j+1}} \exp(B_{j+1}^D(i))]$$

$$B_j^I(i) = \log [e_{M_{j+1}}(x_{i+1}|y_{j+1})a_{I_j,M_{j+1}} \exp(B_{j+1}^M(i+1)) \\ + e_{I_j}(x_{i+1})a_{I_j,I_j} \exp(B_j^I(i+1)) + a_{I_j,D_{j+1}} \exp(B_{j+1}^D(i))]$$

$$B_j^D(i) = \log [e_{M_{j+1}}(x_{i+1}|y_{j+1})a_{D_j,M_{j+1}} \exp(B_{j+1}^M(i+1)) \\ + e_{I_j}(x_{i+1})a_{D_j,I_j} \exp(B_j^I(i+1)) + a_{D_j,D_{j+1}} \exp(B_{j+1}^D(i))]$$

$$B^L(i) = \log [e_L(x_{i+1})a_{L,L} \exp(B^L(i+1)) \\ + e_{M_1}(x_{i+1}|y_1)a_{L,M_1} \exp(B_1^M(i+1)) + a_{L,D_1} \exp(B_1^D(i))]$$

Termination

$$\log(P(X|Y)) = \log[e_B(x_1)a_{B,B} \exp(B^B(0))]$$

5.2.3 LogSum Trick

To avoid underflow problems, the sum of log of probabilities is computed using the following equality. Let x be the $\max(x, y)$, we have: $\log(e^x + e^y) = x + \log(1 + e^{y-x})$. For the general case, given a vector of log of probabilities (x_1, \dots, x_m) , let y be the $\max(x_i)$, the log sum in this case will be:

$$\log\left(\sum_{i=1}^m e^{x_i}\right) = y + \log\left(\sum_{i=1}^m e^{(x_i-y)}\right) \quad (5.1)$$

5.3 Baum-Welch

Forward and Backward probabilities are obtained from logs through exponentiation. Given K pairs of mRNAs and microRNAs $\{\langle X_1, Y_1 \rangle, \dots, \langle X_K, Y_K \rangle\}$, the expected emission counts for all the sequences in the training set are obtained with the following equations:

$$\begin{aligned} E_B(a) &= \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_{i|x_i^k=a} \exp(F_k^B[i] + B_k^B[i]) \\ E_L(a) &= \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_{i|x_i^k=a} \exp(F_k^L[i] + B_k^L[i]) \\ E_T(a) &= \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_{i|x_i^k=a} \exp(F_k^T[i] + B_k^T[i]) \end{aligned}$$

$$E_{M_j}(a|y_j = b) = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_{i|x_i^k=a, y_j^k=b} \exp(F_k^{M_j}[i] + B_k^{M_j}[i])$$

$$E_{I_j}(a) = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_{i|x_i^k=a} \exp(F_k^{I_j}[i] + B_k^{I_j}[i])$$

The expected transition counts for all the sequences in the training set are obtained using the following equations:

$$A_{B,B} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^B[i]) a_{B,B} e_B(x_{i+1}) \exp(B_k^B[i+1])$$

$$A_{T,B} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^T[i]) a_{T,B} e_B(x_{i+1}) \exp(B_k^B[i+1])$$

$$A_{B,L} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^B[i]) a_{B,L} e_L(x_{i+1}) \exp(B_k^L[i+1])$$

$$A_{L,L} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^L[i]) a_{L,L} e_L(x_{i+1}) \exp(B_k^L[i+1])$$

$$A_{T,L} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^T[i]) a_{T,L} e_L(x_{i+1}) \exp(B_k^L[i+1])$$

$$A_{L,M_1} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^L[i]) a_{L,M_1} e_{M_1}(x_{i+1}|y_1) \exp(B_k^{M_1}[i+1])$$

$$A_{L,D_1} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^L[i]) a_{L,D_1} \exp(B_k^{D_1}[i])$$

$$A_{M_j,M_{j+1}} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^{M_j}[i]) a_{M_j,M_{j+1}} e_{M_{j+1}}(x_{i+1}|y_{j+1}) \exp(B_k^{M_{j+1}}[i+1])$$

$$A_{I_j,M_{j+1}} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^{I_j}[i]) a_{I_j,M_{j+1}} e_{M_{j+1}}(x_{i+1}|y_{j+1}) \exp(B_k^{M_{j+1}}[i+1])$$

$$A_{D_j,M_{j+1}} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^{D_j}[i]) a_{D_j,M_{j+1}} e_{M_{j+1}}(x_{i+1}|y_{j+1}) \exp(B_k^{M_{j+1}}[i+1])$$

$$A_{M_N,T} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^{M_N}[i]) a_{M_N,T} e_T(x_{i+1}) \exp(B_k^T[i+1])$$

$$A_{D_N,T} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(D_k^{M_N}[i]) a_{D_N,T} e_T(x_{i+1}) \exp(B_k^T[i+1])$$

$$A_{T,T} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^T[i]) a_{T,T} e_T(x_{i+1}) \exp(B_k^T[i+1])$$

$$A_{M_j,D_{j+1}} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^{M_j}[i]) a_{M_j,D_{j+1}} \exp(B_k^{D_{j+1}}[i])$$

$$A_{I_j,D_{j+1}} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^{I_j}[i]) a_{I_j,D_{j+1}} \exp(B_k^{D_{j+1}}[i])$$

$$A_{D_j,D_{j+1}} = \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^{D_j}[i]) a_{D_j,D_{j+1}} \exp(B_k^{D_{j+1}}[i])$$

$$\begin{aligned}
A_{M_j, I_j} &= \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^{M_j}[i]) a_{M_j, I_j} e_{I_j}(x_{i+1}) \exp(B_k^{I_j}[i+1]) \\
A_{I_j, I_j} &= \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^{I_j}[i]) a_{I_j, I_j} e_{I_j}(x_{i+1}) \exp(B_k^{I_j}[i+1]) \\
A_{D_j, I_j} &= \sum_{k=1}^K \frac{1}{P(X_k|Y_k)} \sum_i \exp(F_k^{D_j}[i]) a_{D_j, M_{j+1}} e_{I_j}(x_{i+1}) \exp(B_k^{I_j}[i+1])
\end{aligned}$$

The expected counts of A and E are initialized using the data from the training set regularized with α s from Dirichlet. The Baum-Welch Algorithm pseudocode used for the training is given below.

Initialization:

Pick as parameters the counts from training set regularized with the Dirichlet priors.

Recurrence:

1. Set all A and E variables to their pseudocounts.
2. For each sequence in the training set $j = 1, \dots, n$.
 - 2.1 Compute forward probability for sequence j .
 - 2.2 Compute backward probability for sequence j .
 - 2.3 Add the contribution of sequence j to A and E according the the equations above.

3. Update the model parameters.
4. Compute the new log likelihood.

Termination:

Stop if the change in the likelihood is less than $1e^{-5}$, or the number of iterations exceeds 100.

5.4 Gamma-Centroid Decoding

As in Viterbi we have 3 vectors concerning the non-positional part of the machine (G^B , G^L , G^T) and 3 for the profile one (G^M , G^D , G^I).

Let $X = x_1, \dots, x_T$ be the mRNA, and let $Y = Y_1, \dots, Y_N$ be microRNA.

Initialization ($i = 0$):

$$G^B(0) = 0, G^L(0) = 0, G^T(0) = -\infty,$$

$$G_j^M(0) = -\infty \forall j = 1 \dots N,$$

$$G_j^I(0) = -\infty \forall j = 1 \dots N - 1,$$

$$G_j^D(0) = -\infty \forall j = 1 \dots N$$

Recursion ($i = 1, \dots, T, j = 1, \dots, N$):

$$G^B(i) = \frac{f_B(i) \cdot b_B(i)}{P(X|Y)} + \max(G^B(i-1), G^T(i-1))$$

$$G^L(i) = \frac{f_L(i) \cdot b_L(i)}{P(X|Y)} + \max(G^L(i-1), G^B(i-1), G^T(i-1))$$

$$G^T(i) = \frac{f_T(i) \cdot b_T(i)}{P(X|Y)} + \max(G^T(i-1), G_N^M(i-1), G_N^D(i-1))$$

$$G_1^M(i) = (\gamma + 1) \cdot \frac{f_{M_1}(i) \cdot b_{M_1}(i)}{P(X|Y)} - 1 + G^L(i - 1)$$

$$G_1^D(i) = (\gamma + 1) \cdot \frac{f_{D_1}(i) \cdot b_{D_1}(i)}{P(X|Y)} - 1 + G^L(i)$$

$$G_j^M(i) = (\gamma + 1) \cdot \frac{f_{M_j}(i) \cdot b_{M_j}(i)}{P(X|Y)} - 1 + \max(G_{j-1}^M(i - 1), G_{j-1}^I(i - 1), G_{j-1}^D(i - 1))$$

$$G_j^I(i) = (\gamma + 1) \cdot \frac{f_{I_j}(i) \cdot b_{I_j}(i)}{P(X|Y)} - 1 + \max(G_j^M(i - 1), G_j^I(i - 1), G_j^D(i - 1))$$

$$G_j^D(i) = (\gamma + 1) \cdot \frac{f_{D_j}(i) \cdot b_{D_j}(i)}{P(X|Y)} - 1 + \max(G_{j-1}^M(i), G_{j-1}^I(i), G_{j-1}^D(i))$$

5.4.1 microRNA Design

Let $X = x_1, \dots, x_T$ be the mRNA, then:

$$P(X = x_i) = \sum_{M=m} P(M, R = r) \quad (5.2)$$

$$P(M^*|R) = \operatorname{argmax}_M P(M|R) \quad (5.3)$$

5.5 Stochastic backtrace procedure

Here, are defined the probabilities that are in the stochastic *traceback* procedure in order to determinate the hidden states.

- **Backward**

$$P(B_{i-1}|B_i) = \frac{f_B[i-1] \times \hat{a}_{B,B}}{f_B[i-1] \times \hat{a}_{B,B} + f_T[i-1] \times \hat{a}_{T,B}}$$

$$P(T_{i-1}|B_i) = \frac{f_T[i-1] \times \hat{a}_{T,B}}{f_B[i-1] \times \hat{a}_{B,B} + f_T[i-1] \times \hat{a}_{T,B}}$$

- *Leader*

$$P(L_{i-1}|L_i) = \frac{f_L[i-1] \times \hat{a}_{L,L}}{f_L[i-1] \times \hat{a}_{L,L} + f_B[i-1] \times \hat{a}_{B,L} + f_T[i-1] \times \hat{a}_{T,L}}$$

$$P(B_{i-1}|L_i) = \frac{f_B[i-1] \times \hat{a}_{B,L}}{f_L[i-1] \times \hat{a}_{L,L} + f_B[i-1] \times \hat{a}_{B,L} + f_T[i-1] \times \hat{a}_{T,L}}$$

$$P(T_{i-1}|L_i) = \frac{f_T[i-1] \times \hat{a}_{T,L}}{f_L[i-1] \times \hat{a}_{L,L} + f_B[i-1] \times \hat{a}_{B,L} + f_T[i-1] \times \hat{a}_{T,L}}$$

- *Trailer*

$$P(T_{i-1}|T_i) = \frac{f_T[i-1] \times \hat{a}_{T,T}}{f_T[i-1] \times \hat{a}_{T,T} + f_{M_N}[i-1] \times \hat{a}_{M_N,T} + f_{D_N}[i-1] \times \hat{a}_{D_N,T}}$$

$$P(M_{N,i-1}|T_i) = \frac{f_{M_N}[i-1] \times \hat{a}_{M_N,T}}{f_T[i-1] \times \hat{a}_{T,T} + f_{M_N}[i-1] \times \hat{a}_{M_N,T} + f_{D_N}[i-1] \times \hat{a}_{D_N,T}}$$

$$P(D_{N,i-1}|T_i) = \frac{f_{D_N}[i-1] \times \hat{a}_{D_N,T}}{f_T[i-1] \times \hat{a}_{T,T} + f_{M_N}[i-1] \times \hat{a}_{M_N,T} + f_{D_N}[i-1] \times \hat{a}_{D_N,T}}$$

$$P(L_{i-1}|D_{1,i}) = \frac{f_L[i] \times \hat{a}_{L,D_1}}{f_L[i] \times \hat{a}_{L,D_1}} = 1$$

$$P(L_{i-1}|M_{1,i}) = \frac{f_L[i-1] \times \hat{a}_{L,M_1}}{f_L[i-1] \times \hat{a}_{L,M_1}} = 1$$

- Base case of *Insertion State*

$$P(M_{1,i-1}|I_{1,i}) = \frac{f_{M_1}[i-1] \times \hat{a}_{M_1,I_1}}{f_{M_1}[i-1] \times \hat{a}_{M_1,I_1} + f_{D_1}[i-1] \times \hat{a}_{D_1,I_1} + f_{I_1}[i-1] \times \hat{a}_{I_1,I_1}}$$

$$P(D_{1,i-1}|I_{1,i}) = \frac{f_{D_1}[i-1] \times \hat{a}_{D_1,I_1}}{f_{M_1}[i-1] \times \hat{a}_{M_1,I_1} + f_{D_1}[i-1] \times \hat{a}_{D_1,I_1} + f_{I_1}[i-1] \times \hat{a}_{I_1,I_1}}$$

$$P(I_{1,i-1}|I_{1,i}) = \frac{f_{I_1}[i-1] \times \hat{a}_{I_1,I_1}}{f_{M_1}[i-1] \times \hat{a}_{M_1,I_1} + f_{D_1}[i-1] \times \hat{a}_{D_1,I_1} + f_{I_1}[i-1] \times \hat{a}_{I_1,I_1}}$$

- Match state

$$P(M_{j-1,i-1}|M_{j,i}) = \frac{f_{M_{j-1}}[i-1] \times \hat{a}_{M_{j-1},M_j}}{f_{M_{j-1}}[i-1] \times \hat{a}_{M_{j-1},M_j} + f_{I_{j-1}}[i-1] \times \hat{a}_{I_{j-1},M_j} + f_{D_{j-1}}[i-1] \times \hat{a}_{D_{j-1},M_j}}$$

$$P(D_{j-1,i-1}|M_{j,i}) = \frac{f_{D_{j-1}}[i-1] \times \hat{a}_{D_{j-1},M_j}}{f_{M_{j-1}}[i-1] \times \hat{a}_{M_{j-1},M_j} + f_{I_{j-1}}[i-1] \times \hat{a}_{I_{j-1},M_j} + f_{D_{j-1}}[i-1] \times \hat{a}_{D_{j-1},M_j}}$$

$$P(I_{j-1,i-1}|M_{j,i}) = \frac{f_{I_{j-1}}[i-1] \times \hat{a}_{I_{j-1},M_j}}{f_{M_{j-1}}[i-1] \times \hat{a}_{M_{j-1},M_j} + f_{I_{j-1}}[i-1] \times \hat{a}_{I_{j-1},M_j} + f_{D_{j-1}}[i-1] \times \hat{a}_{D_{j-1},M_j}}$$

- Deletion state

$$P(M_{j-1,i}|D_{j,i}) = \frac{f_{M_{j-1}}[i] \times \hat{a}_{M_{j-1},D_j}}{f_{M_{j-1}}[i] \times \hat{a}_{M_{j-1},D_j} + f_{I_{j-1}}[i] \times \hat{a}_{I_{j-1},D_j} + f_{D_{j-1}}[i] \times \hat{a}_{D_{j-1},D_j}}$$

$$P(D_{j-1,i}|D_{j,i}) = \frac{f_{D_{j-1}}[i] \times \hat{a}_{D_{j-1},D_j}}{f_{M_{j-1}}[i] \times \hat{a}_{M_{j-1},D_j} + f_{I_{j-1}}[i] \times \hat{a}_{I_{j-1},D_j} + f_{D_{j-1}}[i] \times \hat{a}_{D_{j-1},D_j}}$$

$$P(I_{j-1,i}|D_{j,i}) = \frac{f_{I_{j-1}}[i] \times \hat{a}_{I_{j-1},D_j}}{f_{M_{j-1}}[i] \times \hat{a}_{M_{j-1},D_j} + f_{I_{j-1}}[i] \times \hat{a}_{I_{j-1},D_j} + f_{D_{j-1}}[i] \times \hat{a}_{D_{j-1},D_j}}$$

- **Insertion state**

$$P(M_{j,i-1}|I_{j,i}) = \frac{f_{M_j}[i-1] \times \hat{a}_{M_j,I_1}}{f_{M_j}[i-1] \times \hat{a}_{M_j,I_j} + f_{D_j}[i-1] \times \hat{a}_{D_j,I_j} + f_{I_j}[i-1] \times \hat{a}_{I_j,I_j}}$$

$$P(D_{j,i-1}|I_{j,i}) = \frac{f_{D_j}[i-1] \times \hat{a}_{D_j,I_j}}{f_{M_j}[i-1] \times \hat{a}_{M_j,I_j} + f_{D_j}[i-1] \times \hat{a}_{D_j,I_j} + f_{I_j}[i-1] \times \hat{a}_{I_j,I_j}}$$

$$P(I_{j,i-1}|I_{j,i}) = \frac{f_{I_j}[i-1] \times \hat{a}_{I_j,I_j}}{f_{M_j}[i-1] \times \hat{a}_{M_j,I_j} + f_{D_j}[i-1] \times \hat{a}_{D_j,I_j} + f_{I_j}[i-1] \times \hat{a}_{I_j,I_j}}$$

The forward probability is obtained from Log-odds:

$$F(X) = \log \left(\frac{P(X)}{P^R(X)} \right)$$

thus,

$$P(X) = \exp(F(X) + \log(P^R(X)))$$

$$P^R(X) = \prod_{i=1}^n q_{x_i}$$

5.6 Results

The HMM has been trained and tested using a dataset of experimentally validated Human microRNA-gene interaction binding sites.

The data have been downloaded from miRecords¹ [186]. The validated targets component of miRecords hosts a large, high-quality manually curated database of experimentally validated miRNA-gene interactions with systematic documentation of experimental support for each interaction. The experimentally validated human miRNA-gene are 1,631.

From this dataset we selected only the 102 pairs referring to those data with experimentally validated binding sites. This 102 pairs consist of 126 binding sites drawn from 59 different microRNAs and 76 different mRNAs.

Two kinds of experiments have been done:

- (a) The first one uses the original mRNAs. The flanking sequences, of 10 nucleotides, preceding and succeeding the binding sites are set as Leader and Trailer states, the rest of the messenger is set as background.
- (b) In the second experiment we shuffle the mRNA keeping only the original binding site. In this case the flanking site length has been set to 1 or 2.

To perform a blind test analysis, the dataset has been randomly partitioned into training set (66% - 81 binding sites from 64 miRNA-gene)

¹It is available in: <http://mirecords.biolead.org/>

and testing set (34% - 45 binding sites from 38 miRNA-gene). In the *Figure 5.2*, are shown the results as *Positive Predicted Values/Sensitivity* (*PPV/SEN*) by varying the γ parameter after the training of the model with 10 iterations of Baum-Welch algorithm with properly provided prior distribution based on Dirichlet pseudo counts.

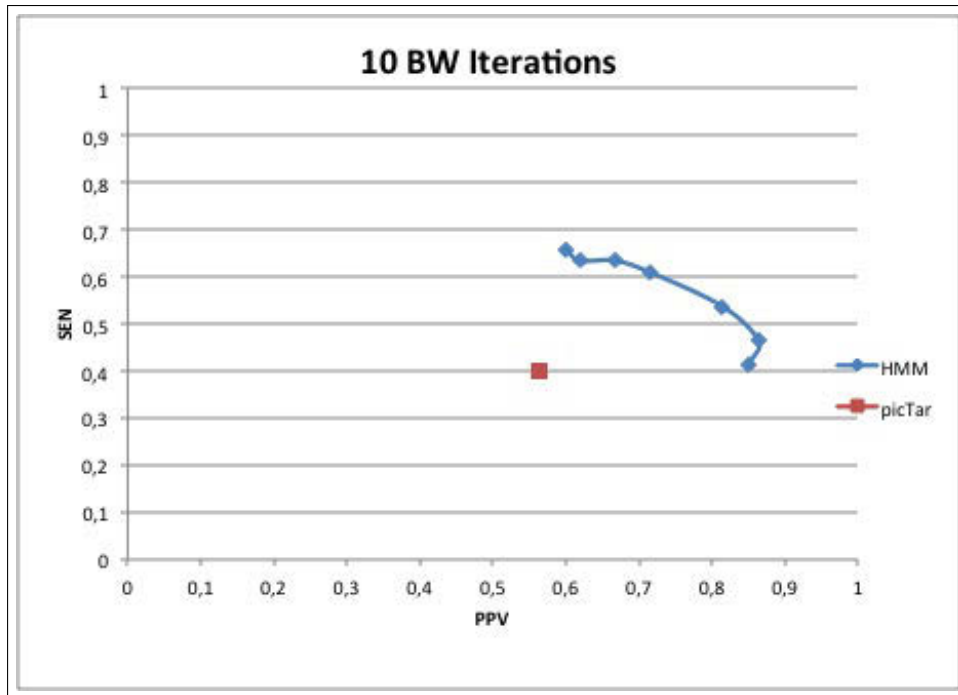


Figure 5.2: Comparison PicTar and Profile HMM for miRNA targeting (10 *Baum-Welch* iterations).

In the *Figure 5.3* it is possible to see the convergence of *Baum-Welch* algorithm after 70 iterations, considering the second training set.

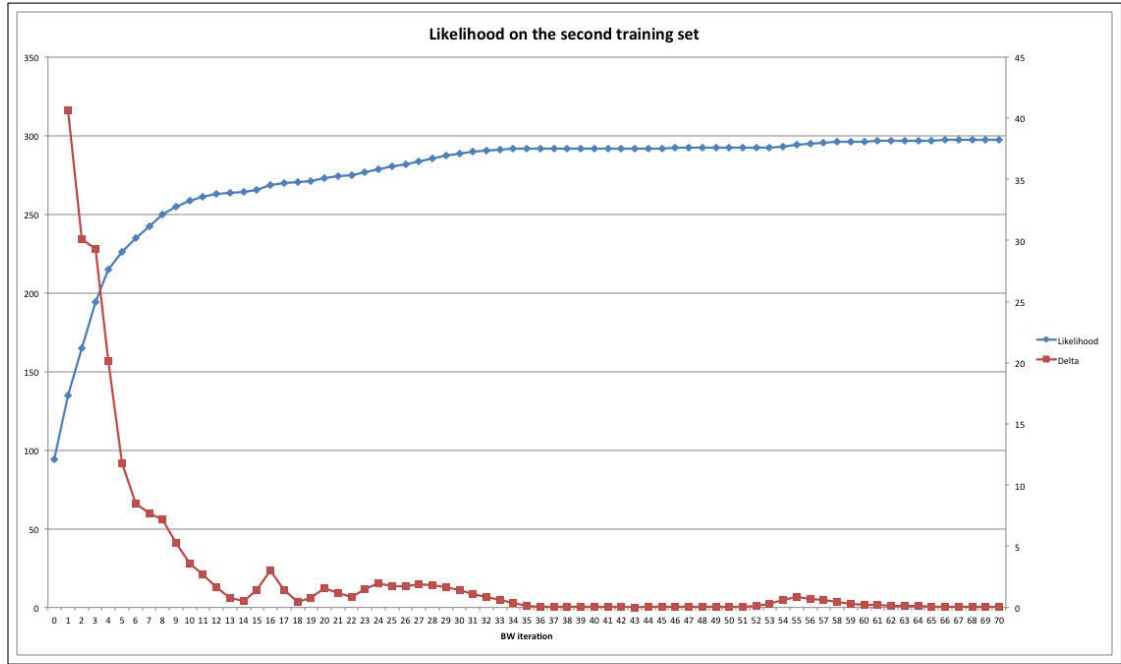


Figure 5.3: Convergence of Baum-Welch algorithm.

The HMM has also been tested on a dataset of experimentally validated false negatives (the training has been done on the training set of the point (a) above). The data consist of 33 pair of miRNA-gene for which the target gene expression is unaffected by the presence of the miRNA. These sequences have been downloaded from *TarBase* [187]. TarBase hosts detailed information for each miRNA/gene interaction, ranging from miRNA- and gene-related facts to information specific to their interaction, the experimental validation methodologies and their outcomes.

Part IV

Motif Discovery in RNA Editing Phenomenon

Chapter 6

A Systematic Method to Find Motifs Characterizing the A-to-I RNA Editing

A universe with a God would look quite different from a universe without one. A physics, a biology where there is a God is bound to look different. So the most basic claims of religion are scientific. Religion is a scientific theory.

Richard Dawkins

Oxford University

DESPITE the enormous efforts made in the last two decades, the real biological function of the RNA editing as well as the features of the substrates of the ADAR still remain unknown. This fourth part is dedicated to the

presentation of a preliminary methodological workflow for the identification of RNA editing structural motifs. The main object of the project is to discover potential sequence signals that appear only in genomic regions subject to editing.

If on the one hand the motif discovering is very challenging because the A-to-I editing in human often occurs in repetitive regions, on the other hand, if we focus on non-repetitive flanking regions of the editing sites we could identify not biased signals related to editing.

The A-to-I RNA editing occurs in regions of double strand RNA (dsRNA). The A-to-I editing can be either *specific* (if a single adenosine is edited within the dsRNA) or *promiscuous* (in the case that the adenosines edited are up to 50%). The specific editing occurs within a short double-strand region, as it happens for those editing sites that are formed in a mRNA in which the bases of the intronic sequences pair in a complementary way to the bases of the exon sequences. The promiscuous editing, instead, occurs within large duplex regions.

6.1 Description of the methodology

6.1.1 Preparation of the dataset

The first phase consists in the preparation of the dataset and, to ensure the reliability of the results, experimental validated (detected by the Sanger method [71]) editing sites have been collected by using the literature. In particular, these editing sites were divided into two categories: (**I**) *true-*

positive (**TP**) and (II) false-positive (**FP**) editing sites.

After collecting experimental validated editing sites, the flanking regions of 2,000 nucleotides downstream and upstream of editing sites were exacted (see *Figure 6.1*).

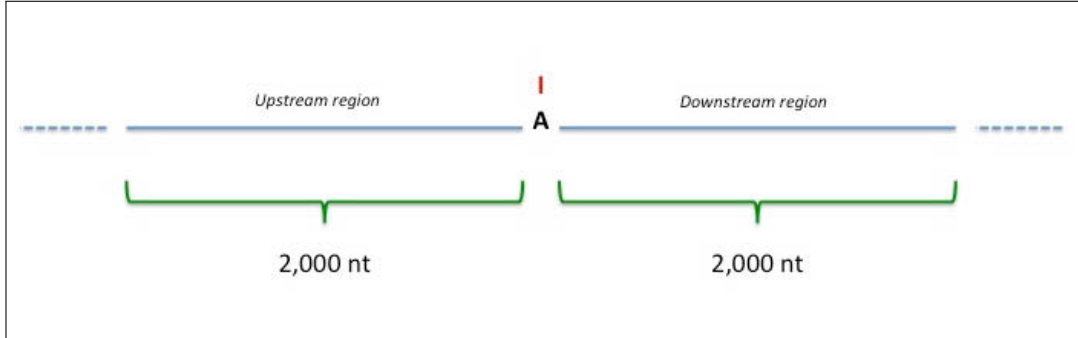


Figure 6.1: Upstream and downstream regions of editing site.

A gene may also contain a few hundreds of sites, as shown in the *Table 6.1*, where there are some examples of experimental validated editing sites that occur in *5HT_{2C}* gene. For this reason for each gene only one randomly *true-positive* editing site was taken into consideration as sample site, together with its flanking sequences (that we consider sample edited sequences). In total, we have 30 sample edited sequences.

6.1.2 Searching for motifs in edited sequences

The searching of motifs characterizing the A-to-I RNA editing phenomenon has been done through **MEME** (*Multiple EM for Motif Elicitation*), a software able to discover one or more motifs in a collection of DNA or protein sequences [188]. MEME uses the technique of the *expectation maximization* (as it happens in Baum-Welch algorithm, see in 2.7.5), adopted to fit a two-

<i>Genomic Position (hg19)</i>	<i>Chr</i>	<i>Strand</i>	<i>% of editing</i>	<i>Region</i>	<i>Amino Acid Change</i>	<i>Tissue</i>	<i>Pubmed ID</i>
114082684	X	+	41,78	EC	I -> M	Cerebellum	19478186, 22912834
114082688	X	+	15,12	EC	N -> D	Cerebellum	19478186, 22912834
114082689	X	+	45,87	EC	N -> S	Cerebellum	19478186, 22912834
114082694	X	+	45,87	EC	I -> V	Cerebellum	19478186, 22912834

Table 6.1: Examples of experimental validated editing sites in $5HT_{2C}$ gene. The *EC* value in *Region* column indicates that the editing site occurs in a *coding* region.

component finite mixture model to the set of the sequences. The algorithm discovers the number of times a motif takes place in each sequence of the dataset and it outputs an alignment of the occurrences of the motif.

Then, it has been applied a stand-alone version of MEME program to search for both 50 palindromic and 50 non-palindromic motifs in our sample edited sequences. In *Figures 6.2* and *6.3*, it can be seen the mapping of palindromic and non-palindromic motifs on sample edited sequences, respectively. In these two figures we can observe that non-palindromic motifs are significantly more present than palindromic ones on based their *combined p-value*.

6.2 Preliminary results

Once obtained palindromic and non-palindromic motifs these have been to mapped into the genome. In order to do this task, we initially subdivided each chromosome in regions of 4,000 nucleotides (in the *Table 6.2* is shown the number of regions for each chromosome).



Figure 6.2: Mapping of palindromic motifs on sample edited sequences.

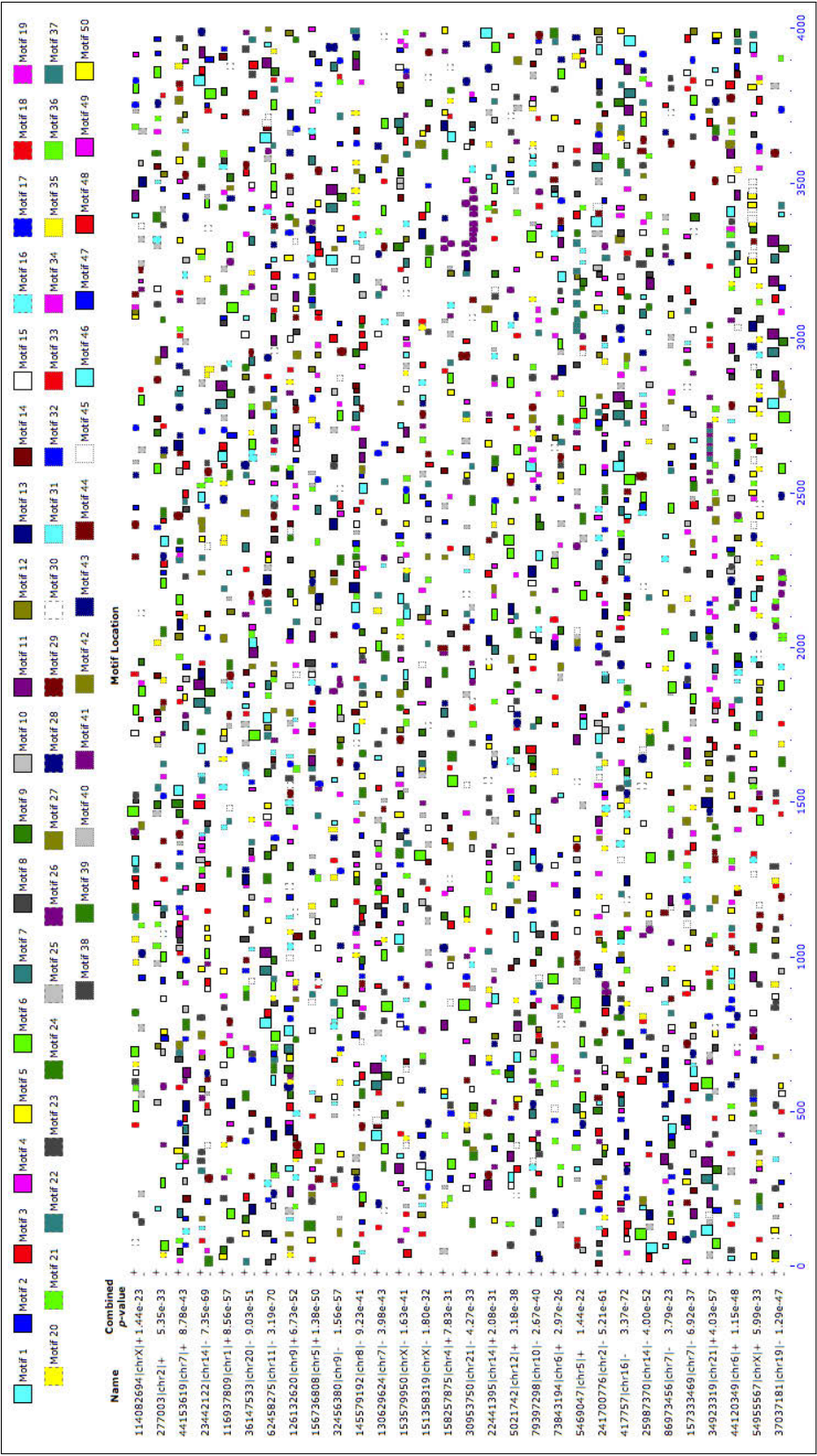


Figure 6.3: Mapping of non-palindromic motifs on sample edited sequences.

<i>Chromosome</i>	<i># nucleotides</i>	<i># of regions</i>
1	249,250,621	62,314
2	243,199,373	60,800
3	198,022,430	49,506
4	191,154,276	47,789
5	180,915,260	45,229
6	171,115,067	42,779
7	159,138,663	39,785
8	146,364,022	36,592
9	141,213,431	35,304
10	135,534,747	33,884
11	135,006,516	33,884
12	133,851,895	33,463
13	115,169,868	28,793
14	107,349,540	26,838
15	102,531,392	25,633
16	90,354,753	22,589
17	81,195,210	20,299
18	78,077,248	20,299
19	59,128,983	14,783
20	63,025,520	15,757
21	48,129,895	12,033
22	51,304,566	12,827
X	155,270,560	38,818
Y	59,373,566	14,844

Table 6.2: Number of region of 4, 000 nucleotides in each human chromosome.

For each of these genomic regions, considering both positive and negative strands, we mapped the discovered motifs by using FIMO tool [189] in order to identify those with the abundance of motifs. In the case of palindromic motifs the results did not change between positive and negative strand regions, since a palindromic sequence is the same whether read 5' to 3' on one strand or 5' to 3'. In *Figures 6.4* and *6.5* it can be seen the mapping of non-palindromic motifs in both strands of chromosome 1,

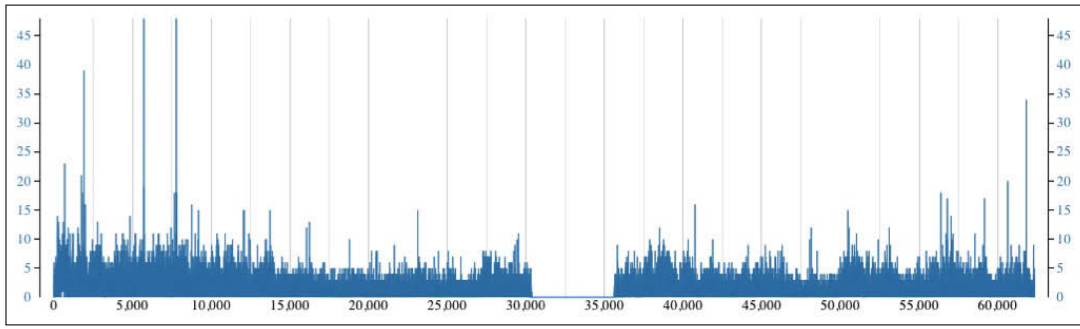


Figure 6.4: Example of mapping of non-palindromic motifs on the positive strand of the chromosome 1.

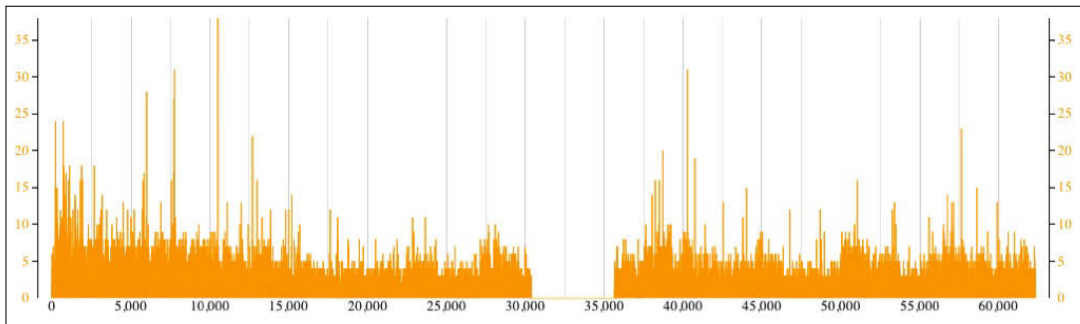


Figure 6.5: Example of mapping of non-palindromic motifs on the negative strand of the chromosome 1.

Similarly, the *Figure 6.6* shows the mapping of palindromic motifs but

only in positive strand of chromosome 1 (as said above, the result does not change).

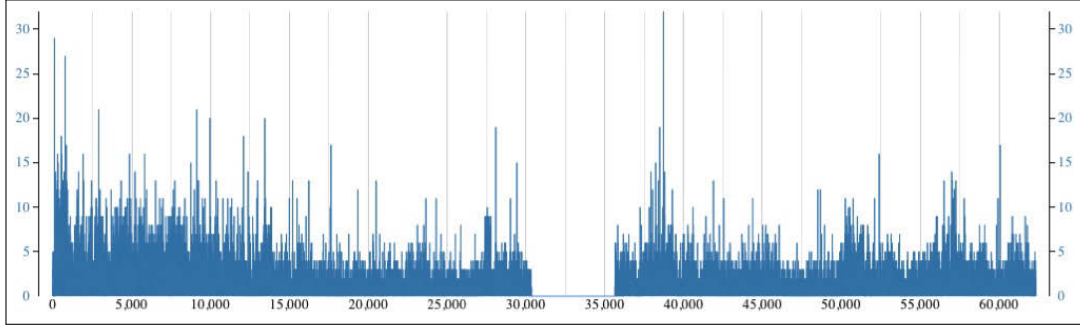


Figure 6.6: Example of mapping of palindromic motifs on the positive strand of the chromosome 1.

In all three figures it can be noticed a gap between 30,000 and 35,000 positions, which is caused by the presence of the centromere that is still not able to sequence.

Thanks to the mapping of motifs on the genome, some regions can be identified. There, it is possible to see an overlap between editing sites and motifs, as shown in *Figure 6.7*.

Then, in the next step it is necessary to automatically extract those genomic regions that contain an overlap between presence of motifs and abundance of A-to-I editing sites. This important phase could help us to establish which of the motifs may actually be correlated with the editing phenomenon.

This preliminary results show the distribution of editing and motifs. Although a global correlation is not present the analysis of local portions of the list could highlight some correlation between motifs regions and editing abundance. Such analysis is on going.

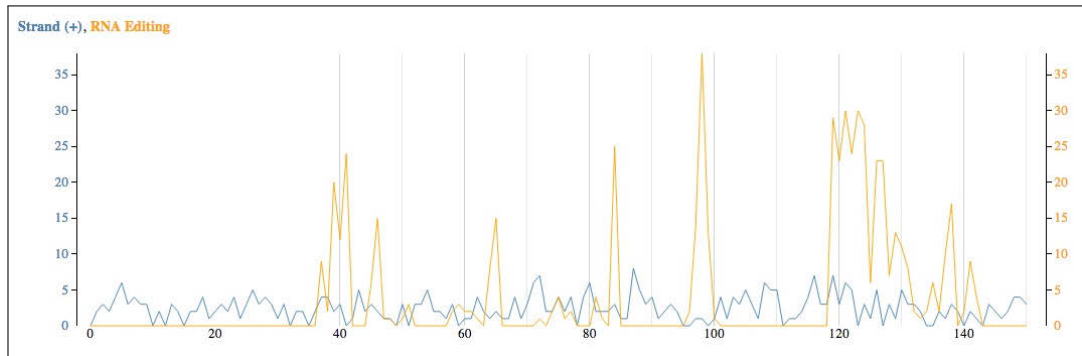


Figure 6.7: An example of overlap between A-to-I editing sites and motifs on the positive strand in the chromosome 1, at positions 24,600,000-24,200,000. The *blue* line indicates the number of motifs in a given region, while the *orange* line expresses the abundance of editing site in a genomic region.

Part V

Conclusions

Conclusions

I_N this thesis I have presented three novel biological databases: miRandola, miReditar, and Virgo.

- MiRandola is the first database of extracellular/circulating miRNA, able to indicate not only the role of miRNAs as extracellular biomarkers but also their physiological role and their involvement in diseases. The article has been published on PlosOne [\[154\]](#).
- MiReditar is the first database containing a collection of predicted human miRNA binding sites in A-to-I edited 3' UTR sequences. The article has been published on Bioinformatics [\[155\]](#).
- VIRGO is a web-based tool that maps A-to-G mismatches between genomic and EST sequences as candidate A-to-I editing sites. The article has been published on the journal BMC Bioinformatics [\[156\]](#).

Then I have presented a novel probabilistic method for the microRNA targeting, based only on sequence knowledge. Results clearly show that it performs better than *Pictar*, another system based on HMMs.

Finally, I focused on the challenging problem consisting in the identification of structural predicting motifs for the A-to-I RNA editing problem.

Future directions

Here I will sketch some future research directions on a few of the projects I have presented. While the current version of miRandola is focused on human circulating miRNAs, in the future it would be possible to introduce information on other species and different types of circulating miRNAs.

Among the future development for VIRGO, there are updates related to the inclusion of new edited sites, the prediction of the secondary structure of edited mRNA, and mapping of functional motifs in flanking regions of edited sequences. Moreover, to create a *2.0 version*, it would be important to analyze the editing phenomenon in the whole human genome, extending the analysis to all the available expressed sequences (EST) belonging to several species. In this case, the goal is to study the conservation of specific sites among different species. The availability of editing data produced from organisms other than man can give a real contribution to the study, for example, of particular diseases. It might be positive to build the relationships between the various species, observing, for example, possible sites maintained among them. A careful analysis has shown the possibility to parallelize the various

modules that make up the structure of the VIRGO, achieving concrete computational improvements. In fact, it would be useful to parallelize both the process of alignment of genomic sequences with the ESTs, and the process of prediction of the secondary structure of RNA. Another important step would be the integration between the databases VIRGO and OMIM [190] in order to associate phenotypes to editing events.

Concerning the HMM, the next step will consist in the construction of a model capable to design artificial microRNAs. In this case the profile HMM will be not conditioned and the relative math will be properly adapted.

Concerning the motif discovery problems on edited RNA sequences the the research is more tricky. The editing is a dynamic process, therefore some false positive could become true positive in a second sequencing. The goal will be the identification of shared structural motifs conserved in several species close to editing events. This will allow the identification of conserved motif which in principle could allow us to clearly understand the biological mechanism behind this complex phenomenon.

Bibliography

- [1] M. O. Dayhoff. *Atlas of protein sequence and structure, 1966*. National Biomedical Research Foundation, January 1966.
- [2] G.G. Kneale and O. Kennard. The EMBL nucleotide sequence data library. *Biochemical Society transactions*, 245(4925):1011–1014, December 1984.
- [3] D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, and E. Sayers. GenBank. *Nucleic Acids Research*, 39(Database issue):D32–D37, January 2011.
- [4] Y. Tateno, T. Imanishi, S. Miyazaki, K. Fukami-Kobayashi, N. Saitou, H. Sugawara, and T. Gojobori. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic acids research*, 30(1):27–30, January 2002.
- [5] T. Etzold and P. Argos. SRS—an indexing and retrieval tool for flat file data libraries. *Computer Applications In The Biosciences: CABIOS*, 9(1):49–57, February 1993.

- [6] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March 1970.
- [7] A. J. Gibbs and G. A. McIntyre. The diagram, a method for comparing sequences. its use with amino acid and nucleotide sequences. *European journal of biochemistry / FEBS*, 16(1):1–11, September 1970.
- [8] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, March 1981.
- [9] W. J. Wilbur and D. J. Lipman. Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences*, 80(3):726–730, February 1983.
- [10] D. J. Lipman and W. R. Pearson. Rapid and sensitive protein similarity searches. *Science (New York, N.Y.)*, 227(4693):1435–1441, March 1985.
- [11] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, October 1990.
- [12] J. W. Fickett. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research*, 10(17):5303–5318, September 1982.
- [13] M. Gribskov, J. Devereux, and R. R. Burgess. The codon preference plot: graphic analysis of protein coding sequences and prediction of

- gene expression. *Nucleic acids research*, 12(1 Pt 2):539–549, January 1984.
- [14] E. Zuckerkandl and L. Pauling. Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8(2):357–366, March 1965.
- [15] R. V. Eck and M. O. Dayhoff. Atlas of protein sequence and structure. *National Biomedical Research Foundation, Silver Spring, Maryland*, 1966.
- [16] W. M. Fitch. On the Problem of Discovering the Most Parsimonious Tree. *The American Naturalist*, 111(978):223–257, 1977.
- [17] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. *Mammalian Protein Metabolism*, pages 21–132, 1969.
- [18] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120, December 1980.
- [19] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, January 1981.
- [20] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621, July 1984.
- [21] P. Y. Chou and G. D. Fasman. Prediction of protein conformation. *Biochemistry*, (2):222–245, January 1974.

- [22] P. Y. Chou and G. D. Fasman. Empirical Predictions of Protein Conformation. *Annual Review of Biochemistry*, 47(1):251–276, 1978.
- [23] P. Y. Chou and G. D. Fasman. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol*, 47:45–148, 1978.
- [24] M. Kanehisa, S. Goto, S. Kawashima, Yasushi Okunoa, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(suppl 1):D277–D280, January 2004.
- [25] E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel, and A. Bairoch. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research*, 31(13):3784–3788, July 2003.
- [26] P. Flicek, B.L. Aken, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Gräf, S. Haider, M. Hammond, K. Howe, A. Jenkinson, N. Johnson, A. Kähäri, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, G. Koscielny, E. Kulesha, D. Lawson, I. Longden, T. Massingham, K. Megy W. McLaren, B. Overduin, B. Pritchard, D. Rios, M. Ruffier, M. Schuster, G. Slater, D. Smedley, G. Spudich, Y. A. Tang, S. Trevanion, A. Vilella, J. Vogel, S. White, S. P. Wilder, A. Zadissa, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suarez, J. Herrero, T. J. Hubbard, A. Parker, G. Proctor, J. Smith, and b S. M. Searle. Ensembl’s 10th

- year. *Nucleic acids research*, 38(Database issue):D557–562, January 2010.
- [27] P. J. Kersey, D. M. Staines, D. Lawson, E. Kulesha, P. Derwent, J. C. Humphrey, D. S. Hughes, S. Keenan, A. Kerhornou, G. Koscielny, N. Langridge, M.D. McDowall, K. Megy, U. Maheswari, M. Nuhn, M. Paulini, H. Pedro, I. Toneva, D. Wilson, A. Yates, and E. Birney. Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic acids research*, November 2011.
- [28] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and Haussler D. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, June 2002.
- [29] UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic acids research*, 41(Database issue):D43–D47, January 2013.
- [30] C. Wu and D.W. Nebert. Update on genome completion and annotations: Protein Information Resource. *Human genomics*, 3(1), March 2004.
- [31] Y. Nakamura, G. Cochrane, and I. Karsch-Mizrachi. The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, November 2012.
- [32] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(suppl 1):D140–D144, January 2006.

- [33] A. Kozomara and S. Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, 39(suppl 1):D152–D157, January 2011.
- [34] M. Magrane and U. Consortium. UniProt Knowledgebase: a hub of integrated protein data. *Database*, 2011(0):bar009, March 2011.
- [35] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics (Oxford, England)*, 23(10):1282–1288, May 2007.
- [36] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki. The Protein Data Bank. *Acta crystallographica. Section D, Biological crystallography*, 58(Pt 6 No 1):899–907, June 2002.
- [37] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, and J. M. Cherry H. Butler, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):25–29, May 2000.
- [38] The Gene Ontology Consortium. The Gene Ontology project in 2008. *Nucleic Acids Research*, 36(suppl 1):D440–D444, January 2008.

- [39] K. R. Rosenbloom, C. A. Sloan, V. S. Malladi, T. R. Dreszer, K. Learned, V. M. Kirkup, M. C. Wong, M. Maddren, R. Fang, S. G. Heitner, B. T. Lee, G. P. Barber, R. A. Harte, M. Diekhans, J. C. Long, S. P. Wilder, A. S. Zweig, D. Karolchik, R. M. Kuhn, D. Haussler, and W. J. Kent. ENCODE Data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Research*, 41(D1):D56–D63, January 2013.
- [40] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, January 2002.
- [41] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(D1):D991–D995, January 2013.
- [42] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, and A. Brazma. ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, 35(Database issue), January 2007.
- [43] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton,

- T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature genetics*, 29(4):365–371, December 2001.
- [44] P. Collas. The current state of chromatin immunoprecipitation. *Molecular biotechnology*, 45(1):87–100, May 2010.
- [45] M. J. Hawrylycz, S. Lein, A. L. Guillozet-Bongaarts, E. H. Shen, L. Ng, J. A. Miller, L. N. van de Lagemaat, K. A. Smith, A. Ebbert, Z. L. Riley, C. Abajian, C. F. Beckmann, A. Bernard, D. Bertagnolli, A. F. Boe, P. M. Cartagena, M. M. Chakravarty, M. Chapin, J. Chong, Rachel A. Dalley, B. D. Daly, C. Dang, S. Datta, N. Dee, T. A. Dolbeare, V. Faber, D. Feng, D. R. Fowler, J. Goldy, B. W. Gregor, Z. Haradon, D. R. Haynor, J. G. Hohmann, S. Horvath, R. E. Howard, A. Jeromin, J. M. Jochim, M. Kinnunen, C. Lau, E. T. Lazarz, C. Lee, T. A. Lemon, L. Li, Y. Li, J. A. Morris, C. C. Overly, P. D. Parker, S. E. Parry, M. Reding, J. J. Royall, J. Schulkin, P. A. Sequeira, C. R. Slaughterbeck, S. C. Smith, A. J. Sodt, S. M. Sunkin, B. E. Swanson, M. P. Vawter, D. Williams, P. Wohnoutka, H. R. Zielke, D. H. Geschwind, P. R. Hof, S. M. Smith, C. Koch, S. G. N. Grant, and A. R. Jones. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*, 489(7416):391–399, September 2012.
- [46] E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, and et al. Genome-wide atlas of gene expression in the adult mouse

- brain. *Nature*, 445(7124):168–176, December 2006.
- [47] C. J. Sigrist, E. de Castro, L. Cerutti, B. A. CuChe, N. Hulo, A. Bridge, L. Bougueleret, and I. Xenarios. New and continuing developments at PROSITE. *Nucleic Acids Research*, 41(Database issue):D344–D347, January 2013.
- [48] A. Mathelier and F. Parcy R. Worsley-Hunt D. J. Arenillas S. Buchman C. Y. Chen A. Chou H. Ienasescu J. Lim C. Shyr G. Tan M. Zhou B. Lenhard A. Sandelin W. W. Wasserman X. Zhao, A. W. Zhang. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, November 2013.
- [49] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.
- [50] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, November 2003.
- [51] N. Tanimura A. Funahashi, M. Morohashi, and H. Kitano. Celldesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1:159–162, 2003.

- [52] K. W. Kohn. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol Biol Cell*, 10(8):2703–2734, August 1999.
- [53] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, N. xand Voss M. Hornischer, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Winger. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(Database issue):D108–D110, January 2006.
- [54] B. Lenhard, A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg, and W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *Journal of Biology*, 2(2):13, 2003.
- [55] Ö. Johansson, W. Alkema, W. W. Wasserman, and J. Lagergren. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, 19(suppl 1):i169–i176, July 2003.
- [56] T. L. Bailey and W. S. Noble. Searching for statistically significant regulatory modules. *Bioinformatics (Oxford, England)*, 19 Suppl 2, October 2003.
- [57] S. Aerts, P. Van Loo, G. Thijs, Y. Moreau, and B. De Moor. Computational detection of cis -regulatory modules. *Bioinformatics*, 19 Suppl 2, October 2003.

- [58] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, January 2005.
- [59] M. Maragkakis, P. Alexiou, G. Papadopoulos, M. Reczko, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, V. Simossis, P. Sethupathy, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, and A. Hatzigeorgiou. Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*, 10(1):295, 2009.
- [60] A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky. Combinatorial microRNA target predictions. *Nature genetics*, 37(5):495–500, May 2005.
- [61] T. Obayashi, Y. Okamura, S. Ito, S. Tadaka, I. N. Motoike, and K. Kinoshita. COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Research*, 41(D1):D1014–D1020, January 2013.
- [62] A. Chatr-aryamontri, A. Ceol, L. M. M. Palazzi, G. Nardelli, M. V. V. Schneider, L. Castagnoli, and G. Cesareni. MINT: the Molecular Interaction database. *Nucleic acids research*, 35(Database issue):D572–D574, January 2007.
- [63] S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, and et al. Human protein reference database as a

- discovery resource for proteomics. *Nucleic Acids Research*, 32(suppl 1):D497–D501, January 2004.
- [64] G. D. Bader, D. Betel, and C. W. V. Hogue. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research*, 31(1):248–250, January 2003.
- [65] D. Croft, G. O’Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, S. Jupe, I. Kalatskaya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt, V. Shamovsky, C. Yung, E. Birney, H. Hermjakob, P. D’Eustachio, and L. Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(Database issue):D691–D697, January 2011.
- [66] J. L. Snoep and B. G. Olivier. Java Web Simulation (JWS); A Web Based Database of Kinetic Models. *Molecular Biology Reports*, 29(1):259–263, March 2002.
- [67] Vijayalakshmi Chelliah, Camille Laibe, and Nicolas Novère. BioModels Database: A Repository of Mathematical Models of Biological Processes. In Maria V. Schneider, editor, *In Silico Systems Biology*, volume 1021 of *Methods in Molecular Biology*, pages 189–199. Humana Press, 2013.
- [68] C. Schaab, T. Geiger, G. Stoehr, J. Cox, and M. Mann. Analysis of High Accuracy, Quantitative Proteomics Data in the MaxQB Database. *Molecular & Cellular Proteomics*, 11(3), March 2012.

- [69] W. Gilbert and A. Maxam. The nucleotide sequence of the lac operator. *Proceedings of the National Academy of Sciences of the United States of America*, 70(12):3581–3584, December 1973.
- [70] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–448, May 1975.
- [71] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, December 1977.
- [72] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2):560–564, February 1977.
- [73] O. Morozova and M. A. Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264, November 2008.
- [74] M. L. Metzker. Sequencing technologies [mdash] the next generation. *Nat Rev Genet*, 11(1):31–46, January 2010.
- [75] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)*, 309(5741):1728–1732, September 2005.

- [76] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, and et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, September 2005.
- [77] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, and et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, November 2008.
- [78] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, and S. M. Johnson. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome research*, 18(7):1051–1063, July 2008.
- [79] C. S. Pareek, R. Smoczynski, and A. Tretyn. Sequencing technologies and genome sequencing. *Journal of applied genetics*, 52(4):413–435, November 2011.
- [80] M. Quail, M. Smith, P. Coupland, T. Otto, S. Harris, T. Connor, A. Bertoni, H. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13(1):341+, July 2012.
- [81] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012:1–11, 2012.

- [82] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25–10, March 2009.
- [83] M. Brown, R. Hughey, A. Krogh, I. S. Mian, K. Sjölander, and D. Haussler. Using Dirichlet mixture priors to derive hidden Markov models for protein families. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 1:47–55, 1993.
- [84] K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, and D. Haussler. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, 12(4):327–345, August 1996.
- [85] Y. Ding, C. Y. Chan, and C. E. Lawrence. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, 11(8):1157–1166, August 2005.
- [86] L. E. Carvalho and C. E. Lawrence. Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proceedings of the National Academy of Sciences*, 105(9):3209–3214, March 2008.
- [87] M. Hamada, H. Kiryu, W. Iwasaki, and K. Asai. Generalized centroid estimators in bioinformatics. *PloS one*, 6(2):e16450+, February 2011.
- [88] M. Hamada, H. Kiryu, K. Sato, T. Mituyama, and K. Asai. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, 25(4):465–473, February 2009.

- [89] C. B. Do, D. A. Woods, and S. Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics (Oxford, England)*, 22(14):e90–e98, July 2006.
- [90] L. He and G. J. Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nature reviews. Genetics*, 5(7):522–531, July 2004.
- [91] V. N. Kim. Small RNAs: classification, biogenesis, and function. *Molecules and cells*, 19(1):1–15, February 2005.
- [92] C. P. Petersen, M. Bordeleau, J. Pelletier, and P. A. Sharp. Short RNAs Repress Translation after Initiation in Mammalian Cells. *Molecular Cell*, 21(4):533–542, February 2006.
- [93] V. Ambros. The functions of animal microRNAs. *Nature*, 431:350–355, 2004.
- [94] W. P. Kloosterman and R. H. Plasterk. The diverse functions of microRNAs in animal development and disease. *Developmental cell*, 11(4):441–450, October 2006.
- [95] T. M. Rana. Illuminating the silence: understanding the structure and function of small RNAs. *Nature Reviews Molecular Cell Biology*, 8(1):23–36, January 2007.
- [96] D. Catalucci, P. Gallo, and G. Condorelli. MicroRNAs in Cardiovascular Biology and Heart Disease. *Circulation Cardiovascular Genetics*, pages 402–408, June 2012.

- [97] W. J. Lukiw and A. I. Pogue. Induction of specific micro RNA (miRNA) species by ROS-generating metal sulfates in primary human brain cells. *Journal of inorganic biochemistry*, 101(9):1265–1269, September 2007.
- [98] H. Xie, B. Lim, and H. F. Lodish. MicroRNAs induced during adipogenesis that accelerate fat cell development are downregulated in obesity. *Diabetes*, 58(5):1050–1057, May 2009.
- [99] A. Esquela-Kerscher and F. J. Slack. Oncomirs - microRNAs with a role in cancer. *Nature reviews. Cancer*, 6(4):259–269, April 2006.
- [100] D. P. Bartel. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, 116:281–297, 2004.
- [101] R. F. Ketting, S. E. J. Fischer, E. Bernstein, T. Sijen, G. J. Hannon, and R. H. A. Plasterk. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes & Development*, 15(20):2654–2659, October 2001.
- [102] D. S. Schwarz, G. Hutvágner, T. Du, Z. Xu, N. Aronin, and P. D. Zamore. Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, 115(2):199–208, October 2003.
- [103] X. Li and R. W. Carthew. A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the *Drosophila* eye. *Cell*, 123(7):1267–1277, December 2005.

- [104] K. Seggerson, L. J. Tang, and E. G. Moss. Two genetic circuits repress the *Caenorhabditis elegans* heterochronic gene *lin-28* after translation initiation. *Developmental Biology*, 243:215–225, 2002.
- [105] C. B. Nielsen, N. Shomron, R. Sandberg, E. Hornstein, J. Kitzman, and C. B. Burge. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA (New York, N.Y.)*, 13(11):1894–1910, November 2007.
- [106] G. Easow, A. A. Teleman, and S. M. Cohen. Isolation of microRNA targets by miRNP immunopurification. *RNA*, 13:1198–1204, August 2007.
- [107] S. Bagga, J. Bracht, S. Hunter, and et al. Regulation by *let-7* and *lin-4* miRNAs results in target mRNA degradation. *Cell*, 122(4):553–563, August 2005.
- [108] I. Behm-Ansmant, J. Rehwinkel, T. Doerks, A. Stark, P. Bork, and E. Izaurralde. mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. *Genes & development*, 20(14):1885–1898, July 2006.
- [109] L. M. Alemán, J. Doench, and P. A. Sharp. Comparison of siRNA-induced off-target RNA and protein effects. *RNA*, 13(3):385–395, March 2007.
- [110] K. A. O’Donnell, E. A. Wentzel, K. I. Zeller, C. V. Dang, and J. T. Mendell. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, 435(7043):839–843, June 2005.

- [111] C. Conaco, S. Otto, J. Han, and G. Mandel. Reciprocal actions of REST and a microRNA promote neuronal identity. *PNAS*, 103:2422–2427, February 2006.
- [112] J. Krol, I. Loedige, and W. Filipowicz. The widespread regulation of microRNA biogenesis, function and decay. *Nature reviews. Genetics*, 11(9):597–610, September 2010.
- [113] Z. Paroo, X. Ye, S. Chen, and Q. Liu. Phosphorylation of the Human MicroRNA-Generating Complex Mediates MAPK/Erk Signaling. *Cell*, 139(1):112–122, October 2009.
- [114] J. G. Doench and P. A. Sharp. Specificity of microRNA target selection in translational repression. *Genes & Development*, 18(5):504–511, March 2004.
- [115] M. Kiriakidou, P. T. Nelson, A. Kouranov, P. Fitziev, C. Bouyioukos, Z. Mourelatos, and A. Hatzigeorgiou. A combined computational-experimental approach predicts human microRNA targets. *Genes and Development*, 18(10):1165–1178, May 2004.
- [116] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA (New York, N.Y.)*, 10(10):1507–1517, October 2004.
- [117] B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798, December 2003.

- [118] P. S. Mitchell, R. K. Parkin, E. M. Kroh, B. R. Fritz, S. K. Wyman, E. L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K. C. O'Briant, A. Allen, D. W. Lin, N. Urban, C. W. Drescher, B. S. Knudsen, D. L. Stirewalt, R. Gentleman, R. L. Vessella, P. S. Nelson, D. B. Martin, and M. Tewari. Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences*, 105(30):10513–10518, July 2008.
- [119] Y. Tomimaru, H. Eguchi, H. Nagano, H. Wada, S. Kobayashi, S. Marubashi, M. Tanemura, A. Tomokuni, I. Takemasa, K. Umeshita, Y. Doki T. Kanto, and M. Mori. Circulating microRNA-21 as a novel biomarker for hepatocellular carcinoma. *Journal of Hepatology*, 56:167–175, July 2011.
- [120] S. K. Gupta, C. Bang, and T. Thum. Circulating MicroRNAs as Biomarkers and Potential Paracrine Mediators of Cardiovascular Disease. *Circ Cardiovasc Genet*, 3:484–488, 2010.
- [121] A. M. Zahm, M. Thayu, N. J. Hand, A. Horner, and J. R. Friedman M. B. Leonard. Circulating microRNA is a biomarker of pediatric Crohn disease. *J Pediatr Gastroenterol Nutr.*, 53(1), July 2011.
- [122] B. Février and G. Raposo. Exosomes: endosomal-derived vesicles shipping extracellular messages. *Current Opinion in Cell Biology*, 16(4):415–421, August 2004.
- [123] J. D. Arroyo, J. R. Chevillet, E. M. Kroh, I. K. Ruf, C. C. Pritchard, D. F. Gibson, P. S. Mitchell, C. F. Bennett, E. L. Pogosova-

- Agadjanyan, D. L. Stirewalt, J. F. Tait, and M. Tewari. Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proceedings of the National Academy of Sciences*, 108(12):5003–5008, March 2011.
- [124] A. Turchinovich, L. Weiz, A. Langhein, and B. Burwinkel. Characterization of extracellular circulating microRNA. *Nucleic Acids Research*, 39(16):7223–7233, September 2011.
- [125] J. J. Song, J. Liu, N. H. Tolia, J. Schneiderman, S. K. Smith, R. A. Martienssen, G. J. Hannon, and L. Joshua-Tor. The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nat Struct Biol*, 10(12):1026–1032, December 2003.
- [126] J. Ma, K. Ye, and D. J. Patel. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature*, 429(6989):318–322, May 2004.
- [127] H. Valadi, K. Ekström, A. Bossios, M. Sjöstrand, J. J. Lee, and J. O. Lötvall. Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nature cell biology*, 9(6):654–659, June 2007.
- [128] N. Kosaka, H. Iguchi, Y. Yoshioka, F. Takeshita, Y. Matsuki, and T. Ochiya. Secretory Mechanisms and Intercellular Transfer of MicroRNAs in Living Cells. *Journal of Biological Chemistry*, 285(23):17442–17452, June 2010.

- [129] R. Benne. Rna editing in trypanosomes. *Biochimica et biophysica*, 221(1):9–23, Apr 1989.
- [130] R. Benne. *RNA Editing: The Alteration of Protein Coding Sequences of RNA*. Prentice Hall, 1993.
- [131] H. Grosjean and R. Benne. *Modification and Editing of RNA*. American Society of Microbiology Press, 1997.
- [132] R. Hiesel, B. Wissinger, W. Schuster, and A. Brennicke. Rna editing in plant mitochondria. *Science*, 246(4937):1632–1634, Dec 1989.
- [133] R. Bock, M. Hermann, and M. Fuchs. Identification of critical nucleotide positions for plastid rna editing site recognition. *RNA*, 3(10):1194–1200, Oct 1997.
- [134] B. Wissinger, A. Brennicke, and W. Schuster. Regenerating good sense: Rna editing and trans splicing in plant mitochondria. *Trends Genet*, 8(9):322–328, Sep 1992.
- [135] D. Pring, A. Brennicke, and W. Schuster. Rna editing gives a new meaning to the genetic information in mitochondria and chloroplasts. *Plant Mol Biol*, 21(6):1163–1170, Mar 1993.
- [136] B. Hoch, R. M. Maier, K. Appel, G. L. Igloi, and H. Kössel. Editing of a chloroplast mrna by creation of an initiation codon. *Nature*, 353(6340):178–180, Sep 1991.
- [137] J. Curran, R. Boeck, and D. Kolakofsky. The sendai virus p gene expresses both an essential protein and an inhibitor of rna synthesis by

- shuffling modules via mrna editing. *EMBO J*, 10(10):3079–3085, Oct 1991.
- [138] B. Sommer, M. Köhler, R. Sprengel, and P. H. Seeburg. Rna editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell*, 67(1):11–19, Oct 1991.
- [139] M. S. Hershfield. Adenosine deaminase deficiency: clinical expression, molecular basis, and therapy. *Semin Hematol*, 35(4):291–298, Oct 1998.
- [140] B. L. Bass and H. Weintraub. A developmentally regulated activity that unwinds rna duplexes. *Cell*, 48(4):607–613, Feb 1987.
- [141] R. W. Wagner and K. Nishikura. Cell cycle expression of rna duplex unwindase activity in mammalian cells. *Mol Cell Biol*, 8(2):770–777, Feb 1988.
- [142] H. U. Weier, C. X. George, K. M. Greulich, and C. E. Samuel. The interferon-inducible, double-stranded rna-specific adenosine deaminase gene (dsrad) maps to human chromosome 1q21.1-21.2. *Genomics*, 30(2):372–375, Nov 1995.
- [143] U. Kim, T. L. Garner, T. Sanford, D. Speicher, J. M. Murray, and K. Nishikura. Purification and characterization of double-stranded rna adenosine deaminase from bovine nuclear extracts. *J Biol Chem*, 269(18):13480–13489, May 1994.

- [144] M. A. O'Connell and W. Keller. Purification and properties of double-stranded rna-specific adenosine deaminase from calf thymus. *Proc Natl Acad Sci U S A*, 91(22):10596–10600, Oct 1994.
- [145] N. Navaratnam, T. Fujino, J. Bayliss, A. Jarmuz, A. How, N. Richardson, A. Somasekaram, S. Bhattacharya, C. Carter, and J. Scott. Escherichia coli cytidine deaminase provides a molecular model for apob rna editing and a mechanism for rna substrate recognition. *J Mol Biol*, 275(4):695–714, Jan 1998.
- [146] K. M. Lonergan and M. W. Gray. Predicted editing of additional transfer rnas in acanthamoeba castellanii mitochondria. *Nucleic Acids Res*, 21(18):4402–4402, Sep 1993.
- [147] P. S. Covello and M. W. Gray. On the evolution of rna editing. *Trends Genet*, 9(8):265–268, Aug 1993.
- [148] Q. Wang, Z. Zhang, K. Blackwell, and G. G. Carmichael. Vigilins bind to promiscuously a-to-i-edited rnas and are involved in the formation of heterochromatin. *Curr Biol*, 15(4):384–391, Feb 2005.
- [149] D. J. Luciano, H. Mirsky, N. J. Vendetti, and S. Maas. Rna editing of a mirna precursor. *RNA*, 10(8):1174–1177, Aug 2004.
- [150] S. Pfeffer, A. Sewer, M. Lagos-Quintana, R. Sheridan, C. Sander, F. A. Grässer, L. F. van Dyk, C. K. Ho, S. Shuman, M. Chien, J. J. Russo, J. Ju, G. Randall, B. D. Lindenbach, C. M. Rice, V. Simon, D. D. Ho, M. Zavolan, and T. Tuschl. Identification of micrornas of the herpesvirus family. *Nat Methods*, 2(4):269–276, Apr 2005.

- [151] W. Yang, T. P. Chendrimada, Q. Wang, M. Higuchi, P. H. Seeburg, R. Shiekhataar, and K. Nishikura. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat Struct Mol Biol*, 13(1):13–21, Jan 2006.
- [152] M. J. Blow, R. J. Grocock, S. van Dongen, A. J. Enright, E. Dicks, P. A. Futreal, R. Wooster, and M. R. Stratton. RNA editing of human microRNAs. *Genome Biol*, 7(4), 2006.
- [153] G. M. Borchert, B. L. Gilmore, R. M. Spengler, Y. Xing, W. Lanier, D. Bhattacharya, and B. L. Davidson. Adenosine deamination in human transcripts generates novel microRNA binding sites. *Hum Mol Genet*, 18(24):4801–4807, Dec 2009.
- [154] F. Russo, S. Di Bella, G. Nigita, V. Macca, A. Laganà, R. Giugno, A. Pulvirenti, and A. Ferro. miRandola: Extracellular Circulating MicroRNAs Database. *PLoS ONE*, 7(10):e47786+, October 2012.
- [155] A. Laganà, A. Paone, D. Veneziano, L. Cascione, P. Gasparini, S. Carasi, F. Russo, G. Nigita, V. Macca, R. Giugno, A. Pulvirenti, D. Shasha, A. Ferro, and C. M. Croce. miR-EdiTar: a database of predicted A-to-I edited miRNA target sites. *Bioinformatics*, 28(23):3166–3168, December 2012.
- [156] V. Macca, A. Laganà, R. Giugno, A. Pulvirenti, A. Ferro, R. Distefano, G. Nigita. VIRGO: Visualization of A-to-I RNA editing sites in genomic sequences. 14(Suppl 7), April 2013.

- [157] A. Giudice M.R. Arena P.L. Puglisi R. Giugno A. Pulvirenti D. Shasha A. Ferro A. Laganà, S. Forte. miRò: a miRNA knowledge base. 2009(bap008), August 2009.
- [158] K. C. Vickers, B. T. Palmisano, B. M. Shoucri, R. D. Shamburek, and A. T. Remaley. MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. *Nature cell biology*, 13(4):423–433, April 2011.
- [159] J. L.Hood, R. S. San, and S. A. Wickline. Exosomes Released by Melanoma Cells Prepare Sentinel Lymph Nodes for Tumor Metastasis. *Cancer Research*, 71(11):3792–3801, June 2011.
- [160] D. Dominissini, S. Moshitch-Moshkovitz, N. Amariglio, and G. Rechavi. Adenosine-to-inosine RNA editing meets cancer. *Carcinogenesis*, 32(11):1569–1577, June 2011.
- [161] A. Gallo and F. Locatelli. ADARs: allies or enemies? The importance of A-to-I RNA editing in human disease: from cancer to HIV-1. *Biological Reviews*, 87(1):95–110, 2012.
- [162] K. Nishikura. Functions and regulation of RNA editing by ADAR deaminases. *Annual review of biochemistry*, 79:321–349, 2010.
- [163] S. Alon, E. Mor, F. Vigneault, G. Church, F. Locatelli, F. Galeano, A. Gallo, N. Shomron, and E. Eisenberg. Systematic identification of edited microRNAs in the human brain. *Genome research*, April 2012.

- [164] Y. Kawahara, B. Zinshteyn, P. Sethupathy, H. Iizasa, A. G. Hatzigeorgiou, and K. Nishikura. Redirection of Silencing Targets by Adenosine-to-Inosine Editing of miRNAs. *Science*, 315(5815):1137–1140, February 2007.
- [165] E. Y. Levanon, E. Eisenberg, R. Yelin, S. Nemzer, M. Hallegger, R. Shemesh, Z. Y. Fligelman, A. Shoshan, S. R. Pollock, D. Sztybel, M. Olshansky, G. Rechavi, and M. F. Jantsch. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nature biotechnology*, 22(8):1001–1005, August 2004.
- [166] A. Athanasiadis, A. Rich, and S. Maas. Widespread A-to-I RNA Editing of Alu-Containing mRNAs in the Human Transcriptome. *PLoS Biol*, 2(12):e391+, November 2004.
- [167] H. Liang and L. F. Landweber. Hypothesis: RNA editing of microRNA target sites in humans? *RNA*, 13(4):463–467, April 2007.
- [168] A. Kiran and P. V. Baranov. DARNED: a DAtabase of RNa EDiting in humans. *Bioinformatics (Oxford, England)*, 26(14):1772–1776, July 2010.
- [169] A. Laganà, S. Forte, F. Russo, R. Giugno, A. Pulvirenti, and A. Ferro. Prediction of human targets for viral-encoded microRNAs by thermodynamics and empirical constraints. *J RNAi Gene Silencing*, 6(1):379–385, May 2010.

- [170] S. H. Bernhart, I. L. Hofacker, and P. F. Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics (Oxford, England)*, 22(5):614–615, March 2006.
- [171] R. M. Marin and J. Vanicek. Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Research*, 39(1):19–29, January 2011.
- [172] R. Lorenz, S. H. Bernhart, C. Honer, Zu Siederdisen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB*, 6(1):26+, November 2011.
- [173] J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen. Principles of MicroRNA,ÀTarget Recognition. *PLoS Biol*, 3(3):e85+, February 2005.
- [174] M. Halvorsen, J. S. Martin, S. Broadaway, and A. Laederach. Disease-Associated Mutations That Alter the RNA Structural Ensemble. *PLoS Genet*, 6(8):e1001074+, August 2010.
- [175] L. X. Shen, J. P. Basilion, and V. P. Jr. Stanton. Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc Natl Acad Sci U S A.*, 96(14):7871–7876, July 1999.
- [176] U. Haas, G. Sczakiel, and S. D. Laufer. MicroRNA-mediated regulation of gene expression is affected by disease-associated SNPs within the 3'-UTR via altered RNA structure. *RNA Biol.*, 9(6):924–937, June 2012.

- [177] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal. The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39(10):1278–1284, September 2007.
- [178] T. He, P. Du, and Y. Li. dbRES: a web-oriented database for annotated RNA editing sites. *Nucleic Acids Res.*, 35(Database issue):D141–D144, November 2006.
- [179] A. M. Kiran, J. J. O’Mahony, K. Sanjeev, and P. V. Baranov. Darned in 2013: inclusion of model organisms and linking with Wikipedia. *Nucleic Acids Res*, 41(Database issue), October 2013.
- [180] E. Picardi, M. D’Antonio, D. Carrabino, T. Castrignanà, and G. Pesole. ExpEdit: a webserver to explore human RNA editing in RNA-Seq experiments. *Bioinformatics*, 27(9):1311–1312, March 2011.
- [181] G. Ramaswami and J. B. Li. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res*, October 2013.
- [182] Y. Neeman, E. Y. Levanon, M. F. Jantsch, and E. Eisenberg. RNA editing level in the mouse is determined by the genomic repeat repertoire. *RNA (New York, N.Y.)*, 12(10):1802–1809, October 2006.
- [183] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, January 1998.
- [184] S. W. Burge, J. Daub, R. Eberhardt, J. Tate, L. Barquist, E. P. Nawrocki, S. R. Eddy, P. P. Gardner, and A. Bateman. Rfam 11.0:

- 10 years of RNA families. *Nucleic Acids Research*, 41(D1):D226–D232, January 2013.
- [185] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn. The Pfam protein families database. *Nucleic Acids Research*, 40(D1):D290–D301, January 2012.
- [186] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li. miRecords: an integrated resource for microRNA-target interactions. *Nucleic acids research*, 37(Database issue):D105–110, January 2009.
- [187] T. Vergoulis, I. S. Vlachos, P. Alexiou, G. Georgakilas, M. Maragkakis, M. Reczko, S. Gerangelos, N. Koziris, T. Dalamagas, and A. G. Hatzi-georgiou. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic acids research*, 40(Database issue):D222–D229, January 2012.
- [188] C. Elkan T.L. Bailey. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.
- [189] W.S. Noble C.E. Grant, T.L. Bailey. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, February 2011.
- [190] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowl-

edgebase of human genes and genetic disorders. *Nucleic acids research*,
33(Database issue), January 2005.